UNIVERSIDADE DE SÃO PAULO
ESCOLA DE COMUNICAÇÕES E ARTES
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

JANAILTON LOPES SOUSA

PRINCÍPIOS PARA AVALIAÇÃO DE VOCABULÁRIOS CONTROLADOS

São Paulo
2023

UNIVERSIDADE DE SÃO PAULO
ESCOLA DE COMUNICAÇÕES E ARTES
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

JANAILTON LOPES SOUSA

PRINCÍPIOS PARA AVALIAÇÃO DE VOCABULÁRIOS CONTROLADOS

VERSÃO CORRIGIDA
(VERSÃO ORIGINAL DISPONÍVEL NA BIBLIOTECA DA ECA/USP)

Tese apresentada ao Programa de Pós-graduação em Ciência da Informação da Escola de Comunicações e Artes da Universidade de São Paulo, para obtenção do título de Doutor.
Programa: Ciência da Informação
Área de concentração: Cultura e Informação
Linha de pesquisa: Organização da Informação e do Conhecimento.
Orientadora: Prof.ª Dr.ª Vânia Mara Alves Lima

São Paulo
2023

Catalogação na Publicação
Serviço de Biblioteca e Documentação
Escola de Comunicações e Artes da Universidade de São Paulo
Dados inseridos pelo(a) autor(a)

Sousa, Janailton Lopes

Princípios para avaliação de vocabulários controlados / Janailton Lopes Sousa; orientadora, Vânia Mara Alves Lima. - São Paulo, 2023. 172 p.

Tese (Doutorado) - Programa de Pós-Graduação em Ciência da Informação / Escola de Comunicações e Artes / Universidade de São Paulo.
Bibliografia Versão corrigida

1. Vocabulário controlado. 2. Organização do Conhecimento. 3. Tesauros. 4. Avaliação de tesauros. I. Lima, Vânia Mara Alves . II. Título.

CDD 21.ed. - 020

Elaborado por Alessandra Vieira Canholi Maldonado - CRB-8/6194

SOUSA, Janailton Lopes. **Princípios para avaliação de vocabulários controlados**. Orientação: Vânia Mara Alves Lima. 2023. 172 f. Tese (Doutorado) – Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 2023.

Aprovado em: / /

Banca Examinadora:

_____

Profª. Drª. Vânia Mara Alves Lima - Orientadora

Universidade de São Paulo (USP)

_____

Prof. Dr. Marivalde Moacir Francelin

Universidade de São Paulo (USP)

_____

Prof. Dr. Rogério Aparecido Sá Ramalho

Universidade de Federal de São Carlos (UFSCar)

_____

Profª. Drª. Brígida Maria Nogueira Cervantes

Universidade Estadual de Londrina (UEL)

_____

Profª. Drª. Valdirene Pereira da Conceição

Universidade Federal do Maranhão (UFMA)

Dedico à minha família

AGRADECIMENTOS

Agradeço a Deus pelo dom da vida e pelo conhecimento que tenho acesso.

À minha amada e querida esposa Joyce, pelo apoio incondicional e ao meu querido filho Isaac, que impulsionou a conclusão deste trabalho, aos amigos e familiares que de alguma forma contribuiram diretamente ou indiretamente.

Agradeço a Profª. Vânia Lima pela orientação, paciência e sabedoria para me guiar durante a pesquisa. Ao prof. Fernando Modesto pelo período de aprendizado no Programa de Estágio Supervisionado.

Minha Gratidão a banca do Exame de Qualificação e a Banca de avaliação por aceitarem o convite e desafio.

Agradeço à Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro à pesquisa, o que permitiu dedicação exclusiva à pesquisa.

Agradeço ao Programa de Pós Graduação em Ciência da Informação, a Escola de Comunicação e Artes e a Universidade de São Paulo, assim como as instâncias superiores que permitem o ensino público e gratuito neste país.

"Todo conhecimento comporta o risco do erro e da ilusão" (Edgar Morin, 2000).

RESUMO

SOUSA, Janailton Lopes. **Princípios para avaliação de vocabulários controlados**. Orientação: Vânia Mara Alves Lima. 2023. 172 f. Tese (Doutorado) – Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 2023.

Introdução: Vocabulários controlados são conjunto de termos empregados tanto na indexação, como na recuperação da informação. Utilizados como instrumento de controle terminológico, com intuito de padronizar a comunicação científica. O Resource Description Framework e o Simple Knowledge Organization System são modelos de dados que permitem representar vocabulários controlados em ambientes digitais. **Objetivos**: Discutir a avaliação de vocabulários controlados a partir de três tipos distintos de análise: quantitativa, baseada em modelos de dados e métrica de termos. **Métodos**: Trata-se de uma pesquisa aplicada, cuja análise caracteriza 10 tipos de vocabulários controlados do tipo tesauro disponíveis em RDF e SKOS processados de forma automática com auxílio de linguagem de programação. **Resultados**: Com base em dados estatísticos foi possível mapear tipos de tags por vocabulário controlado, verificar os principais erros associados à estrutura dos modelos de dados e extrair métricas a partir dos termos contidos em cada vocabulário controlado. **Conclusão**: Por meio de um panorama geral de comparação dos vocabulários controlados em cada modalidade de análise, constatando as principais diferenças e similaridades encontradas, por fim apresenta princípios que podem ser adotados para avaliar vocabulários controlados disponíveis em RDF e SKOS.

Palavras-chave: Terminologia. Vocabulários controlados. Tesauros. RDF. SKOS

ABSTRACT

SOUSA, Janailton Lopes. **Principles for evaluating controlled vocabularies**. Orientation: Vânia Mara Alves Lima. 2023. 172f. Thesis (Doctorate in Information Science) – School of Communication and Arts, University of São Paulo, São Paulo, 2023.

Introduction: Controlled vocabularies are a set of terms used in both indexing and information retrieval. Used as a terminological control instrument, with the aim of standardizing scientific communication. The Resource Description Framework and the Simple Knowledge Organization System are data models that allow representing controlled vocabularies in digital environments. **Objectives**: Discuss the evaluation of controlled vocabularies based on three distinct types of analysis: quantitative, based on data models and term metrics. **Methods**: This is an applied research, whose analysis characterizes 10 types of controlled thesaurus-type vocabularies available in RDF and SKOS processed automatically with the aid of a programming language. **Results**: Based on statistical data, it was possible to map types of tags by controlled vocabulary, check the main errors associated with the structure of the data models and extract metrics from the terms contained in each controlled vocabulary. **Conclusion**: Through a general overview of comparison of controlled vocabularies in each analysis modality, noting the main differences and similarities found, it finally presents principles that can be adopted to evaluate controlled vocabularies available in RDF and SKOS.

Keywords: Terminology. Controlled vocabularies. Thesauri. RDF. SKOS

LISTA DE SIGLAS

API Application Programming Interface
ANSI American National Standard Institute
BERT Bidirectional Encoder Representations from Transformers
CSV Comma-separated values
HTML HyperText Markup Language
IRI Identificador de Recursos Internacionalizados
ISKO International Society for Knowledge Organization
ISO International Organization for Standardization
JSON JavaScript Object Notation
KO Knowledge Organization
KOS Knowledge Organization System
NISO National Information Standards Organization
NLP Natural Language Processing
OWL Web Ontology Language
PDF Portable Document Format
PLN Processamento da Linguagem Natural
RDF Resource Description Framework
SGML Standard Generalized Markup Language
SKOS Simple Knowledge Organization System
SOC Sistema de Organização do Conhecimento
SRI Sistema de Recuperação de Informação
TCT Teoria Comunicativa da Terminologia
TGT Teoria Geral da Terminologia
UEL - Universidade Estadual de Londrina
USP Universidade de São Paulo
URI Uniform Resource Identifier
W3C World Wide Web Consortium
WWW World Wide Web
XML eXtensible Markup Language

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

O vocabulário controlado é um instrumento de controle terminológico utilizado para a representação e recuperação da informação. Embora existam controvérsias sobre o controle da língua, não se pode negar um conceito autodeclarado pelo próprio termo. De modo generalista e fora do escopo da Ciência da Informação, Figueiredo (2010) afirma que um vocabulário é um conjunto de termos ou vocábulos, pertencentes a uma arte ou ciência, e esta palavra vem do latim *vocabularium*. No Dicionário de biblioteconomia e arquivologia encontra-se uma definição não muito distinta, onde um vocabulário em seu sentido mais amplo é um conjunto de vocábulos (palavras e termos) de uma língua (Cunha; Cavalcanti, 2008).

Um vocabulário controlado também pode ser considerado como um conjunto de termos empregados tanto no momento da indexação como na recuperação da informação, com o propósito de alinhar a linguagem do pesquisador com a do indexador (Cunha; Cavalcanti, 2008). Neste sentido, os vocabulários controlados podem ser declarados como um conjunto de termos adotados em sistema informação, para controle terminológico, com intuito de possibilitar a representação de conceitos por meio dos termos, assim como a recuperação da informação. Os vocabulários controlados podem ser considerados Sistemas de Organização do Conhecimento (SOC) e, dessa forma, caracterizados como taxonomias, tesauros, listas de assuntos, dicionários de sinônimos, etc.

Um vocabulário controlado, do tipo tesauro, se apresenta como um instrumento de controle terminológico dentro de um domínio do conhecimento científico, utilizado para representação e recuperação da informação. Zeng (2008) em seu gráfico de tipologias de SOCs, enfatiza a diversidade de sistemas simples e complexos, os quais podem variar entre listas de termos, dicionários de sinônimos, taxonomias, tesauros, ontologias e redes semânticas. É importante destacar que nem todos os SOCs são vocabulários controlados, como é o caso das ontologias e redes semânticas, porém, todos os vocabulários controlados são tipos de SOCs.

Como todos os vocabulários controlados são SOCs e todos os SOCs se apresentam de formas diversas, os vocabulários controlados também são apresentados de formas diversas. Também possuem níveis de complexidade em dois eixos: estrutura e função. Conforme ressalta Zeng (2008). Quanto à estrutura podem ser estáveis, com duas ou múltiplas

dimensões, progressivamente apresentada por meio de lista de termos, modelos de metadados, sistemas de classificação, de categorização e modelos de relacionamento. E por função incluem a eliminação de ambiguidade, controle de sinônimos, estabelecimento de relações semânticas, hierárquicas e associativas, a apresentação de relações e propriedades de conceitos nos modelos de conhecimento.

Considerando a estrutura e função de um SOC, um vocabulário controlado pode ser avaliado por meio de diferentes perspectivas. Existem diversos métodos de avaliação de vocabulários controlados, todavia, a proposição de princípios para este processo de avaliação, surge como necessidade implícita à própria manutenção, verificações de erros e mapeamento estrutural. Para uso recorrente neste trabalho adotou-se a expressão "vocabulário controlado", em termos gerais, devido às normas técnicas que orientam o desenvolvimento de diversos tipos de vocabulários controlados e não apenas dos tesauros.

O tesauro é o principal tipo de vocabulário controlado analisado na pesquisa, devido a variedade de estudos que envolvem esse instrumento de representação, assim como a disposição de normas e modelos de dados. O uso da expressão 'avaliação de vocabulários controlados' foi tomado como referência para esta pesquisa e não se limita a um tipo específico de vocabulário controlado, embora tenha explorado os tesauros, os princípios aqui apresentados podem ser aplicados em diferentes tipos de SOCs, desde que estejam disponíveis em modelos de dados equivalentes.

Como as tecnologias estão cada vez mais presentes nos diversos campos do conhecimento, os vocabulários controlados também estão disponíveis por meio de softwares e padronizados em estruturas de dados, como o *Resource Description Framework* (RDF) 1.1 Primer (Manola, Miller e McBride, 2014) e o *Simple Knowledge Organization System* (SKOS) Primer (Isaac e Summers, 2009) nomeados pela *World Wide Web Consortium* (W3C) como modelos de dados. O SKOS surge da tentativa de expressar tesauros em RDF, enquanto o RDF foi concebido para representar informações na web, portanto ambos surgem atrelados aos SOCs e consequentemente aos vocabulários controlados.

A avaliação de vocabulários controlados abrange diversos elementos, como performance, planejamento, construção e manutenção. Em alguns casos inclui também o software utilizado na implementação e disponibilização, tendo em vista que muitos

vocabulários são desenvolvidos a partir de softwares específicos, por exemplo *Tematres*[1], *Skosmos*[2], *VocBench*[3], dentre outros, pode-se pensar também uma possível avaliação dos softwares utilizados no desenvolvimento de vocabulários controlados.

A maioria dos softwares geram algum tipo de dado, que pode ser apenas um *log* de registro das atividades, ou até relatórios com índices categorizados sobre processos ou produtos. Neste sentido, os dados contidos nos vocabulários controlados também são gerados por ferramentas de softwares e possuem um grande potencial de análise, assim como a avaliação da qualidade de dados e processos envolvidos na sua construção e manutenção.

Dentre os elementos que constituem um vocabulário controlado, os agrupamentos de tags e os termos se destacam, os quais podem ser identificados como tipos dados, codificados em modelos de dados como o RDF e SKOS. Independente da codificação, os dados de um vocabulário controlado podem ser combinados com outros atributos e gerar novos *insights,* por exemplo, utilizando o enriquecimento de dados para contextualizar informações, comparando padrões de buscas com listas de termos autorizados, etc. Esses *insights* podem desencadear políticas de gestão e preservação, governança e vinculação de dados, além de abrir possibilidades para distintos tipos de avaliações.

O acesso a um vocabulário pode ser realizado tanto de forma interna quanto externa. Tradicionalmente, vocabulários controlados são utilizados em bibliotecas por especialistas na padronização de termos, para efetuar buscas em suas bases de dados, caracterizando-se como um acesso interno. O acesso externo pode ser realizado via consulta à base de dados em nível de usuário e de desenvolvedor, desde que disponibilizado para este fim, por exemplo, utilizando uma *Application Programming Interface* (API), desde que exista uma cultura de compartilhamento ou vinculação de dados com outras instituições.

O acesso externo a um vocabulário controlado também pode ser realizado por meio de modelos de exportação de dados, como é possível verificar em muitos softwares que exportam dados em RDF ou SKOS, entre outros. Uma vez disponível, um vocabulário controlado pode ser reaproveitado, mas se não há essa possibilidade de exportação para determinados formatos já conhecidos é importante também deixar opções de integração desses vocabulários por meio de APIs, que são mecanismos de integração utilizados em vários sistemas da atualidade para acessar determinados tipos de dados ou *web services*.

[1] https://vocabularyserver.com/web/
[2] https://skosmos.org/
[3] https://vocbench.uniroma2.it/

Os vocabulários controlados despertam contínuo interesse devido a sua capacidade de representar conceitos por meio de termos. Ryan (2014) destaca que um vocabulário controlado é usado para fornecer consistência durante uma pesquisa, além de descrever coisas, lugares, formas, assuntos entre outros tipos de registros relevantes que são representados em termos únicos. O uso dos vocabulários controlados é disseminado com maior intensidade em ambientes acadêmicos, pois seu objetivo é a redução de ambiguidade entre os termos utilizados na pesquisa, ocorre que isso nem sempre significa maior precisão nos resultados da busca. O que permite questionar também a funcionalidade social deste instrumento terminológico de cunho científico, uma vez que um conjunto de termos técnicos de domínios do conhecimento podem não possuir uma sensibilidade popular.

Talvez seja oportuno citar a Socioterminologia como a adaptação de termos aos contextos em que são inseridos ou até mesmo a folksonomia que seria uma atribuição livre de termos sugerida pelos próprios usuários. Embora permaneça este questionamento da importância social, destaca-se que esses instrumentos de uso acadêmico são de grande importância para a sociedade, pois os vocabulários controlados são instrumentos terminológicos que padronizam a terminologia de um domínio do conhecimento. Os vocabulários controlados aumentam a precisão de recuperação da informação por reduzirem a sinonímia e a polissemia de palavras que permeiam a comunicação científica e não-científica.

As ciências avançam porque padronizam os seus processos e métodos, os vocabulários controlados são resultados dessa padronização conceitual e terminológica. Logo, o avanço das ciências possuem uma importância social, tecnológica e cultural. Neste sentido, avaliar um vocabulário possui certa relevância, por isso é indispensável questionar o uso de técnicas e padrões adotados para esse tipo de instrumento de representação terminológica no âmbito da Organização do Conhecimento e da Ciência da Informação .

Mas se há algum mérito ou relevância para os vocabulários controlados é preciso investigar seus processos e estruturas, tomando como base os fundamentos da Terminologia como acolhedora dos estudos de termos especializados. As ciências revisam seus fundamentos para validar, refutar ou expandir seus pressupostos e com a Ciência da Informação não pode ser diferente, tendo em vista que ela também deve se manter à prova, como tantas outras e dialogar em um sentido contributivo com outras ciências, intercalando técnicas e métodos que possam apoiar essa relação interdisciplinar.

Muitos sistemas de buscas na internet em sua maioria utilizam uma opção mais aberta de busca, porque, digitar uma palavra errada não é um grande problema, devido às alternativas e sugestões de resultados. Por isso é comum obter um resultado parecido com o termo que poderia ser utilizado corretamente, ou seja, palavras com erros ortográficos ou incompletas que são empregadas como sinônimos ou termos relacionados também podem ser consideradas relevantes dependendo do contexto de pesquisa.

Grande parte dos mecanismos de buscas utilizam algum sistema de inteligência, como o *Bidirectional Encoder Representations from Transformers* (BERT) projetado para pré-treinar representações bidirecionais profundas de texto não rotulado, processamento da linguagem natural e inferências de linguagem. (Devlin; Chang; Lee; Toutanova, 2018). Tais sistemas de inteligência garantem certa precisão na hora de efetuar uma busca, considerando erros e palavras com algum grau de similaridade entre os termos de buscas.

Embora os vocabulários controlados tenham essa característica de redução de ambiguidades, este controle terminológico é feito por um profissional especializado. O processamento automatizado de termos na classe dos milhares por segundo é distinto de um processamento manual, que pode considerar fatores subjetivos para elencar um candidato a termo. Neste sentido, os princípios de avaliação apresentados abrangem a identificação de dados estatísticos que representam o design de um vocabulário assim como possíveis erros que podem comprometer um vocabulário controlado e métricas baseadas em Processamento da Linguagem Natural (PLN) em inglês *Natural Language Processing* (NLP). Também é necessário destacar que há uma grande necessidade de explorar métodos que possam abranger as diferentes nuances do processo de avaliação.

Por isso é necessário ampliar o uso da tecnologia em atividades complexas, pois o tempo de processamento de uma informação está avançando cada vez mais. Logo, a tecnologia é uma aliada útil para otimizar tarefas que demandam muito tempo de especialistas. Os vocabulários controlados precisam também adotar novas formas de aliar suas propriedades para mecanismos de busca que são nativos da web. Por isso a necessidade de se utilizar cada vez mais modelos de dados com viés semânticos ou padrões integrados à web para que esses vocabulários possam ser vistos pela sua natureza qualitativa, quer seja, a precisão, redução de ambiguidade, representação adequada dos termos ou conceitos, entre outros.

O viés normativo é outra perspectiva que precisa ser considerada no processo de avaliação, porque ela baliza um nível mínimo de definições que podem ser aderidas. Apesar do uso de normas como a Z.39.19 da *National Information Standards Organization* (NISO) existem outras normas que foram elaboradas pela *International Organization for Standardization* (ISO) que reforçam o apoio teórico normativo desta pesquisa no âmbito da Terminologia e da Ciência da Informação. No Brasil instituições como Associação Brasileira de Normas Técnicas (ABNT) enfatizam a importância das normas ISO, tendo em vista que algumas normas brasileiras são adaptações de normas ISO.

Com efeito, além da garantia normativa e teórica desta pesquisa, o uso de uma linguagem de programação para realização das análises dos vocabulários controlados é uma etapa de conhecimentos técnicos e científicos explorados a partir de Bird; Klein e Loper (2009) que explanam sobre *Natural Language Processing* (NLP) utilizando a biblioteca NLTK em *Python*, outras bibliotecas também foram exploradas, como *rdfLib*. Para o ambiente de desenvolvimento e execução das análises foi utilizado o *Jupyter Notebook*, versão online disponibilizada pelo *Google Cloud.* E para execuções locais utilizou-se a distribuição de pacotes Anaconda[4], por ser reconhecida na computação científica e ter o código fonte aberto.

A avaliação dos vocabulários foi realizada por meio de três análises distintas que se mostraram viáveis ao longo da pesquisa. A primeira se chama análise estrutural quantitativa e se baseia na contagem de tags. A segunda chama-se análise baseada em modelos de dados, cujo propósito é identificar erros em vocabulários publicados em SKOS e RDF, para isso foi utilizado como referência o *qSKOS*, uma ferramenta desenvolvida em linguagem Java por Christian Mader da Universidade de Viena (Mader; Haslhofer; Isaac, 2012), pois tal ferramenta permite identificar problemas de qualidade associados à etiquetagem dos dados. E por fim, a análise métrica de termos é realizada por meio de recursos disponíveis na biblioteca NLTK, extraindo métricas com base na análise dos termos extraídos dos vocabulários controlados.

Portanto, enfatiza-se que esta pesquisa não abrangerá todo o escopo envolvido na caracterização de um vocabulário controlado, mas pretende apresentar métodos objetivos e

[4] https://www.anaconda.com/

práticos, para a avaliação de vocabulários controlados, por meio de análises quantitativas, estruturais e métricas de termos, com o auxílio de programação em *Python*.

1.1 Hipótese

A hipótese defendida nesta pesquisa parte do pressuposto de que há poucos estudos práticos no Brasil, que envolvem o processamento automático de vocabulários controlados em RDF e SKOS, com auxílio de ferramentas de softwares ou linguagens de programação, demonstrando de forma prática como o processo de avaliação pode ser conduzido.

1.2 Questão da pesquisa

A questão levantada nesta pesquisa refere-se a forma de implementação do processo de avaliação de um vocabulário controlado em ambientes digitais, especificamente aqueles que estão codificados em RDF e SKOS.

1.3 Justificativa

Os desdobramentos da pesquisa realizada durante o mestrado no Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de São Carlos, sobre a avaliação do padrão SKOS para representação de vocabulários controlados, financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo. Despertou inquietações sobre o estado da arte dos vocabulários controlados no âmbito da Ciência da Informação, no que diz respeito aos métodos e instrumentos adotados para avaliar vocabulários controlados. Os desdobramentos da pesquisa de mestrado consistiram em apresentar uma relação teórico-metodológica sobre o que é conceito e representação, analisar e descrever a composição do SKOS e identificar na literatura científica, diretrizes para avaliação de vocabulários em SKOS.

Em meio a esta inquietação e observando as diretrizes para avaliação de KOS identificadas por Ramalho e Sousa (2019) aplicáveis aos vocabulários codificados em SKOS, percebeu-se a necessidade de expandir este processo de avaliação, tendo em vista que muitos vocabulários não estão disponíveis apenas em SKOS, mas também em RDF. Por isso, entende-se a necessidade de ampliar estas diretrizes e contribuir com princípios de avaliação que podem ser aplicados em diferentes tipos de vocabulários controlados disponíveis em RDF

e SKOS. Neste ensejo, constata-se a também possibilidade de realizar análises, que incluem o Processamento da Linguagem Natural.

Outro fato, deve-se a leitura dos artigos de Mader; Haslhofer; Isaac, (2012) e Pastor-Sanchez;Martinez-Mendez, e Rodriguez-Munoz (2012), pois notava-se uma certa complexidade nas análises de arquivos RDF e SKOS e dificuldade de entendimento sobre a forma como foram extraídas tais informações de dados que até então pareciam inertes, tal complexidade impulsionou a motivação desta pesquisa para compreender e desenvolver formas próprias de análises que pudessem ser compreendidas e analisadas em território nacional. Não por acaso, esta pesquisa busca viabilizar a compreensão desses tipos de análises e alinhar a pesquisa nacional aos modelos desenvolvidos pela comunidade científica internacional no âmbito da Organização do Conhecimento.

1.4 Objetivos

1.4.1 Objetivo geral

O objetivo desta pesquisa é discutir sobre avaliação de vocabulários controlados, com base em três tipos de análise: estrutural quantitativa, baseada em modelos de dados e métrica de termos.

1.4.2 Objetivos específicos

Para atingir o objetivo geral proposto, colocam-se como objetivos específicos:

- selecionar vocabulários controlados do tipo tesauros, que estão indexados na Basic Register of Thesauri, Ontologies & Classifications (BARTOC) e disponíveis em RDF ou SKOS;
- analisar os vocabulários selecionados com base na contagem de tags, nos problemas associados ao tagueamento e recursos de NLP;
- extrair métricas de caracterização dos vocabulários controlados por tipo de análise;
- apresentar uma visão geral de avaliação dos vocabulários controlados.

1.5 Métodos e Procedimentos

Esta é uma pesquisa exploratória porque segundo Selltiz; Wrightsman; Cook (1965), (Gil, 2008) e (Costa; Costa, 2001), porque objetiva adquirir maior familiaridade com o fenômeno pesquisado permitindo a formulação mais precisa de problemas, criação de novas hipóteses e realização de pesquisas mais estruturadas. Caracteriza-se, também, como pesquisa de natureza aplicada, uma vez que "objetiva gerar conhecimentos para aplicação prática, dirigidos à solução de problemas específicos" (Silveira, Córdova, 2009, p.35).

De modo geral, este trabalho apresenta procedimentos que podem ser implementados em um processo de avaliação de vocabulários controlados, com ênfase na análise estrutural quantitativa, baseada em modelos de dados e métrica de termos em ambientes digitais.

As etapas desta pesquisa consistiram em realizar uma pesquisa bibliográfica utilizando o Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). As bases selecionadas foram *Directory of Open Access Journals* (DOAJ), *Scientific Electronic Library Online* (SCIELO), *Science Direct* e SCOPUS da *Elsevier, PubMed, Gale Academic One File*, *Springerlink* da *Nature* e PROQUEST antes findar o contrato com a CAPES. A pesquisa bibliográfica também foi realizada na Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI), na Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), no Banco de Dados Bibliográficos da USP - Dedalus, no Portal de Busca Integrada da USP e no *Google Scholar*.

Os termos utilizados na pesquisa foram selecionados de acordo com a abrangência temática da pesquisa que inclui a terminologia, terminografia, web semântica, métodos de avaliação, vocabulário controlado, tesauro e Processamento da Linguagem Natural. As palavras-chave utilizadas nas buscas foram expressas em português e inglês e de acordo com os temas macro da pesquisa, que incluem a terminologia e a avaliação de vocabulário controlado. Os termos de busca empregados resumem-se em: avaliação de vocabulários, assessment of vocabulary, assessment of vocabularies controlled, evaluate of vocabularies controlled, evaluate of vocabulary, análise de vocabulário e analysis of vocabulary. Outros termos de buscas utilizados foram terminologia, *terminology*, terminografia, *terminography*, lexicologia, *lexicology*, lexicografia e *lexicography,* Processamento da Linguagem Natural e *Natural Processing Language*, aplicando o mesmo modelo de busca anteriormente citado.

Utilizando o modo busca avançada com o uso do operador booleano OR utilizou-se respectivamente todos os termos supracitados, aplicados a todos os idiomas, sem restrição de ano, tipo de documento ou base de dados. Posteriormente foram aplicados os filtros de buscas para somente artigos e revisados por pares.

Por conseguinte foram aplicados os filtros de busca em cada base supracitada que está contida no portal de periódicos da CAPES. E este mesmo método foi aplicado também individualmente a cada base de dados exceto *PubMed*, justificado mais adiante. Durante a pesquisa o Portal de Periódicos CAPES estava sofrendo muita instabilidade o que dificultou o acesso a muitos artigos via Portal. Por isso também foi realizado o acesso às bases de dados via VPN da USPnet para ter acesso institucional às publicações científicas.

Na Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação (Brapci) e na Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) foram utilizados apenas termos em português e no Google Scholar português e inglês, com busca aberta em qualquer idioma aplicando os mesmos termos de busca. Foi possível recuperar documentos em outros idiomas além daqueles utilizados na busca e em diversos formatos, como livros e relatórios.

Na etapa de seleção do referencial teórico foi possível identificar algumas peculiaridades temáticas sobre vocabulários controlados. Por exemplo, o resultado mais comum para o termo vocabulário resulta em publicações na área de Ciências da Saúde aplicada a fonoaudiologia e desenvolvimento cognitivo, por isso não foi realizada nenhuma busca exaustiva na PubMed, porque foram utilizados apenas os resultados oferecidos via Portal de Periódicos CAPES que porventura estivessem indexados nessa base de dados. Outra característica comum é a replicação de publicações em bases distintas, o que permitiu realizar o download do mesmo item mais de uma vez. Além de identificar itens publicados simultaneamente em bases privadas e públicas, por meio do *Google Scholar.* Por esse motivo, um dos critérios de seleção adotados foi utilizar publicações acessíveis por meio de acesso institucional.

Também foram realizadas buscas separadas pelas normas no portal da National Information Standards Organization (NISO) para verificar e analisar a versão mais recente da norma ANSI/NISO Z39, assim como da norma ISO 25964 no portal da International Organization for Standardization (ISO) e das recomendações da W3C para SKOS Simple

Knowledge Organization System Primer (Isaac e Summers, 2009), RDF 1.1 Primer (Manola, Miller e McBride, 2014), RDF 1.1 Concepts and Abstract Syntax (Cyganiak, Wood e Lanthaler, 2014), RDF Schema 1.1 (Brickley e Guha, 2014), RDF 1.1 XML Syntax (Gandon, Schreiber, 2014), RDF 1.1 Turtle (Prud'hommeaux e Carothers, 2014) e SPARQL 1.1 Query Language (Harris e Seaborne, 2013) para auxílio do escopo teórico e prático.

Os textos selecionados foram armazenados e organizados por meio do Mendeley Reference Manager for Desktop. Após o processo de seleção e leitura dos textos foi baseada na análise de conteúdo Bardin (1977) considerando as fases de pré-análise e exploração do material, fazendo uso da leitura flutuante, escolha dos documentos e da regra de pertinência. Para redação de escopo referencial do teórico e da seleção de vocabulários controlados que possuem registros na literatura. Por isso, a seleção dos vocabulários controlados avaliados também foi definida de acordo com parâmetros identificados na literatura científica, normas e recomendações técnicas, nesta etapa foram desenvolvidos os primeiros experimentos práticos baseados na utilização de softwares, ferramentas de programação e nos modelos de dados em que estão disponibilizados, favorecendo os vocabulários passíveis de análise.

Nesta pesquisa, a *Basic Register of Thesauri, Ontologies & Classifications* (BARTOC) foi utilizada como referência para a seleção dos vocabulários controlados, pelo fato de ser um grande repositório aberto com dados indexados sobre vocabulários controlados. Até a presente pesquisa a BARTOC possui 3.427 registros de vocabulários controlados de diversos tipos, entre dicionários, listas de assunto, taxonomias, tesauros e ontologias. Deste total, 196 registros são de vocabulários diversos (glossários, esquemas de classificação, dicionários, etc.) em língua portuguesa, que incluem Brasil e outros países, como Portugal. Quando aplicado o filtro apenas para tesauros, esse número reduz para 46 registros.

Utilizando esse modelo de busca somente para tesauros, acrescentando apenas o idioma espanhol os resultados sobem para 181 tesauros e ao acrescentar a língua inglesa, os três idiomas resultam em 608 resultados. Devido a diversidade idiomática foram considerados apenas os tesauros em português (Brasil e Portugal) na tentativa de identificar e analisar tesauros brasileiros ou que tivesse alguma relação com a língua portuguesa. Tendo em vista a importância local de estudos na área da Terminologia e da Organização do Conhecimento, com intuito de fortalecer e fomentar pesquisas futuras nacionais.

Por isso, esta pesquisa considerou a princípio, somente os 46 registros apresentados, no entanto, foi necessário verificar quais vocabulários controlados estavam disponíveis em RDF ou SKOS dentre esses registros, por isso foi necessário verificar um a um e conferir quais estavam acessíveis e nesses dois formatos, o que é explicado posteriormente na análise preliminar. E depois de identificar somente os vocabulários controlados em RDF e SKOS foi necessário verificar quais poderiam ser analisados integralmente, por isso dentre os 46 resultados somente 10 vocabulários controlados foram selecionados e analisados.

A proposta desta pesquisa de avaliar os vocabulários controlados com ênfase em três modalidades, uma análise estrutural quantitativa, análise baseada em modelos de dados e uma análise métrica de termos desdobra-se da seguinte maneira: a análise estrutural quantitativa corresponde a contagem de *tags* incorporadas no vocabulário controlado de acordo com a sua função, por exemplo, tags de documentação, etiquetagem, relacionamentos, entre outros; a análise baseada em modelos de dados consiste em identificar problemas relacionados à estruturação dos dados em RDF ou SKOS, pois muitos vocabulários controlados foram publicados nestes modelos de dados. Por fim, a análise métrica de termos compreende o uso de recursos computacionais para extrair métricas a partir dos termos alocados nos vocabulários selecionados, com auxílio de técnicas e ferramentas de NLP.

As análises e revisão teórico-metodológica, realizadas para sistematizar as informações advindas desta investigação, assim como os resultados obtidos a partir da revisão bibliográfica realizada, foi organizada seguindo as investigações no âmbito da Terminologia e da Organização do Conhecimento, como a base teórica da Ciência da Informação no que compete os estudos sobre vocabulários controlados. Favorecendo assim, criação de subsídios para um melhor entendimento sobre a avaliação e utilização dos padrões que abrangem os vocabulários controlados.

1.6 Estrutura do trabalho

Esta tese foi dividida em seis seções, como apresentado a seguir:

Na seção 1 é apresentada a introdução da pesquisa quantitativa por meio da contextualização do tema, percorrendo sobre o problema e motivações da pesquisa, os objetivos e os procedimentos metodológicos.

Na seção 2 discute-se a compreensão da Terminologia, as principais normas utilizadas no desenvolvimento de vocabulários controlados e a caracterização dos vocabulários controlados e de um tesauro.

Na seção 3 apresenta algumas considerações atribuídas ao Processamento da Linguagem Natural, posteriormente descreve as principais características dos modelos de dados RDF e SKOS por meio de suas propriedades, Em seguida é apresentada uma forma de análise de termos estabelecida na década de 90 intitulada Onometria.

Na seção 4 é apresentado o processo de avaliação de vocabulários controlados considerando os principais registros na literatura científica sobre tipos de avaliação congêneres e uma análise preliminar que permitisse selecionar integralmente os vocabulários controlados, por meio da análise estrutural quantitativa, baseada em modelos de dados e métrica de termos.

Na seção 5 sintetiza os principais resultados obtidos em cada tipo de análise por meio de uma comparação geral, assim como as premissas que podem ser adotadas no processo de avaliação.

Na seção 6 conclui este trabalho, oferecendo uma visão ampla da pesquisa, dificuldades do processo e perspectivas de estudos futuros.

2 TERMINOLOGIA, NORMAS E VOCABULÁRIOS CONTROLADOS

A Terminologia não é uma ciência nova, embora sua teoria tenha surgido por volta de 1931, com a tese de doutorado de Eugen Wüster (1898-1977). Para ratificar esta afirmação é necessário revisitar algumas publicações que antecedem esta teoria. Não há uma intenção de discutir a etimologia da palavra, mas as implicações históricas que contextualizam o surgimento, o compromisso e a importância da Terminologia registrada em breves momentos históricos.

O surgimento interdisciplinar da Terminologia abre espaço para o diálogo com outras áreas, por meio de instrumentos, teorias e métodos que auxiliam na compreensão do termo e dos arquétipos conceituais que influenciam a elaboração dos instrumentos de controle terminológico. Por exemplo, o remoto surgimento da taxonomia, exercida pelos mais inveterados filósofos ocidentais.

A breve incursão teórica apresentada a seguir, permite identificar e compreender alguns registros históricos que antecedem a criação da Teoria Geral da Terminologia por meio de publicações congêneres relativas ao controle de termos no âmbito das ciências. Pois é notório que a padronização de termos possui um caráter científico, com interesse de nomear, classificar e organizar o avanço das descobertas científicas de modo gradual e constante.

Dentre os fragmentos históricos que atestam a importância da padronização terminológica, destaca-se um artigo do *The British Medical Journa*l de 1871 sobre *Pharmacopœial Nomenclature* afirmando que "[...] não há pouca dificuldade de adoção para os nomes químicos farmacêuticos, explícitos de fácil compreensão e inequívocos" (PHARMACOPŒIAL [...], 1871, p.373, tradução própria). Embora a padronização terminológica seja uma tendência nas áreas médicas e farmacêuticas, há registros na Botânica, Filosofia e Música.

Por exemplo, outro registro que enfatiza a necessidade da padronização terminológica foi publicado por na revista *Nature* em 1871, sobre a classificação das frutas, em resposta a um pesquisador chamado Dr. Dickson que criticou a "extensão desnecessária da terminologia existente e a variação dos termos empregados" (FRUIT [..], 1871, p.347, tradução própria). Kitchener (1872) também apresenta críticas sobre a terminologia botânica afirmando que é uma terminologia tão poliglota em sua derivação como a torre de Babel e muitas vezes envolvendo questionáveis hipóteses de função ou ideias incorretas de

morfologia. Kitchener (1872) estava preocupado sobre o ponto de vista didático, tendo em vista a dificuldade de apreensão dos termos por parte dos alunos.

Dyer (1872) em resposta a Kitchener (1872) que também publicou um artigo na *Nature* sobre a terminologia botânica afirmou que, quando os termos são concebidos pela primeira vez, eles refletem as idéias científicas da época. Mas à medida que o conhecimento progride, surgem novas visões, embora as coisas permaneçam as mesmas, as ideias sobre elas mudam, e os nomes outrora recebidos com um significado inteligível, se tornam arbitrários. Dyer (1872) defende principalmente que não se pode apenas descartar um termo ou seu antecessor, pois os registros de termos antecessores e predecessores também fazem parte da evolução terminológica da área.

Para além da Botânica, a terminologia também esteve presente nas discussões filosóficas, ou, pelo menos, na redação de textos filosóficos. Um texto publicado por Vere (1873) no qual faz algumas observações sobre a importância e a dificuldade de lidar com a linguagem popular de maneira estritamente filosófica. Onde também considera que o título de "Terminologia Filosófica", sob o qual o texto foi redigido limita seu conteúdo, por isso adiciona algumas outras considerações sobre o mesmo assunto, para que seus leitores encontrem nestas observações adicionais uma confirmação e um desenvolvimento adicional sobre a necessidade de uma linguagem filosófica mais copiosa. Isto é, a Terminologia precisa apresentar ferramentas de controle terminológico replicável e fácil de reproduzir.

Hullah (1874) também faz uma crítica quanto a terminologia utilizada na área musical. Expressando sua preocupação ao tratar de certos nomes ou termos e epítetos em uso entre os músicos ingleses. Para isto, propõe um entendimento quanto à nomenclatura musical. Um ponto perceptível e dedutível na crítica de Hullah (1874) refere-se aos termos não apenas como espelhos de definições abstratas, mas de funções, atributos de qualificação e classificação. Devido à complexidade da terminologia musical para tons, notas, escalas, intervalos, entre outros.

Lewis (1879) publica um texto sobre os males decorrentes do uso de nomes históricos nacionais como termos científicos, com isso revela uma preocupação da Antropologia na caracterização de povos e localizações geográficas. Por isso, propõe a criação de comitês nomeados pelas várias Sociedades Antropológicas da Europa para definir os principais tipos de habitantes de seus respectivos países e também a formação de uma Comissão Internacional para padronizar as definições e nomenclaturas, livres de objeções. Em

tese, os nomes atuais, na época de sua publicação eram imprecisos, e levavam a erros carregados com consequências graves para toda a raça humana.

O que Lewis (1879) propõe é que certos tipos abstratos fossem cuidadosamente considerados e definidos, sem levar em conta a nacionalidade, para evitar quaisquer questões de natureza política ou pessoal, substituindo os nomes por números ou letras. Ressalta ainda que quando isso fosse realizado, e não antes, teria o direito de considerar que os estudos antropológicos começaram a se cristalizar em uma ciência tão digna de nome como astronomia ou geologia.

Os registros científicos em diversas áreas do conhecimento apontam sobre a necessidade da padronização de termos científicos. Deste modo, percebe-se que a Terminologia surge por meio de uma demanda científica, para possibilitar a evolução das ciências. Quando Thomas Kuhn explicita em seu livro "A estrutura das revoluções científicas", sobre a mudança de paradigma é possível deduzir que a Terminologia surge decorrente de uma crise terminológica em diversas ciências. O surgimento científico da terminologia não se limitou a uma área específica, como se percebe por meios dos trabalhos dos autores (Kitchener, 1872; Dyer, 1872; Vere, 1873; Hullah, 1874; Lewis, 1879).

Lewis (1879) também reitera que é notável a preocupação com a nomenclatura baseada em ideologias de povos antigos, o que poderiam indicar uma suposta autoridade científica devido a sua localização geográfica ou dominação de outros povos. Antes da TGT houve uma preocupação em padronizar as terminologias das ciências. Depois que este objetivo foi alcançado é necessário tratar com acuidade esses termos e suas formas de representação, pois a linguagem é dinâmica.

Embora a padronização de termos tenha uma preocupação registrada no ocidente, esta é uma demanda que emergiu em vários países ao ponto do Ministério da Educação do governo da Índia publicar em 1968 um documento chamado "*Evolving a National Terminology*" recomendando que os termos científicos e técnicos adotados na Índia devem ser comuns, na medida do possível.

Uma afirmação defendida por Jawaharlal Nehru logo na apresentação do documento é que não deveria haver tal coisa como "a ciência indiana", assim como se faz com a tecnologia. A percepção ao olhar para essas questões de uma maneira estreita e nacionalista acaba levando ao estreitamento de sua ciência e ao estreitamento de sua tecnologia e de seu próprio trabalho (ÍNDIA, 1968). Nehru ressalta que esse negócio de

evoluir termos especiais que não são conhecidos do público em geral nem de qualquer outra pessoa no mundo realmente significa que você está se isolando da corrente geral do conhecimento e, ao mesmo tempo, dissociando-se de seu próprio povo que não entende seus termos técnicos. E assim você se converte em algo que ninguém entende e que ninguém se importa. (ÍNDIA, 1968).

A visão de Nehru sobre a ciência é uma visão globalizada, onde não há limites geográficos e linguísticos na padronização dos termos, por isso, afirma que a ciência e a tecnologia não conhecem fronteiras. Ninguém fala ou deveria falar sobre ciência inglesa, ciência francesa, ciência americana, ciência chinesa. A ciência é algo maior que os países, destaca também que existe uma certa vantagem em adotar o que é chamado de terminologia internacional ao lidar com termos científicos e técnicos. (ÍNDIA, 1968).

A visão de Nehru não é totalmente utópica, e os Descritores em Ciências da Saúde - DECs são a prova disso, assim como o Sistema Internacional de Unidades e Medidas e outros padrões internacionais que consolidam o avanço das ciências. Não há dúvidas que as pesquisas científicas na área da terminologia são contribuições diretas para as ciências, ainda que específicas e não populares. Gilreath (1992) ao explanar sobre a harmonização terminológica destaca que os conceitos não estão vinculados a idiomas específicos, por exemplo, o idioma é o inglês, mas os princípios se aplicam a qualquer idioma. O que ratifica a visão Nehru, quando enfatiza sobre o isolamento terminológico.

Percebe-se que a Terminologia surge a partir de uma demanda científica, logo o termo técnico-científico é o elemento central de estudo (Krieger; Finatto, 2004). Estes termos são atribuídos a determinados conceitos, o que significa que estes termos são selecionados com uma perspectiva onomasiológica, ou seja, do conceito para o termo. Gomes (2021) destaca o conceito como elemento comum entre a terminologia e a teoria do conceito, mas não se restringe a isso, tendo vista a sincrética relação com a Linguística e suas especificidades, a saber a Lexicologia, que se ocupa em estudar o repertório de palavras que compõem o léxico de uma língua e a Lexicografia que lida com os instrumentos produzidos com base na Lexicologia, como os dicionários de línguas.

Não por acaso, a Terminologia também engloba a Terminografia, que lida com um conjunto de termos. Esta subdivisão de especialidades entre Lexicologia e Terminologia possuem relações de similaridades e contrastes entre si conforme. Aubert (2001) afirma que as bases da terminologia consistem no ordenamento e transferência de conhecimentos, na

formulação e disseminação de informações especializadas, na transferência de textos científicos para outros idiomas e na armazenagem e recuperação de informação. Apesar de fundamentar-se em objetivos claramente definidos, a Terminologia possui uma estreita relação com a Lexicografia.

> A lexicografia considera as palavras enquanto parte do léxico, ou seja, como fazendo parte do conjunto de unidades de que uma determinada comunidade dispõe para se comunicar por intermédio da língua. Já a terminologia considera as palavras enquanto um conjunto delimitado por uma situação concreta de utilização (Aubert, 2001, p. 26).

O que pode ser claramente contrastante entre esses dois campos de estudo é que a Lexicografia se preocupa com a amplitude léxica das palavras, enquanto a terminologia foca na especificidade da palavra em relação ao seu conceito. Há diversas características que poderiam ser elencadas nesta relação de oposição e similaridade, como é possível observar (Aubert, 2001). A palavra terminologia é um termo polissêmico segundo Krieger e Finatto (2004), porém é possível inferir que esta palavra também pode ser um homônimo perfeito, porque possui uma grafia igual de si mesma com significados diferentes, o que eleva a excentricidade da Terminologia.

Uma atividade comum no âmbito da Ciência da Informação é o uso de diversas palavras que muitas vezes se refere ao mesmo objeto, porém em contextos diferentes, por exemplo, Linguagem Controlada, Linguagem Terminológica e Linguagem de indexação, que podem ser comumente encontradas em publicações que se referem ao processo de indexação de termos, o que pode ampliar as confusões teóricas que cerceiam as discussões sobre Terminologia. O que quer dizer que a Terminologia não está imune às ambiguidades e aliterações.

Compreender as teorias que compõem a gênese da Terminologia é um dos principais desafios para elaboração de técnicas de avaliação de instrumentos terminológicos para reduzir a possibilidade de apresentar métodos de análises descontextualizadas com o objeto de estudo. Por isso Aubert (2001) apresenta procedimentos de pesquisa que podem ser utilizados na Terminologia, o que chama a atenção nesse ponto, é o fato de vários instrumentos terminológicos possuírem métodos de elaboração, porém não se percebe comumente o uso de estratégias de avaliação dos mesmos. Obviamente não será possível obter uma compreensão total da Terminologia, tendo em vista o alcance das suas

ramificações, e porque tampouco foi realizada uma revisão sistemática das publicações de Wuster. Porém, existe uma possibilidade de ampliar estes estudos em pesquisas futuras, identificando documentos históricos e contribuições já realizadas que podem ter sido ofuscadas ao longo do tempo.

O surgimento da Teoria Geral da Terminologia (TGT) de Wuster deu origem a outras teorias complementares que tentaram suprir as lacunas da TGT. Especialistas como Teresa Cabré, cujo sobrenome foi alterado para Cabré Castellví, assumiram essa missão, ao propor a Teoria Comunicativa da Terminologia (TCT) como alternativa à TGT.

O reconhecimento da Terminologia enquanto uma ciência foi influenciada pela tese de doutorado de Eugen Wüster (1898-1977) em 1931, o qual, de acordo com Trojar (2017), Wüster é considerado o pai da terminologia moderna e responsável por preparar seis recomendações ISO e um padrão ISO por meio do comitê ISO/TC37. Wüster provocou uma calorosa discussão entre os linguistas, além de ser criticado sob a perspectiva onomasiológica[5], que particularmente pode ser entendida como um ramo da lexicografia. A lexicografia dedica-se principalmente ao estudo de dicionários, que podem ser classificados em diversos tipos, mas não se restringe somente a estes, de acordo com Biderman (1984) o léxico constitui um conjunto aberto em qualquer sistema linguístico e sua expansão é contínua.

Segundo Barros (2004) a terminologia tem como unidade padrão o termo, que é uma unidade lexical com um conteúdo específico dentro de um determinado domínio do conhecimento. Logo, um conjunto de termos pode ser chamado de terminologia. O vocábulo também é entendido como uma unidade lexical na visão de Barros (2004). A terminologia não se restringe somente a um conjunto de termos, ela pode ser concebida sobre três perspectivas teóricas segundo Cabré (1995), enquanto objeto, como disciplina e como prática.

A terminologia enquanto objeto pode ser observada do ponto de vista da linguística, da filosofia e por diferentes disciplinas técnico-científicas. De acordo com Cabré (1995, p.3) do ponto de vista linguístico “[...] os termos são o conjunto de signos linguísticos que constituem um subconjunto dentro do componente lexical da gramática do falante.” Para a filosofia a terminologia é um conjunto de unidades cognitivas que representam o

[5]Ramo da linguística que examina como uma determinada ideia ou conceito encontrou expressão na palavra e passou a ser representada por ela. Ver Babini (2006).

conhecimento especializado. Por fim, para diferentes disciplinas técnico-científicas, a terminologia é o conjunto de unidades de expressão e comunicação que permitem a transferência do pensamento especializado (Cabré, 1995).

A Terminologia como disciplina é concebida sobre três posições distintas, que defendem uma visão autônoma e autossuficiente e com fundamentos próprios cujos fundamentos são encontrados na TGT. A segunda perspectiva reconhece a terminologia como uma disciplina dependente ou parte de outra, podendo ser a Linguística ou Filosofia. Por fim, a terceira visão reconhece a autonomia da terminologia e sua natureza interdisciplinar (Cabré, 1995). A interdisciplinaridade da terminologia pode ser observada por meio do diálogo com diversas áreas do conhecimento.

Neste sentido, Cabré (1995) reconhece a natureza interdisciplinar da Terminologia, e enquanto disciplina, ressalta que possui as bases teóricas limitadas a um objeto de estudo definido, por isso, suas características são similares às outras disciplinas que possuem vertentes teóricas e práticas. Cabré (1995) ressalta que embora a Linguística contribua para a base conceitual da Terminologia é preciso destacar algumas diferenças, que se aplicam na concepção da linguagem, do objeto de estudo e nos objetivos teóricos-descritivos.

A Terminologia é uma disciplina aplicada e possui um ramo específico que trabalha com o desenvolvimento de dicionários especializados ou glossários terminológicos, chamado Terminografia, que aplica as teorizações da Terminologia (Bevilacqua; Finatto, 2006). De acordo com Cabré (1995) na prática da terminologia, o terminólogo segue um processo onomasiológico, diferente da lexicografia, cujo processo é semasiológico. A abordagem semasiológica parte do significante para o significado e opõe-se a onomasiologia que parte do conceito para a palavra (Babini, 2006; Couto, 2012).

Diferente de Cabré (1995) que considera a Terminologia por meio de três perspectivas, enquanto objeto, disciplina e prática. Aubert (2001) considera apenas duas, a terminologia objeto e como instrumento. Como objeto se refere ao conjunto de termos característicos de determinada área ou subárea. E como instrumento, um conjunto de pressupostos, métodos e representações que permitem a descrição das linguagens ditas de especialidade (Aubert, 2001).

A Terminologia trabalha com a normalização das unidades terminológicas e seus produtos não costumam ser polissêmicos (Cabré, 1995). A normalização assegura a qualidade e eficiência da comunicação especializada, principalmente no âmbito científico. Uma das grandes funções atribuídas à terminologia refere-se a representação do conhecimento especializado e sua transferência por meio da comunicação especializada (Cabré, 1999). A normalização possui uma relação intrínseca com a Terminologia, tendo em vista que a TGT de Wuster influenciou a criação de normas.

De acordo com Barros (2004) as funções da terminologia podem ser observadas por meio de três dimensões, a saber: metalinguística, comunicativa e política identitária. E essas dimensões se ramificam em três funções da terminologia enquanto disciplina. A função conceitual ou cognitiva, ligada à análise e descrição terminológica, a função comunicacional relacionada a comunicação e transferência de conhecimento científico/tecnológico e a função simbólica ou identitária que se refere a uma identidade nacional, regional ou de grupo (Barros, 2004).

A Terminologia ampliou sua extensão além do que Wuster propôs na TGT, que consistia em eliminar a ambiguidade por meio da normalização terminológica e estabelecer a terminologia como ciência para todos os fins práticos (Cabré Castellví[6], 2019). Quando questionada sobre o foco da terminologia Cabré (2005) ressalta que há divergências teóricas, para uns, o conceito concebido como universal é anterior ao termo, e para outros o termo é concebido como uma unidade de forma e conteúdo simultaneamente. Mas para responder esse questionamento destaca que o uso do conceito ou do termo é justificado pela adequação ao contexto e propósito que se pretende alcançar.

A Terminologia é uma ciência interdisciplinar que pode ter sido desenvolvida a partir de três grandes teorias, a mais remota seria a teoria do conhecimento que explica os conceitos, a teoria da comunicação, e a teoria da linguagem para incluir as unidades terminológicas (Cabré, 1999). A Terminologia também possui uma estreita relação com a teoria da classificação do conhecimento, devido a uma figura em comum, Conrad Gessner (1516 - 1565). De acordo com Barbosa (1969), Conrad Gessner oferece uma grande contribuição à história da classificação. Além disso, o naturalista suíço foi responsável por escrever a *Historiae Animalium*, a *Bibliotheca Universalis* e a *Historia Plantarum*.

[6] Cabré Castelví trata-se-se da mesma autora Cabré citada anteriormente, no entanto optou-se por citar o sobrenome conforme foi registrado na publicação consultada.

Barbosa (1969) ressalta que o sistema usado por Gessner é considerado como o primeiro esquema de classificação bibliográfica. Porém, a contribuição de Conrad Gessner não se limita à classificação, conforme explicita Nunes (2019) citando que Gesner é uma autoridade incontornável, pois foi uma figura chave na história natural renascentista, além de publicar em várias línguas (latim, grego, alemão e francês), que “por mais imperfeito que fosse” – escreve ele – mostrou muitos avanços na pesquisa de plantas.

Nunes (2019) ao citar a contribuição de Conrad Gessner no âmbito lexicográfico cita a dificuldade de tradução terminológica e desenvolvimento de vocabulários onomásticos a partir da obra de Michael Toxites, um médico, alquimista e poeta romano que viveu entre (1514-1581). Nunes (2019) ressalta que os vocabulários de Onomástica são "índices", um léxico destinado a médicos, farmacêuticos e acadêmicos de medicina que precisam conhecer os vários nomes de plantas, em que a prioridade são as línguas vulgares.

É muito comum encontrar publicações que enfatizam apenas a especialidade da Terminologia, desconsiderando o fato histórico que explicita sua origem a partir de uma crise na padronização de termos científicos. Não por acaso Zamorano Aguila (2013) sugere uma investigação epistemológica da Terminologia a partir da teoria do caos. Quando se fala de "caos" nas teorias do caos - o "caos determinístico", aquele que rege os sistemas dinâmicos não lineares. No conceito de ‘caos matemático’, refere-se ao ‘comportamento estocástico’ [aleatório] que ocorre num sistema determinístico, governado por leis exatas e imóveis (Zamorano Aguila, 2013).

Com base nesta premissa, Zamorano Aguila (2013) afirma que a partir de Wüster surgiu o desenvolvimento de modelos teóricos que não são onomasiológicos, mas semasiológicos, e não universalistas ou a-históricos. Ou seja, a teoria wüsteriana passou para um espaço caótico enriquecido pela presença de várias disciplinas, o que levou à auto-organização da terminologia para um programa de ciência interdisciplinar (portanto, autônomo, mas não auto suficiente, com base na TGT (Zamorano Aguila, 2013).

O que se percebe na perspectiva de Zamorano Aguila (2013) é que embora Wüster busque uma padronização teórica, os textos publicados em uma língua e mesmo os textos de especialidades possuem uma dinâmica própria, não previsível por normas ou teorias determinísticas. O que remete uma dualidade teórica que também foi enfrentada por Ludwig Wittgenstein, que em um primeiro momento interpreta os problemas da linguagem como um

problema de lógica e escreve *Tractatus Logico-Philosophicus,* depois repensa suas interpretações e publica a obra *Investigações Filosóficas* criticando a si mesmo e desconstruindo seu discurso.

Wittgenstein reconhece a complexidade da linguagem e variação contextual, e considera não apenas a lógica, mas a semântica. Outra dualidade comum, encontra-se no pensamento racionalista de René Descartes e o pragmatismo de Charles Sanders Peirce ao apresentar a semiótica como análise complexa de textos (Wittgenstein, 1968; Descartes, 1996; Peirce, 2015; Santaella, 2004). Não por acaso Wittgenstein foi influenciado por Immanuel Kant que por sua vez foi influenciado por Descartes. Logo o pensamento lógico da linguagem possui um sólido embasamento, porém não absoluto. Por isso, Zamorano Aguila (2013) afirma que este fato cria um espaço mais adequado do que o aparente determinismo de Wüster para explicar a criação de um texto especializado e assim analisar a natureza complexa, dinâmica e poliédrica dos termos.

Zamorano Aguila (2013) sugere que o texto especializado, como o texto geral, é uma rede de múltiplos elementos em contínua interação e também um sistema claramente aberto, dissipativo e adaptativo. Neste aspecto, a utilização de textos especializados sujeita-se tanto à uniformidade quanto à deformidade de uma língua, podendo inclusive sofrer reinterpretações em outro idioma distinto. Segundo Zamorano Aguila (2013) o texto especializado cumpre a dupla função de representar, cognitivamente, um setor da ciência, de objetos naturais ou culturais e transmitir o conhecimento.

Sabendo desta dupla função dos textos especializados e a da necessidade de estabelecer alguma coerência terminológica nos diversos campos científicos. A Terminologia surge da necessidade de denominar os sistemas de conceitos das diferentes disciplinas, com o objetivo de permitir uma comunicação eficiente entre especialistas. (Kobashi; Smit; Tálamo, 2001). Embora o caos esteja presente em várias disciplinas, ele não representa a totalidade de uma disciplina, por isso, os instrumentos terminológicos mesclam um estado lógico e semântico dos termos.

Dubuc (2002) afirma que a unidade terminológica é a denominação de uma noção única para o domínio de estudo, porque ela pertence exclusivamente a este domínio. Enquanto Barros (2006) defende que o conceito não é imutável e nem universal, pois eles sofrem mudanças diacrônicas, diastráticas, diafásicas. Embora pareçam afirmações distintas,

entende-se que o termo é uma unidade terminológica e representa um conceito, logo são afirmações intrinsecamente relacionadas. Ou seja, ao mudar um conceito, o termo também pode ser alterado, tendo em vista que não será representativo para o conceito modificado.

Porém, se o termo antigo não representa o novo conceito e há um novo termo e um novo conceito, então há uma espécie de evolução terminológica, o que relembra o posicionamento de Dyer (1872) em relação aos termos antecessores e sucessores de um domínio. Isto é, tal situação não é apenas substitutiva, devido à exclusividade ou mudança dos termos, mas um componente do ciclo terminológico, que não pode ser apenas descartado ou ignorado, por isso, a constituição de um instrumento terminológico vai além da seleção de termos gerais e específicos. Não por acaso, as normas técnicas padronizam o desenvolvimento de instrumentos de representação terminológica, estabelecendo procedimentos que podem ser adotados.

2.1 Normas para construção de vocabulários

A partir desta concepção entende-se que a Terminologia prescreve as discussões que sucedem a reflexão teórica até então apresentada e abre caminho para identificar algumas ramificações que sucederam sua origem. Dentre elas é importante destacar que o trabalho de Wüster influenciou as normas desenvolvidas pela *International Organization for Standardization* (ISO). Inclusive, as normas ISO 1087:1990 cuja versão mais recente é do ano de 2019, e a ISO/R 704:1968, que fornecem uma descrição sistemática dos conceitos relacionados aos trabalhos desenvolvidos no escopo da Terminologia.

Uma norma que também precisa ser destacada é a ISO 704: 2022 voltada para princípios e métodos dos trabalhos de terminologia. Esta norma indica os princípios e métodos que devem ser observados para a manipulação da informação, planejamento e tomada de decisões envolvidas na gestão terminológica. Há de se considerar que a terminologia em certo nível abstrato também lida com uma gestão de conceitos expressos por meio de termos.

Por isso as atividades da terminologia segundo a International Organization for Standardization (2022) consistem em identificar conceitos e suas relações, analisar e modelar sistemas conceituais, estabelecer representações de sistemas conceituais por meio de diagramas, definir conceitos; atribuir designações (predominantemente termos) a cada

conceito em uma ou mais línguas; registrar e apresentar dados terminológicos, ou seja, exercer a terminografia.

De acordo com a International Organization for Standardization (2022), para analisar terminologias, enquanto conjuntos de termos é necessário identificar o contexto ou campo de assunto, as propriedades atribuídas aos objetos no campo assunto. Além de determinar as propriedades que são abstraídas por meio de suas características, combinação das características para formar um conceito e atribuição de uma designação. De forma abstrata, a análise terminológica deve começar com os objetos em questão e o campo de assunto, contextualizando esses objetos (International Organization for Standardization, 2022).

Outros pontos identificados na ISO 704: 2022 referem-se às definições de conceitos gerais e individuais, intensão e extensão e tipos de relações entre conceitos. O que remete à Teoria do conceito de Dahlberg (1978) e a relação intrínseca entre a Terminologia e os SOCs, assim como os vocabulários controlados e os tesauros. A Terminologia é o elo teórico que liga o conceito ao termo e subsidia a prática dos instrumentos de controle terminológico, principalmente por meio de normas e convenções técnicas.

Existem outras normas associadas a Terminologia, dentre elas podemos citar a ISO 10241-1:2011 que define as entradas terminológicas em normas e a ISO 860:2007, cujo objetivo é a harmonização de conceitos e termos dos trabalhos de terminologia. Neste trabalho é necessário explorar ainda normas específicas que tratam da elaboração do principal instrumento de representação aqui analisado, o tesauro, enquanto tipo de vocabulário controlado.

A primeira norma a ser explorada é a norma ANSI/NISO Z39.19 revisada em 2010, que estabelece as diretrizes para a construção, formato e gerenciamento de vocabulários controlados monolíngues. Embora o alcance desta norma seja para diversos tipos de vocabulários, ela não abrange alguns tipos de SOCs, como ontologias ou redes semânticas. Por isso os vocabulários controlados amparados por esta norma são as listas de termos controlados, anéis de sinônimos, taxonomias e tesauros.

Esta norma se divide em onze partes por isso destacamos apenas alguns aspectos relevantes para esta pesquisa, a primeira introduz sobre a necessidade do vocabulário controlado e de que forma se alcança o controle vocabular, a segunda parte trata sobre o

âmbito de cobertura da norma, por isso inclui diretrizes e convenções que vão desde o conteúdo, exibição, construção, teste, manutenção e gerenciamento de vocabulários controlados (National Information Standards Organization, 2010).

Um aspecto importante que chama a atenção no início da norma refere-se à manutenção, tendo em vista que a linguagem se transforma e os vocabulários mudam podendo se tornar relevantes ou obsoletos. No tópico sobre testes e avaliação destaca-se a necessidade de realizar testes periódicos que permitirão mensurar algum nível de qualidade dos vocabulários controlados.

A partir do quinto item da norma inicia-se os principais esclarecimentos sobre os vocabulários controlados. O primeiro é que eles atendem a cinco propósitos, que são: O primeiro é tradução, ou seja, uma conversão da linguagem natural, o segundo é a consistência, para uniformizar os termos, o terceiro é a indicação de relacionamentos entre os termos. A quarta é a etiquetagem e navegação por meio de hierarquias e por último a recuperação.

Também existem quatro princípios importantes que orientam o desenvolvimento dos vocabulários controlados e consistem em eliminar a ambiguidade, controlar sinônimos, estabelecer relações entre termos e realizar testes e validação dos termos. Com base nesses quatro princípios é possível deduzir que também podem ser aplicados aos tesauros, tendo em vista a abrangência da ANSI/NISO Z39.19 de 2010.

Os metadados são descrições de dados criados a partir de um dicionário digital de dados (Souza, Catarino; Santos, 1997). Os metadados também são explorados no âmbito dos vocabulários controlados e podem ser utilizados segundo a National Information Standards Organization (2010) usando um vocabulário controlado como fonte de termos, usando os metadados como descrição para descoberta de recursos. E esquemas de metadados para representar todo o conteúdo do vocabulário.

O desenvolvimento de um vocabulário controlado é caracterizado por várias etapas que envolve desde a seleção de termos, que abrangem o domínio do vocabulário, especificidade, granularidade e relacionamentos com outros vocabulários. Além do uso de notas de escopo que fornecem mais informações sobre um determinado termo, também existem termos simples e compostos que precisam ser bem definidos como tais.

Os relacionamentos são de equivalência e hierarquia podendo ser genérica e de instância, assim como de associação, quando não são equivalentes nem hierárquicos. A

National Information Standards Organization (2010) apresenta uma tabela comparativa entre os distintos tipos vocabulários, a partir desta tabela é possível elencar alguns elementos básicos que compõem um tesauro. No quadro abaixo é apresentado algumas dessas propriedades básicas.

Quadro 01 – Propriedades básicas dos tesauros

CLASSE	PROPRIEDADE
TERMOS	Termos preferidos
	Termos de entrada
RELACIONAMENTOS	Relacionamento de equivalência
	Relacionamento hierárquico
	Parte/Inteira
	IsA
	HasA
	Termos Relacionados

Fonte: Adaptado de National Information Standards Organization (2010)

Percebe-se que um tesauro precisa definir seus termos preferidos e os distintos tipos de relacionamentos, observando que os termos relacionados devem possuir ligações recíprocas. As propriedades de Parte/ Inteira, *IsA* e *HasA* indicam tipos de relacionamentos que no caso de Parte/ Inteira abrange situações nas quais um conceito é inerentemente incluído em outro. Enquanto *IsA* e *HasA* indicam a relação genérica, de instância ou partitivas e podem ser reduzidas aos termos genéricos ou específicos.

Com base nestas informações é possível sintetizar que um tesauro precisa apresentar as seguintes características: em relação aos termos preferidos podem ser classificados em genéricos, específicos e relacionados, quanto ao tipo de relacionamento pode ser hierárquico, de equivalência ou partitivo. Devido também às características dos tipos de termos, os relacionamentos também podem ser genéricos e específicos, invocando propriedades hierárquicas e recíprocas, quando são relacionados entre si. Para melhor compreensão se apresenta o quadro a seguir com as siglas e propriedades que podem ser utilizadas nos tesauros de acordo com a National Information Standards Organization (2010).

Quadro 02 – Terminologia adaptada segundo a ANSI/NISO Z.39.19

ABREVIAÇÃO/ SIGLA	DEFINIÇÃO/TRADUÇÃO
TG	Termo Genérico
TGI	Termo Genérico (instância)
TGP	Termo Genérico (partitivo)

TE	Termo Específico
TEI	Termo Específico (instância)
TEP	Termo Específico (partitivo)
TR	Termo relacionado
NE	Nota de Escopo
U	USAR
UF	USADO PARA
NH	Nota de História

Fonte: adaptado de National Information Standards Organization (2010)

Uma vez definida a sintaxe terminológica que será adotada nos vocabulários controlados, outros aspectos entram em cena, principalmente no âmbito da gestão e apresentação. Por isso, a partir da National Information Standards Organization (2010) é possível elencar as principais diretrizes que definem o desenvolvimento dos tesauros, entre eles a apresentação do ponto de vista tipográficos, regras de arquivamento, entre outros aspectos, dos quais acrescenta-se que a prototipagem de um vocabulário pode ser um recurso útil.

A National Information Standards Organization (2010) também se refere ao tipo de exibição considerado formas simples e complexas, todavia podemos enfatizar que há uma tendência dos sistemas se tornarem cada vez mais responsivos, intuitivos e acessíveis. O formato é outra preocupação desta norma, ainda que apresente o formato impresso, o formato eletrônico também é enfatizado por meio do uso de hiperlinks. Devido ao ano de publicação desta norma muitos outros recursos não previstos podem ser acrescentados, principalmente devido à variedade formatos disponíveis no âmbito digital.

A documentação é outra diretriz básica da National Information Standards Organization (2010) e afirma que todos os vocabulários controlados devem fornecer documentação para o usuário indicando a forma de uso do vocabulário. Este ponto não se restringe apenas ao usuário, mas aos desenvolvedores de sistemas congêneres, especialistas e pesquisadores, que muitas vezes não obtêm informações suficientes para promover mudanças em determinados sistemas.

Outra recomendação desta norma é que todos os vocabulários controlados devem incluir na documentação de suporte o objetivo do vocabulário controlado, o escopo, convenções e abreviações e regras adotadas, se o vocabulário controlado está em conformidade com algum padrão nacional ou internacional, regras de arquivamento, número total de termos, data da última atualização, declaração sobre a política de atualização,

informações de contato e quaisquer informações adicionais necessárias para navegação ou utilização de recursos adicionais.

A pesquisa é um recurso que também precisa ser adotado nos vocabulários controlados, principalmente em sistemas eletrônicos, para que isto ocorra é necessário fornecer recursos de pesquisa e navegação, como pesquisa pelo termo completo ou um truncamento do termo, pesquisa por termos não preferenciais para recuperar termos preferenciais relacionados, pesquisa sem distinção entre maiúsculas e minúsculas, navegação em telas hierárquicas e alfabéticas, exibição de termos no contexto de seus relacionamentos e seu registro completo.

A princípio os recursos de pesquisa parecem demasiados para os dias atuais, tendo em vista que as opções de pesquisa simples e avançada se tornaram muito mais populares. Por meio da adição de operadores booleanos, o que também poderia ser aplicado neste caso, adicionando ou excluindo tipos de termos ou relacionamentos ao efetuar uma busca.

Por fim, uma diretriz de suma importância para gestores de vocabulários controlados são os relatórios, porque os relatórios fornecem informações necessárias para otimização dos sistemas de que gerenciam vocabulários controlados eletronicamente. Diferente da versão impressa, que precisaria de um sistema auxiliar para obter tais informações, mas independente do formato, os relatórios precisam ser claros e legíveis. A National Information Standards Organization (2010) já reconhecia que praticamente todos os vocabulários controlados desenvolvidos hoje são criados e mantidos usando algum tipo de software, por isso a implementação de campos de relatórios é plenamente viável.

Outra norma utilizada nesta pesquisa é a ISO 25964, uma norma que também aborda a interoperabilidade dos tesauros com outros tipos de vocabulários, dividida em duas partes e publicada respectivamente nos anos de 2011 e 2013. A norma ISO 25964 fornece recomendações para o desenvolvimento e manutenção de tesauros, entretanto estas partes possuem ênfases distintas embora possuam o mesmo objetivo. A parte 1 da norma, publicada em 2011, engloba os aspectos do desenvolvimento de tesauros, monolíngues e multilíngues, incentivando a interoperabilidade de rede, inclusive por meio de um modelo de dados e esquemas XML.

Diferente da norma ANSI/NISO Z39.19 que se restringe a vocabulários controlados monolíngues a norma ISO 25964-1:2011 também aborda sobre vocabulários controlados multilíngues, que podem surgir por meio de traduções para outros idiomas, por combinação de vários tesauros monolíngues e construção simultânea de várias versões linguísticas de um tesauro. Além desta diferença também aponta algumas diretrizes para os softwares de gerenciamento de vocabulários controlados, tendo em vista que muitos tesauros são elaborados a partir destes tipos de softwares.

Outra novidade que essa norma apresenta é a inclusão de padrões que são citados como formatos de intercâmbio e protocolos de troca de dados. Entre eles *Machine-Readable Cataloging* (MARC), *o Metadata Authority Description Schema* (MADS), o *Simple Knowledge Organization Systems* (SKOS), *Zthes*, o *Terminological Markup Framework* (TMF), XML, APIs, URIs, RDF e *SPARQL*, embora não forneça maiores informações.

A segunda parte da norma ISO 25964-2: 2013 explica como estabelecer mapeamento entre os conceitos de tipos distintos de vocabulários controlados, apresentando modelos estruturais para mapeamento entre vocabulários, tipos de mapeamento que podem ser por equivalência, hierárquico ou associativo, além de equivalência exata, inexata e parcial. Também discute sobre o uso de mapeamento na recuperação da informação. Esta segunda parte da norma, denominada ISO 25964-2:2013 apresenta formas de mapear vocabulários distintos como taxonomias e tesauros e integrá-los por meio de equivalências conceituais.

De certo modo, o mapeamento de vocabulários permite a reunião de um conglomerado de vocabulários controlados distintos interligados como uma espécie de vocabulários vinculados. O que remete a iniciativa do *Linked Open Vocabularies* (LOV) que reúne uma série de esquemas de dados e os mantém interligados. Em síntese a ISO 25964-2: 2013 recomenda tipos de mapeamento que podem ser estabelecidos entre diferentes tipos de vocabulários controlados.

A ISO 25964 é uma extensa norma aplicável a tesauros e outros tipos de vocabulário, entretanto a primeira parte da norma torna-se muito mais proveitosa para esta pesquisa, principalmente no âmbito da avaliação, desenvolvimento e gestão de vocabulários controlados que inevitavelmente dependem de algum tipo de infraestrutura tecnológica para ser gerenciada. Neste sentido, o uso de frameworks que suportam modelos de dados para sua representação é cada vez mais comum.

2.2 Vocabulários Controlados

Os vocabulários controlados são tipos de SOCs que podem ser considerados como agrupamentos de conceitos, tendo em vista sua composição por meio de termos representativos dos conceitos. Os vocabulários controlados são tipos de Sistemas de Organização do Conhecimentos (SOCs) em inglês *Knowledge Organization Systems* (KOS). Os SOCs não possuem uma definição generalista que dimensione seus limites de abrangência, por isso é mais comum encontrar definições particulares de cada sistema que são classificados como SOCs como se percebe pela definição Hodge (2000). E devido à ausência de uma definição mais precisa sobre o que seria um SOC foi aberta uma lacuna para uma série de termos congêneres que dificultam o seu entendimento conceitual. Lara (2015, p. 92) destaca que "[...] no Brasil, não há consenso sobre a utilização de um termo que abrange o conjunto de instrumentos de organização da informação e do conhecimento".

Outro ponto destacado é a diversidade de termos associados ao principal campo de estudos desses sistemas, que é a Organização do Conhecimento. Não por acaso, Barité (2011) ressalta sobre o uso indiscriminado dessas expressões, que não tem contribuído para unificar critérios sobre o assunto. Por isso, dentre as expressões citadas por Barité (2011), nesta pesquisa adota-se a expressão SOC para designar uma perspectiva macro conceitual.

A expressão vocabulário controlado justifica-se pela referência às normas que estabelecem os critérios para elaboração de tesauros, que é considerado um tipo de vocabulário controlado, do mesmo modo que e o vocabulário controlado é considerado um tipo de SOC, portanto o tesauro também é um tipo de SOC (Zeng, 2008; Souza, Tudhope, Almeida, 2012). Seguindo uma lógica alusiva à Teoria Geral de Sistemas (TGS) de Ludwig von Bertalanffy (Alves, 2012), quando afirma que existem sistemas dentro de outros sistemas, o tesauro se aplica a esta lógica. O tesauro é o principal instrumento de representação analisado nesta pesquisa, tendo em vista que as regras que orientam seu desenvolvimento são bem definidas.

Por isso, é recorrente encontrar documentos de referência como a norma ANSI/NISO Z39.19 de 2010, utilizada como referência para elaboração de tesauros. Assim como publicações sobre metodologias que podem ser adotadas, por exemplo (Campos; Gomes, 2006), além de publicações da década de 90 como (Gomes, 1990) que orienta sobre a elaboração de tesauros monolíngues. Os tesauros estruturam seus termos de forma relacional,

compreendendo três tipos de relações: hierarquia, equivalência e correlação (Tálamo; Lara; Kobashi, 1992). Estas relações entre termos reafirmam o alinhamento dos vocabulários à teoria do conceito que possui relações semelhantes.

Dentre as propriedades dos conceitos que possuem grande influência nos vocabulários controlados destacam-se as relações estabelecidas entre eles. Os conceitos possuem relações lógicas, relações hierárquicas, relações partitivas, relações de oposição e relações funcionais (Dahlberg, 1978). Os conceitos são arquétipos para a construção de vocabulários controlados estabelecendo regras fixas (princípios lógicos) e regras dinâmicas (princípios semânticos). Neste sentido, a concepção do conceito é um aporte teórico necessário para uma compreensão ampla e dinâmica dos vocabulários controlados. Os conceitos possuem algumas características que os definem, Lemos (1986) destaca que são representados por uma palavra ou um conjunto de palavras (termos compostos) e apresentam-se em geral no singular.

Dahlberg (1978) destaca algumas considerações sobre os conceitos, entre elas a importância das linguagens naturais para a formulação de enunciados sobre conceitos. Assim como sobre suas características, pois isso facilita a determinação do número de funções exercidas, que consistem em ordenar e classificar os conceitos e os respectivos índices, definir os conceitos e formar os nomes dos conceitos (Dahlberg, 1978).

Campos e Gomes (2006) descrevem que dois princípios podem se aplicar aos vocabulários controlados, o primeiro consiste no estabelecimento do termo/conceito, e o segundo no estabelecimento das relações entre eles. Por isso, as relações entre os conceitos estão intrinsecamente ligadas às relações estabelecidas entre os termos em um vocabulário controlado. Por exemplo, a relação hierárquica expressa em taxonomias.

Um ponto de flexão da teoria do conceito refere-se ao estabelecimento de categorias ao determinar um conceito, pois elas fornecem princípios para estruturar todas as classes de conceitos de um domínio (Campos; Gomes, 2006). Deste modo, percebe-se a inclusão de classes ou facetas para agrupar conceitos, e as relações entre os termos que representam os conceitos, aproximando os vocabulários das Teorias de Classificação Facetadas (Campos; Gomes, 2006; Tálamo; Lara; Kobashi, 1992).

O vocabulário controlado é o resultado da conversão da linguagem natural, representada por um conjunto de termos ou candidatos a termos. Esses descritores são

bastante utilizados na indexação de recursos informacionais. O que possibilita a recuperação da informação em sistemas manuais e automáticos. Por isso os vocabulários controlados são uma tradução da linguagem natural para uma linguagem controlada.

Para assegurar a representatividade de um termo é necessário estabelecer algum critério de seleção, pois a partir do momento em que é reunido um corpus textual de onde irá se extrair um conjunto de palavras entende-se que é necessário adequá-las aos padrões descritivos que assegurem a eficiência de um vocabulário (Tálamo; Lara; Kobashi, 1992). Por isso, a garantia literária define a perenidade de um vocabulário, além da assimetria teórica de termos padronizados em determinada área do conhecimento.

Kobashi (2008) estabelece que vocabulário controlado é uma linguagem artificial constituída de termos organizados em estrutura relacional. E sua função aplica-se a duas formas, a primeira é representação por meio de um conjunto controlado e finito de termos, também chamados de descritores. E a segunda é a padronização desses termos. Ressalta também que, em geral, os vocabulários controlados são apresentados em ordem hierárquica e alfabética considerando sua macroestrutura e microestrutura (Kobashi, 2008).

Os SOCs são conhecidos pela sua diversidade, pois existem tipos diferentes e com características pertinentes ao grau de complexidade de cada estrutura. Além de identificar o nível de complexidade que cada tipo de vocabulário atende, destaca-se também a sua função, pois as listas de assunto objetivam o controle de ambiguidade, enquanto os tesauros abrangem o controle de ambiguidade, o controle de sinônimos, relacionamentos hierárquicos e associativos.

Souza, Tudhope e Almeida (2012) apresentam algumas dimensões para classificar os KOS tomando como base as representações de (Hodge, 2000; Wright, 2008; Obrst, 2004; Daconta, Obrst, Smith, 2005; Bergman, 2007; Almeida, Souza, Fonseca, 2011; Smith; Welty, 2001; Lassila, McGuiness, 2001; Guarino, 2006; Zeng, 2008). Onde verifica-se claramente a presença do tesauro nessas representações como um tipo de KOS que pode tipificar uma expressividade semântica e um certo nível de complexidade em sua estruturação.

Souza, Tudhope e Almeida (2012) fornecem uma revisão teórica sobre SOC, ao mesmo tempo que propõe algumas dimensões, dentre as quais incluem-se, a estrutura, relações/funções semânticas, domínio, recursos sistemáticos, orientação tecnológica, padrões orientados para linguagem e conhecimento, organismos de padronização, complexidade,

análise conceitual e estrutura conceitual, análise terminológica, acesso e visualização, formato de apresentação do vocabulário, atualização, entre outros. Por isso é perceptível que esta pesquisa de certo modo se alinha à essas dimensões que balizam a construção do escopo desse trabalho de forma geral.

A National Information Standards Organization (2010) por meio da norma ANSI/NISO Z39.19 também define alguns métodos que podem ser empregados na construção de um vocabulário controlado, o primeiro denomina-se abordagem do comitê e consiste na elaboração de uma lista dos termos-chave e relações entre eles, feitos por especialistas do campo. Essa abordagem pode ser *Top Down*, quando os termos genéricos são identificados primeiro e, em seguida, os termos específicos, para atingir o nível de especificidade desejado. E *Bottom Up* que ocorre quando as listas de termos derivam de um corpus de objetos de conteúdo e, então, devem ser incorporados a um vocabulário controlado.

O segundo denomina-se abordagem empírica, caracterizada por dois métodos, o método dedutivo, onde os termos são extraídos de objetos de conteúdo (por humanos ou computadores; e o método indutivo, onde os novos termos são selecionados para inclusão potencial no controle vocabulário conforme são encontrados em objetos de conteúdo. E também a junção dessas duas abordagens que podem ser empregadas em qualquer estágio de elaboração.

O tesauro é classificado simultaneamente como um tipo de vocabulário controlado e SOC, embora existam discussões teóricas e metodológicas que abrangem essas duas esferas, os tesauros possuem concepções próprias que caracteriza de forma mais específica esse tipo de vocabulário controlado, como é apresentado a seguir.

2.3 Tesauros

A elaboração de um tesauro é um processo que sobretudo considera algumas garantias que asseguram sua estabilidade, Moreira e Moura (2006) identificam três pontos fundamentais, a garantia literária, a garantia de uso e a garantia estrutural. A National Information Standards Organization (2010) também cita garantia literária, do usuário e organizacional como justificativas para a representação de conceitos.

Entende-se que a garantia literária é uma concordância terminológica e conceitual entre os pares de uma área do conhecimento, enquanto a garantia de uso demonstra a relevância de um tesauro para o seu público. E a garantia estrutural ou organizacional seria a adoção de padrões e normas que asseguram o design de um tesauro, ou seja, porque ele foi desenhado ou possui determinadas características, como macroestrutura e microestrutura.

Moreira e Moura (2006) citam sua experiência metodológica para elaboração de um tesauro, que consiste primeiramente no estabelecimento dos objetivos, na definição de uma equipe, na seleção de tesauros já existentes na área, na coleta dos dados. No estabelecimento de políticas, tradução dos tesauros de origem, escolha das facetas do futuro tesauro, na junção e comparação entre os termos para a estruturação do novo tesauro. Smit; Kobashi (2003) e Vogel; Kobashi (2019) também apresentam uma abordagem para elaboração de vocabulários controlados porém voltada aos arquivos.

Sánchez-Cuadrado, Colmenero-Ruiz e Moreiro (2012) fazem uma revisão das recomendações e normas sobre KOS aplicada aos tesauros concluindo que os padrões e recomendações ISO se concentraram mais em tesauros do que em outros tipos de KOS, ou seja, os tesauros, enquanto um tipo de vocabulário controlado, têm se destacado como importante KOS, principalmente aplicado à web. De acordo com Mazzocchi (2018) os tesauros são tipos de vocabulários controlados e estruturados, que apresentam relações hierárquicas, associativas e equivalência entre termos/conceitos.

Segundo Tálamo; Lara; Kobashi (1992, p.199) “os tesauros são repertórios ou listas de termos autorizados, constituídos por unidades (descritores e não descritores) pertencentes a um domínio particular do conhecimento, relacionadas semântica e logicamente [...]”. Além de se caracterizarem por sua estrutura terminológica e relacional, os tesauros possuem uma sintaxe de representação que em geral, consiste em identificar os Termos Genéricos (TG), Termos Específicos (TE), Termos Relacionados (TR), entre outros.

Assim como outros tipos de vocabulários controlados, os tesauros tem por objetivo reduzir a ambiguidade, homonímia e sinonímia. Elencando um conjunto de termos autorizados e padronizados em determinada área do conhecimento. Por isso, o seu uso é comum na indexação de sistemas de recuperação da informação, embora seja possível encontrá-lo por muitas vezes isolados a esses sistemas.

Os tesauros são passíveis de estudos envolvendo tecnologias de processamento computacional, tendo em vista que sua estrutura compreende um arranjo lógico hierárquico e também semântico. Desde a década de 90 já existem registros dessa modalidade de pesquisa, por exemplo, análises de similaridade em tesauros como o Roget's thesaurus e o Macquarie thesaurus Grefenstette (1992). Com o crescente avanço da tecnologia e surgimento de novos modelos de dados, os horizontes de aplicabilidade se expandem no que diz respeito ao processamento automático de termos. Neste sentido, o uso de técnicas automatizadas com auxílio de ferramentas de software ou de programação possibilita a implementação de novos tipos de análises, como são apresentadas a seguir.

3 REPRESENTAÇÃO E PROCESSAMENTO DE TERMOS

Os vocabulários controlados podem ser representados por meio de modelos de dados em ambientes digitais, utilizando por exemplo o RDF ou SKOS. A utilização desses modelos de dados reverbera as características semânticas e contextuais sobre as quais foram projetadas. Não por acaso o RDF e o SKOS podem ser utilizados como estruturas de enriquecimento de dados para um vocabulário controlado que não apenas possuirá uma designação de termos, mas tipos de relacionamentos entre termos, notas de edição, classes, arranjos conceituais, entre outros.

A partir desses modelos de dados, os vocabulários controlados passam a assumir uma representação dinâmica da informação, com hiperligações entre si e outros vocabulários, conectando um conglomerado de termos em constante expansão. Neste sentido, é possível extrair um conjunto de métricas referentes ao processamento de um aglomerado de termos, implementado de forma automática. Deste modo, uma análise com base em um modelo de dados pode apresentar resultados com base na estrutura lógica em que os dados foram estruturados, enquanto uma análise com ênfase somente no termo pode apresentar características pertinentes à estrutura do termo.

Durante a pesquisa bibliográfica foi possível perceber uma lacuna existente entre os tipos de análises aplicadas aos modelos de dados a nível internacional e nacional. Este último com publicações que destoam do nível de complexidade internacional, por exemplo, o artigo de Dongo, Cardinale e Chbeir (2018) que analisa a inferência de tipo de dados independentes por meio de um tipo de dados *RDF inFerring Framework* chamado RDF-F. Esses tipos de dados independentes são indicados por Cyganiak, Wood e Lanthaler (2014) na recomendação da W3C *RDF 1.1 Concepts and Abstract Syntax*. Dongo, Cardinale e Chbeir (2018) que apresentam 12 definições divididas em 3 etapas de análises que consistem na análise de informação predicada, na análise de espaço léxico e na análise semântica de predicado.

Ao perceber a necessidade de realizar análises que envolvessem NLP no âmbito da Ciência da Informação foi possível encontrar registros que indicavam que este campo de análise possui publicações da década de 90 voltadas a *Text Retrieval Information Retrieval*. Inclusive Strzalkowski (1999) no prefácio do livro organizado por ele, intitulado *Natural Language Information Retrieval* enfatiza que o NLP é uma tecnologia chave para construir os

sistemas de informação do futuro. Quando aplicada, a Recuperação da Informação (RI) lida com a catalogação, categorização, classificação e busca de grandes quantidades de informação, especialmente na forma textual.

Não obstante, a Ciência da Informação possui registros de estudos voltados para NLP que por sinal pareciam promissores na época, mas com o passar dos anos foram se distanciando e minguando cada vez mais, ao ponto de não ser considerada como parte do escopo da Ciência da Informação. Tais estudos envolviam processamento computacional e técnicas avançadas de cálculos estatísticos. Todavia, é importante ressaltar que pesquisas voltadas para NLP utilizando os substratos da Ciência da Informação não é algo novo, pois em 1994 Gregory Grefenstette publicou um livro chamado *Exploration in automatic thesaurus discovery*, no qual aplica técnicas de NLP para analisar tesauros.

Neste livro Grefenstette (1994) apresenta um sistema intitulado Sextant , um sistema que utilizava contextos sintáticos para descobrir semelhanças entre palavras. Esse sistema era baseado na hipótese de que as palavras que são usadas de maneira semelhante ao longo de um *corpus* são de fato semanticamente semelhantes (Grefenstette, 1992). Para isto, Grefenstette (1994) realiza análises de similaridade em tesauros como o *Roget's thesaurus* e o *Macquarie thesaurus* (Grefenstette, 1992).

Outro ponto que merece destaque é a importância da Ciência da Informação desenvolver seus próprios *datasets* e criar *Gold Standards* e *Gold Datas*, ou seja, em uma tradução literal seriam padrões de ouro e dados de ouro, que são arquivos de referência para análises, por exemplo, Grefenstette (1994) já apresenta essa definição em seu livro, referindo-se à um *corpus* de referência para geração de métricas de acurácia, como *recall, precision* e *f-score* ou *f-measure*.

Ao utilizar bibliotecas como NLTK surge a mesma necessidade, embora exista uma série de *corpus* e palavras que podem ser utilizadas como *gold data* como por exemplo, o *brown corpus*, porém a maioria em língua inglesa e de assuntos diversos, o que dificulta o nível de análise de especialidades. Neste sentido, é importante fomentar a produção de trabalhos que analisam termos e textos de especialidade para construção de parâmetros de referência em língua portuguesa.

3.1 Processamento da Linguagem Natural

Para lidar com dificuldade de suporte no desenvolvimento de análises que envolvem tecnologias computacionais é preciso considerar que os vocabulários controlados expressos em RDF e SKOS são compostos de *word tags*, ou seja palavras ou termos descritos em linguagens de marcação. Por meio dessa característica é possível adotar técnicas de Processamento da Linguagem Natural ou *Natural Language Processing* (NLP). A NLP abrange um conjunto de técnicas computacionais que fornecem mecanismos para a análise e processamento de textos em linguagem natural.

A NLP é um campo de estudo que adota técnicas automáticas de análise da linguagem humana, convertendo para padrões legíveis por máquinas. A NLP envolve diversas aplicações na área da Linguística, Ciência da Computação e Ciência da Informação, essas aplicações envolvem, sobretudo, a análise de dados textuais. Para empregar das abordagens de NLP é necessário o uso de algum software de análise e extração de termos. Ou com o auxílio de alguma linguagem de programação, por exemplo, a linguagem de programação *Python* possui um conjunto de bibliotecas voltadas para este fim, intitulada *Natural Language ToolKit* (NLTK).

De acordo com Nadkarni; Ohno-Machado e Chapman (2011) o NLP começou na década de 1950 como a interseção da Inteligência Artificial e da Linguística. Ressaltam ainda que a NLP era originalmente distinta da Recuperação de Informações de texto (RI), que emprega técnicas baseadas em estatísticas altamente escaláveis para indexar e pesquisar grandes volumes de texto de forma eficiente. A NLP não se restringe somente a Linguística e a Computação mesmo possuindo um número expressivo de publicações sobre o tema nestas áreas. A Ciência da Informação também possui registros de pesquisas voltadas para NLP, mesmo que de forma tímida, comparado aos expoentes desse campo.

Segundo Cambria e White (2014) os problemas semânticos voltados a NLP estavam claros desde o início, mas a estratégia adotada pela comunidade de pesquisa foi abordar a sintaxe primeiro, para a aplicabilidade mais direta das técnicas de *Machine Learning*. Cambria e White (2014) ressaltam que a NLP centrada na sintaxe ainda é a maneira mais popular de gerenciar perguntas como recuperação e extração de informações, categorização automática, modelagem de tópicos, entre outros. Por isso, Cambria e White (2014) apresentam a noção de Curva Sintática e Curva Semântica em NLP. A Curva Sintática

é centrada na sintaxe e pode ser agrupada em três categorias principais: detecção de palavras-chave, afinidade lexical e métodos estatísticos. Enquanto a Curva Semântica é caracterizada por um modelo de conjunto de conceitos.

Alguns termos são comuns empregados em NLP, como a ideia de *corpus*, que é designado como um conjunto de textos que compõem o objeto de análise. Também é comum encontrar palavras como tokenização que se caracteriza como uma transformação de palavra por palavra em *token*, ou seja cada palavra recebe um determinado valor. E a expressão *n-gram*, que é uma sequência de *n* itens, quer seja letras, palavras ou fonemas, esses n-grams também pode assumir certos pares e se tornarem *bigrams*, ou trigrams (Nadkarni; Ohno-Machado; Chapman, 2011).

Leitão (2006) destaca que existem três abordagens para NLP, a abordagem simbólica que é capaz de tratar todos os níveis da análise linguística de uma Linguagem Natural (LN). A abordagem empírica que apoia-se no uso de grandes quantidades de dados e procedimentos estatísticos e a abordagem neural fundamentada em Redes Neurais Artificiais para implementação de redes semânticas.

A NLP pode ser segmentada em vários campos de análises, o *text mining* ou mineração de texto é um desses campos e refere-se ao ato de analisar textos para extrair insights. Esse processo pode ser feito por meio de análises estatísticas e probabilísticas com o objetivo de extrair padrões ou tendências. *Text mining* é um tipo de análise de dados feita em arquivos de texto que podem ser estruturados ou não. Esse processo pode ser dividido em algumas etapas que referem-se desde a seleção de um ou vários textos que formam um *corpus.* Também é caracterizado pelo uso de softwares de análise de texto que podem ser utilizados nesse processo, para executar a tokenização das palavras.

Uma vez tokenizada, cada palavra recebe um valor que permite extrair métricas de processamento linguístico. A mineração de texto, também permite a extração de *insights* e padrões significativos de dados de texto não estruturados. E pode ser aplicada para realizar análise de sentimento, classificação de conteúdo e recuperação de informações. Para que isto aconteça é necessário realizar uma série de etapas, que incluem um pré-processamento, extração de recursos, modelagem e a avaliação.

A etapa de pré-processamento envolve a limpeza e preparação dos dados textuais para análise, o que pode incluir tarefas como organização do *corpus* textual, remoção de

espaços em branco, condicionamento dos dados para um formato específico de análise, por exemplo, (*.txt*), além da tokenização e lematização. A extração de recursos envolve a identificação e extração de recursos relevantes dos dados textuais, que podem incluir *n-grams*, vetores de frequência de documento, inversão de frequência de termos e incorporação de palavras.

A modelagem envolve o uso de técnicas que também são encontradas no processo de *Data Mining*, como a classificação, agrupamento e regressão de dados. Além da construção e o treinamento de modelos de aprendizado de máquina quando for o caso. Porém, não se restringe apenas a isso, pois softwares de análises de textos e bibliotecas como NLTK permitem identificar, a frequência das palavras, características morfológicas, contexto de uso da palavra, entre outros. Por fim, a etapa de avaliação identifica o desempenho dos modelos e algoritmos de mineração de texto e a iteração do método adotado, esta é uma espécie de validação do modelo de análise empregado.

A análise de sentimentos é outra ramificação do NLP, ela considera um valor positivo ou negativo com base em um conjunto de palavras presentes em um texto. A análise de sentimentos também envolve uma variedade de tarefas e técnicas, como as que se aplicam ao *text mining*. E configura-se como uma parte importante do universo da ciência de dados, pois é utilizada para extrair informações de grandes volumes de dados de texto em diversas áreas, como redes sociais, comércio eletrônico, finanças e serviços de *streaming*. Ela pode ser utilizada para entender o sentimento do cliente em relação a um produto ou serviço, monitoramento de marca, atendimento ao cliente e identificar tendências ou padrões de opinião (Cambria; White, 2014).

A NLP é uma área da Computação muita explorada pela Linguística, em especial a Linguística Computacional, porém, não se restringe a isso, podendo ser utilizada por qualquer pesquisador que trabalhe com a manipulação de textos e termos especializados. Por exemplo, Jacquemin e Tzoukermann (1999) apresentam uma abordagem de NLP para indexação automática sobre vocabulário controlado, que leva em consideração a variação de termos em língua francesa. Essa abordagem combina um *tagger* de parte da fala, um gerador de formas morfologicamente relacionadas e um analisador transformacional leve (Jacquemin; Tzoukermann, 1999).

Outro registro que pode comprovar estudos de NLP na CI é destacado por Ruge (1999) ao demonstrar que desde 1986 existem abordagens automáticas para mapear conjuntos de palavras de documentos para termos de tesauros. E dentre as abordagens incluem-se a pseudo classificação, padrões de texto, análise lexical e observação do usuário. Embora coexista um certo desapreço de análises estatísticas, associadas apenas a quantidade e uma certa vulgarização terminológica para análises superficiais. Os estudos métricos em NLP mesmo de forma sintática demonstram justamente o contrário, porque possuem um elevado nível de complexidade para extração das métricas como identificado nos estudos de (Zhou, 1999; Guthrie; Guthrie; Leistensnider, 1999; Riloff; Lorenzen, 1999; Karlgren, 1999), entre outros autores que explicam suas pesquisas no livro editado por Strzalkowski (1999).

Talvez possa ser uma tarefa fácil desenvolver um código de programação para criar uma calculadora que efetue cálculos simples, de soma, subtração, multiplicação e divisão. Porém existe uma curva de dificuldade que varia de acordo com o tipo de cálculo a ser realizado, por exemplo, um número fatorial, associado a alguma potência ou raiz quadrada. Apesar de parecer um assunto que seja mais conveniente tratar no âmbito matemático, não se pode ignorar a necessidade dos pesquisadores lidarem com cálculos básicos e que muitas vezes compõem as equações que permitem obter certos resultados numéricos para a composição de uma tabela de dados.

Por isso, a NLP envolve o desenvolvimento de algoritmos e modelos que podem entender, interpretar e gerar análises sintáticas e semânticas usando uma ampla gama de aplicações, incluindo também a tradução automática, classificação de texto, geração de textos, clusterização extração de informações, *Machine Learning*, segmentação de texto, reconhecimento de entidade nomeada, marcação de contextos, resumo de textos, entre outros.

A NLP é um campo que realizou notáveis progressos nos últimos anos, cujas bases ainda podem ser encontradas em aplicações modernas e atualizadas constantemente, como *NLTK Tookit*, também devido aos avanços no aprendizado de máquina e à disponibilidade de grandes quantidades de dados analisáveis. No entanto, continua sendo uma área de pesquisa desafiadora, pois a linguagem natural é complexa e ambígua e está em constante evolução.

Além do uso de NLP é preciso estar atento a forma em que os dados textuais estão disponibilizados, pois há softwares que processam um número limitado de tipos de arquivos.

Enquanto uma linguagem de programação permite processar arquivos não contemplados por softwares convencionais, como é o caso do RDF e SKOS, por isso é importante conhecer sobretudo a estrutura sobre a qual são disponibilizados esses modelos de dados e consequentemente os vocabulários controlados.

3.2 Modelos de Dados

Um modelo de dados é uma maneira de organizar e representar dados em um determinado sistema, esta expressão é adotada pela W3C para referir-se ao RDF e ao SKOS. Dois modelos de dados que são capazes de definir uma estrutura de relacionamento e regras de manipulação de dados. Mas antes de discutir sobre estes dois modelos de dados é importante ressaltar que esta expressão possui uma ampla gama de definição e uso, que pode incluir representações de dados como uma coleção de tabelas.

Uma tabela estruturada com diferentes tipos de dados especificados em cada campo pode ser entendida como um *data frame*. E uma possibilidade não muito explorada de modelos RDF e SKOS é que eles podem ser convertidos em conjuntos de tabelas definidas com linhas e colunas para formarem *datasets*, especificamente em *Comma-separated values* (CSV), o que permite realizar análises baseadas em machine learning. Outra possibilidade é a conversão de dados para formatos de texto simples, como *Text File* (TXT).

No âmbito da Ciência da Informação também existem linguagens de consulta que podem ser aplicadas a modelos de dados expressos em RDF e SKOS como SPARQL *Query Language for RDF* (Clark; Feigenbaum; Torres, 2008). Esta linguagem permite executar consultas em determinados bancos de dados, como DBpedia[7] por meio de um SPARQL *Endpoint*. Estas buscas permitem apresentar resultados baseados na tripla semântica sobre a qual foi projetado o RDF. A SPARQL possui uma sintaxe de consulta baseada no SQL, porém como as especificações aplicadas aos modelos de dados linkados.

Existem duas formas de acessar integralmente um vocabulário codificado em RDF, por meio de um SPARQL *Endpoint* ou efetuando diretamente o download do arquivo. É importante ressaltar que o RDF é um modelo voltado para máquinas e não humanos, todavia é possível manipular dados em RDF com base nas recomendações disponibilizadas pela W3C. Embora seja indicado apenas para pesquisadores e especialistas. Por isso é

[7] https://www.dbpedia.org/

necessário especificar algumas das características básicas do RDF, assim como do SKOS para compreensão da etapa de análises.

O *Resource Description Framework* (RDF) é um modelo de dados utilizado para representar, organizar e trocar informações sobre recursos por meio de declarações compostas por expressões baseadas em uma tripla semântica, composta por um *subject*, *predicate* e um *object*. O RDF é utilizado para descrever recursos que representam a abstração de uma entidade (documento, conceito abstrato, pessoa, empresa, etc.). O modelo de dados RDF usa o *Internationalized Resource Identifier* (IRI), nós em branco e nós literais como elementos para criar triplas e fornecer relacionamentos entre recursos (Dongo; Cardinale; Chbeir, 2018).

O RDF é baseado na idealização de recursos conectados, cada um identificado por um IRI que é uma derivação do *Uniform Resource Identifier* (URI), relacionado às propriedades de descrição do modelo RDF, contendo os valores textuais ou numéricos e relacionados entre si. De tal forma, que os links entre os recursos expressem o contexto semântico dos recursos descritos.

Dos elementos que compõem a tripla semântica do RDF o *subject* é efetivamente designado por uma URI endereçada por um conjunto de caracteres que apontam um identificador único. O *predicate* é expresso por uma lista de prefixos previamente definidos que compõem o vocabulário do esquema RDF, ou seja, a etiquetagem de um recurso em RDF possui uma sintaxe definida orientada pelas recomendações da W3C. Por fim, o *object* é a parte literal do valor declarado em uma linguagem natural, acompanhada do idioma de origem.

O RDF é usado para representar qualquer tipo de recurso informacional, incluindo documentos, pessoas, objetos físicos e conceitos abstratos (Manola, Miller; Mcbride, 2014). E também pode ser utilizado em ontologias, para fornecer uma representação mais estruturada e formal do conhecimento. O RDF pode ser serializado em cinco formatos distintos, a saber: *Turtle* e and *TriG*, JSON-LD, RDFa, N-Triples e N-Quads e RDF/XML. O RDF possui uma sintaxe própria também chamada de vocabulário RDF, que basicamente diz respeito ao uso de elementos descritivos que expressam atributos pré-definidos.

Estes atributos identificam o recurso descrito, por meio de classes e propriedades. Logo, entendimento do RDF pressupõe a identificação de seus atributos, para isto o quadro 03 apresenta as classes RDF e suas respectivas definições. O RDF/XML possui uma codificação

em linguagem de marcação, que significa que as classes e propriedades são expressas na forma de tags, por exemplo: *<rdf:Resource>*, que representa vinculação de prefixo "rdf:" associado a RDF *Syntax namespace* para contextualização do "*Resource*".

Quadro 03 – Dicionário de Classes RDF

CLASSE	DEFINIÇÃO
rdfs:Resource	Recurso de classe, qualquer recurso.
rdfs:Literal	Classe de valores literais, por exemplo, strings textuais.
rdf:langString	Classe de valores literais de string com tags de idioma.
rdf:HTML	Classe de valores literais HTML.
rdf:XMLLiteral	Classe de valores literais XML.
rdfs:Class	Classe das classes.
rdf:Property	Classe de propriedades RDF.
rdfs:Datatype	Classe de tipos de dados RDF.
rdf:Statement	Classe de instruções RDF.
rdf:Bag	Classe de contêineres não ordenados.
rdf:Seq	Classe de contêineres ordenados.
rdf:Alt	Classe de contêineres de alternativas.
rdfs:Container	Classe de contêineres RDF.
rdfs:ContainerMembershipProperty	Classe de propriedades de associação de contêiner
rdf:List	Classe de Listas RDF.

Fonte: adaptado de Brickley; Guha (2014)

Embora as classes RDF sejam apresentadas e definidas, isto não é suficiente para o entendimento da lógica do RDF, tendo em vista que Brickley e Guha (2014) simplificaram a recomendação de Hayes e Patel-Schneider (2014) que define uma semântica do modelo teórico para grafos RDF e os vocabulários RDF. Tal recomendação faz uso intenso da lógica para interpretação das condições semânticas do RDF.

Além destas classes, o RDF possui propriedades que integram essas classes e determinam seu domínio e alcance conforme o quadro 04. As propriedades seguem a mesma regra para expressão de tags apresentada nas classes, mas é importante ressaltar que as propriedades estão de certa forma subordinadas as classes, por isso elas possuem uma limitação de domínio e alcance dentro das classes, porém de forma relacionada e não necessariamente hierárquica.

Quadro 04 – Dicionário de propriedades RDF

PROPRIEDADE	DEFINIÇÃO
rdf:type	O assunto é uma instância de uma classe.
rdfs:subClassOf	O assunto é uma subclasse de uma classe.
rdfs:subPropertyOf	O sujeito é uma subpropriedade de uma propriedade.
rdfs:domain	Um domínio da propriedade do assunto.
rdfs:range	Um intervalo da propriedade do assunto.
rdfs:label	Um nome legível para o assunto.
rdfs:comment	Uma descrição do recurso de assunto.
rdfs:member	Um membro do recurso de assunto.
rdf:first	O primeiro item na lista RDF de assunto.
rdf:rest	O restante da lista de assuntos RDF após o primeiro item.
rdfs:seeAlso	Mais informações sobre o recurso de assunto.
rdfs:isDefinedBy	A definição do recurso de assunto.
rdf:value	Propriedade idiomática usada para valores estruturados.
rdf:subject	O assunto da instrução RDF de assunto.
rdf:predicate	O predicado da instrução RDF do assunto.
rdf:object	O objeto da instrução RDF do assunto.

Fonte: adaptado de Brickley; Guha (2014)

A identificação dessas classes e propriedades servem como elemento de consulta para a leitura de arquivos que estão disponíveis no formato RDF. Portanto, trata-se de uma versão resumida do vocabulário RDF e RDF Schema. O RDF possui uma robustez semântica e uma sintaxe que parece desafiadora, talvez seja esse o motivo para não ter alcançado tanta popularidade, mesmo que alguns softwares utilizados para a gestão de vocabulários controlados disponibilizem exportações de dados para o formato RDF e SKOS.

O *Simple Knowledge Organization System* (SKOS) também é um modelo de dados que surgiu de uma tentativa de publicar um tesauro em RDF, por isso é um modelo de dados com maior afinidade para vocabulários controlados (Baker; Bechhofer; Isaac; Miles; Schreiber; Summers, 2013). O SKOS é uma espécie de simplificação das propriedades semânticas do RDF, não por acaso, simplificar o RDF é uma tarefa tão hercúlea, que desenvolver um novo modelo parecia mais sensato. O SKOS foi projetado para oferecer

suporte ao desenvolvimento, manutenção e gerenciamento de SOCs, por isso sua utilização é viável para os distintos tipos de vocabulários controlados.

O SKOS manteve a premissa semântica do RDF, porém de forma simplificada, baseado na ideia de representar o conhecimento como uma rede de conceitos, onde cada conceito é definido por um conjunto de propriedades e relacionamentos com outros conceitos (Dahlberg, 1978; Catarino, 2014; Lara, 2015; Ramalho, 2015). Estes relacionamentos podem ser basicamente podem ser estabelecidos de forma ampla ou específica. Assim como o RDF, o SKOS possui uma sintaxe baseada na linguagem de marcação XML e possui um conjunto de prefixos declarativos que compõem a sintaxe do vocabulário SKOS.

Por ser uma recomendação da W3C, o SKOS permite a publicação de SOCs no ambiente *web*, para que os conceitos sejam vinculados entre si e formem esquemas conceituais. Cada conceito é apontado por uma URI e etiquetado por um termo ou *string*. O SKOS permite criar relações semânticas entre conceitos diferentes representados por seus respectivos termos. O SKOS possui um *background* normativo sustentado pela ISO 25964, o que dá a ele uma característica ímpar como formato citado pela ISO além de ser um recomendação da W3C. Além do o fato de possuir equivalências conceituais e estruturais com os principais tipos de vocabulários controlados, permitindo o usufruto de uma tecnologia emergente aplicada aos SOCs.

O SKOS também possui uma lista de *Prefix Vocab* que representa o vocabulário SKOS por meio de *tags* autorizadas. A estrutura lógica e semântica do SKOS baseada em vocabulários controlados pode ser definida pelo usos das seguintes *tags*: <*skos:broader*>, <*skos:narrower*> e <*skos:related*>, para criar relações genéricas, específicas e relacionadas. As propriedades de etiquetagem de recursos conceituais por meio de termos podem ser respectivamente atribuídas as tags <*skos:prefLabel*>, <*skos:altLabel*> e <*skos:hiddenLabel*>, para termos preferenciais, alternativos e ocultos.

O SKOS é caracterizado por uma série de propriedades categorizadas de acordo com a sua função, por exemplo, rótulos lexicais para inserção de termos, propriedades de documentação para inserir notas associadas à um termo, relações semânticas para definir os níveis de relacionamentos entre os termos, e assim por diante, conforme o quadro 05 abaixo. O vocabulário SKOS possui uma abordagem baseada na simplificação de SOCs, por isso as tags autorizadas são auto descritivas.

Quadro 05 - Propriedades SKOS

Classe Conceitual	**Esquemas Conceituais**	**Rótulos Lexicais**	**Notações**
skos:Concept	skos:ConceptScheme skos:inScheme skos:hasTopConcept skos:topConceptOf	skos:altLabel skos:hiddenLabel skos:prefLabel	skos:notation
Propriedades da Documentação	**Relações Semânticas**	**Coleções Conceituais**	**Propriedades de Mapeamento**
skos:changeNote skos:definition skos:editorialNote skos:example skos:historyNote skos:note skos:scopeNote	skos:broader skos:broaderTransitive skos:narrower skos:narrowerTransitive skos:related skos:semanticRelation	skos:Collection skos:OrderedCollection skos:member skos:memberList	skos:broadMatch skos:closeMatch skos:exactMatch skos:mappingRelation skos:narrowMatch skos:relatedMatch

Fonte: Adaptado de Isaac; Summers (2009) e Sousa (2019)

As categorias que definem as propriedades do SKOS enfatizam a representação conceitual, pois meio de classes, esquemas e coleções de conceitos, os rótulos lexicais se encarregam de associar termos a cada conceito, enquanto as propriedades de documentação expandem a descrição e adicionam novas informações (Isaac; Summers, 2009). As propriedades de mapeamento são características aprimoradas das relações semânticas, pois permitem inferências entre os conceitos, entendidas, portanto, como superpropriedades semânticas. Neste sentido, o SKOS é um modelo de dados semântico que incorpora as características essenciais dos SOCs, elencando o conceito como núcleo estruturante.

O SKOS e o RDF são os principais modelos de dados utilizados nesta pesquisa, principalmente para verificar se os vocabulários analisados estão de acordo com as recomendações que estão sendo adotadas para expressar tesauros, quer seja de maneira simplificada por meio do SKOS ou de forma completa utilizando o RDF. Todavia percebe-se a dificuldade de lidar com estes modelos de dados, tanto que *American Library Association* publicou em 2015 um guia para representar vocabulários controlados estruturados em SKOS.

No guia citado existe uma definição detalhada sobre cada *tag* de acordo com a sua função (Frazier, 2015).

Embora seja relevante esta iniciativa de tutoria profissional, ela não supre a lacuna que estes modelos de dados possuem entre a possibilidade de uso e a realidade enfrentada por profissionais da informação, quando o assunto envolve análises e desenvolvimento de instrumentos tecnológicos que envolvem ferramentas computacionais. Além do fato desses modelos de dados não serem de uso comum para maioria dos desenvolvedores de sistemas, o que dificulta o processo de análise e resolução de problemas durante a elaboração de códigos de programação. Por isso, a avaliação de vocabulários controlados por meio de programação e uso de softwares tornou-se um desafio para os profissionais da informação.

Os métodos de avaliação incluídos no escopo da Ciência da Informação possuem maior amplitude em nível internacional, onde é possível identificar dezenas de artigos com técnicas sofisticadas de análise computacional envolvendo diversos tipos de vocabulários controlados. O que se percebe é que não somente do ponto de vista aplicado, mas teórico esses métodos de avaliação foram se consolidando enquanto outros foram se tornando obsoletos ou até mesmo esquecidos. No entanto, é preciso recuperar algumas definições já discutidas e estabelecidas que podem servir de base para novas técnicas de avaliação, como é o caso da onometria, detalhada na próxima seção.

3.3 Avaliação de termos (Onometria)

Utilizar técnicas de NLP em *corpus* textual convencional costuma gerar métricas relacionadas aos tokens analisados, isto também é possível quando aplicado aos vocabulários controlados. Ainda que a existência de 'bancos de dados terminológicos' (Faulstich, 1995), não seja uma novidade. As análises deste tipo de dado não se tornaram comuns no âmbito da Terminologia. Porém, não foi por falta de tentativas, como é possível verificar na publicação de Gilreath (1993), onde propõe uma abordagem formal para a avaliação de termos.

Para designar essa abordagem, Gilreath cunhou o termo *Onometrics* que pode ser traduzido como *Onometria*, um neologismo derivado das raízes gregas onoma (nome) e metron (medida). A onometria é a ciência e a prática da avaliação de termos. Ela se aplica tanto à teoria (princípios básicos) quanto à prática cotidiana de selecionar ou avaliar termos.

(Gilreath, 1993). Embora seja comum o uso de normas para orientar a elaboração de vocabulários controlados, incluindo os processos que abrangem a seleção de termos, há de se questionar os critérios adotados para tal fim.

Por exemplo, a seleção de termos tradicionalmente foi encabeçada por especialistas em terminologia, quer fossem utilizados em vocabulários controlados ou no processo de indexação, recorrendo aos processos cognitivos (Fujita, Boccato, Rubi, Gonçalves, 2009; Alonso-Arroyo, Fujita, Gil-Leiva, Pandiella, 2016). Tal característica está presente em tesauros, taxonomias, entre outros, por isso surge o desafio de objetivar computacionalmente os artefatos oriundos de processos subjetivos. Não reduzindo-se a ideia de afastar a influência humana da produção de coisas artificiais, mas identificar padrões de elaboração e parâmetros que possam ser metrificados.

Logo, de acordo com Gilreath (1993) a onometria pode ajudar terminologias instáveis em direção à padronização e, uma vez que a padronização se torne viável, pode ser útil de duas maneiras: proporcionando aos desenvolvedores de padrões uma ferramenta eficaz para avaliar os termos propostos e implementando padrões onometricamente justificados. A onometria pode ser útil para orientações teóricas envolvendo a nomenclatura de novos conceitos ou a padronização de nomes para conceitos estabelecidos, mas não se restringe a isso. Gilreath (1993) ressalta que os princípios onométricos também se aplicam ao discurso expositivo (escrita e fala) no geral, pois escritores e oradores podem ser tão culpados de usar termos imprecisos, opacos, enganosos ou de outra forma falhos, quanto os nomeadores de conceitos (Gilreath, 1993).

A onometria em si não é uma novidade, pois existem muitos tipos de análises de palavras, conceitos, mas o termo onometria é incomum, principalmente na literatura brasileira, pois, ao realizar uma busca na BDTD em português e inglês sobre esse termo não foi encontrado registros até o ano de 2023, embora tenha originado-se em 1993. É provável que esta seja a primeira pesquisa a mencionar esse termo, mas existem termos com características semelhantes, como a palavra cunhada por Fred W. Riggs designada como onomântica em referência a uma abordagem terminológica na qual primeiro um conceito é dado e então seus termos são procurados (Gilreath, 1993).

Ao contrário da onometria, a onomântica é mencionada na literatura brasileira, especialmente por meio de um artigo publicado por (Gomes; Campos; Guimarães, 2010).

Uma característica comum dessas duas palavras é que ambas não foram registradas ou indexadas em português no portal de periódicos CAPES, apenas em inglês é possível encontrar 7 registros de cada palavra. Também não há registros em português ou inglês na BDTD.

O que se percebe é que Riggs (1996) e Gilreath (1993) apresentam abordagens que foram ofuscadas ao longo do tempo, mas que possuem uma grande relevância nas discussões sobre terminologia. A onomântica por exemplo, possui uma abordagem de nomeação ou busca de termos que vai do conceito ao termo (Gilreath, 1993). Enquanto a onometria foca não apenas em termos, mas bons termos por meio de uma avaliação. Por esse motivo, a onometria pode ser executada por meio de 17 critérios de avaliação de termos, para realizar uma análise sistemática, explícita, objetiva e completa, com intuito de apontar métricas que possam refletir as características de um termo.

Antes de conhecer estes critérios é importante destacar que o fato de utilizar métricas para mensurar as características de um termo, não se restringe apenas ao ato de coletar dados estatísticos. Porém no sentido de medir ou avaliar coisas como ressalta Gilreath (1993), abrangendo termos, fundamentos lógicos, teorias, parágrafos, definições, software ou sistemas. Por isso, há uma certa vantagem no uso de métricas, pois tende a melhorar nossa compreensão de um objeto, aprimora técnicas e produtos semelhantes e eleva a qualidade da comunicação. Fornecendo uma maneira objetiva, sistemática e confiável de avaliar artefatos (Gilreath, 1993).

Heurística, testes, escalas e valores também são mencionados por Gilreath (1993) ao explicar que heurísticas são testes, regras, fórmulas, algoritmos, procedimentos, planos e diretrizes para atividades relacionadas ao artefato que será analisado. Na onometria, a heurística abrange a formação, seleção e reconhecimento de termos bons e ruins por meio de testes ou classificações e outros critérios. Um teste é uma forma prática de medir um critério e, assim, colocá-lo em uma escala, alguns critérios são mais passíveis de quantificação do que outros (Gilreath, 1993). Por isso, uma escala pode servir como base para determinar um intervalo de valores que os objetos medidos podem ter para um determinado critério, ela ainda pode ser binária, contínua ou graduada.

Gilreath (1993) permite afirmar que uma escala binária opera com duas opções opostas entre si, por exemplo: sim ou não. Enquanto as escalas contínuas podem ser

chamadas de escalas analógicas, inclinadas, graduais ou quantificadas, por exemplo uma régua. E as escalas graduadas podem ser chamadas de escalas discretas ou escalonadas. Adotando como referência de medida baixo e alto, para indicar o menor e maior valor. Por fim, os valores são pontuações ou classificações específicas de um objeto para o critério fornecido (Gilreath, 1993). Por meio dessas definições o constructo onométrico começa a tomar forma, no entanto, os critérios podem ser conflitantes entre si. Logo, entende-se que não há uma obrigatoriedade em atender todos os critérios para identificar a qualidade de um termo ou que possa ser reduzido apenas para 17 como é estabelecido a priori.

Gilreath (1993) deixa claro que esses critérios podem aumentar ou diminuir, por isso é necessário ajustar ao propósito que se pretende alcançar com uma abordagem onométrica. Os critérios de avaliação de termos propostos por Gilreath (1993) são: (1) acurácia, (2) precisão, (3) descritividade, (4) inequívoco, (5) mononímia, (6) registro apropriado, (7) precedente, (8) concisão, (9) simplicidade apropriada, (10) correção de forma, (11) pureza etimológica, (12) derivabilidade, (13) inflexibilidade, (14) uniformidade de série, (15) aceitabilidade, (16) eufonia e (17) pronunciabilidade. Para simplificar as características de cada critério, foi elaborado o quadro 06 com as respectivas definições.

Quadro 06- Critérios onométricos

CRITÉRIO	DEFINIÇÃO
Acurácia	Qualidade do termo determinada pela ausência ou presença de elementos enganosos ou incorretos.
Precisão	Grau em que um termo delineia claramente seu conceito.
Descritividade	Grau em que o significado literal de um termo corresponde ao seu significado pretendido.
Inequívoco	Qualidade de um termo que tem apenas um significado dentro de um determinado campo de conhecimento.
Mononímia	Qualidade de um termo que é o único (mono) nome formal (nym) para um determinado conceito.
Registro apropriado	O estilo de um termo (registro) é consistente ou compatível com o contexto de uso

Precedente	Medida em que uma designação proposta está em harmonia com as designações estabelecidas.
Concisão	Comprimento ortográfico de um termo em comparação com outro.
Simplicidade Apropriada	Significa que o número de palavras em um termo é apropriado para o nível de importância do conceito designado.
Correção de forma	Medida em que um termo não tem erros gramaticais, como erros ortográficos. hifenização errada, ordem (invertida) errada.
Pureza etimológica	Termos construídos a partir de elementos derivados de uma única língua.
Derivabilidade	Qualidade dos termos cujos elementos podem ser usados para nomear uma variedade de conceitos relacionados.
Inflexibilidade	Qualidade dos termos que flexionam bem em formas como comparativos, superlativos e negativos (antônimos).
Uniformidade de série	Qualidade de um grupo de termos que usam elementos comuns para nomear conceitos relacionados.
Aceitabilidade	Termos que não são carregados emocionalmente, obscenos, mórbidos, tendenciosos quanto ao gênero, informais, etc.
Eufonia	Qualidade fonética de um termo ter um som agradável.
Pronunciabilidade	Refere-se ao nível de dificuldade para pronunciar um termo.

Fonte: Elaborado a partir de Gilreath (1993)

Os critérios elencados por Gilreath (1993) são baseados nas publicações de diversos autores, como Felber (1989), Nybakken (1959) e também faz uso de normas ISO. Entretanto, alguns critérios podem soar estranhos, por exemplo, eufonia e pronunciabilidade. A proposição de metrificar o som de uma palavra transfere a responsabilidade de selecionar termos que sejam pronunciáveis sabendo que no âmbito científico há muitas designações de

termo, cuja origem remete ao Latim. Assim como palavras demasiadamente grandes ou compostas que estão associadas a um objeto, instrumento, metodologia ou instituição.

A abordagem onométrica pode ser abrangente quando considera todos os critérios ou seletiva, quando opta apenas por alguns. Ela pode ser utilizada por desenvolvedores de padrões, especialistas em terminologia, como pesquisadores independentes. O intuito é minimizar a intuição e a subjetividade na avaliação de termos, fornecendo fundamentos lógicos para a preferência e depreciação de termos. Embora não seja totalmente garantida a eliminação da subjetividade, tendo em vista que há idiomas como a língua portuguesa no Brasil, com um repertório de palavras ambíguas, todavia, esta abordagem objetiva pode melhorar significativamente a comunicação científica.

A avaliação de termos não se restringe apenas a um único termo, embora se reconheça a dificuldade de avaliar termo a termo. Porém, a avaliação também pode ser comparativa - quando um determinado termo é comparado, considerando alternativas para um ou mais critérios (Gilreath, 1993). Um dos grandes desafios da abordagem onométrica é que não há uma ilustração clara de como tornar essa avaliação possível, por isso, Gilreath (1993) até sugere o uso de uma Inteligência Artificial para auxiliar nesse processo. O que permite acrescentar ferramentas e técnicas de análises automatizadas que possam fornecer as métricas que se deseja alcançar para cada critério elencado e torná-los mais objetivos.

A extração de métricas oriundas de termos podem auxiliar na compreensão de instrumentos terminológicos. Para Gilreath (1993) os princípios onométricos devem ajudar os especialistas de praticamente qualquer disciplina a criar, selecionar e usar uma terminologia melhor e facilitar a harmonização da terminologia. A designação de harmonização terminológica é proposta por Gilreath (1992), que a define como o processo no qual diversas posições são amplamente reconciliadas e assimiladas em uma única posição unificada, por exemplo, harmonização de sistemas de conceitos, definições e termos. Também estabelece graus de harmonia que oscilam entre unanimidade, consenso, maioria e pluralidade.

De acordo com Gilreath (1994) outra aplicabilidade para a abordagem onométrica refere-se a disputa de termos, que seria um desacordo sobre como nomear um determinado conceito (sentido), a sinonímia é um exemplo de disputa de termos. Embora sejam elencados 17 critérios, para as disputas de termos, apenas 16 seriam aplicáveis, pois a mononímia, um dos critérios listados, não se aplicaria a esse tipo de análise. Neste sentido, a onometria

permite objetivar o processo de seleção de termos preferenciais para conceitos e pode ser adaptada para que cada critério seja atribuído à um peso de importância para medir (classificar) sinônimos comparados (Gilreath, 1994).

Por entender que as disputas de termos avançam para uma análise mais conceitual, ela não será aprofundada nesta pesquisa. Porém, a abordagem onométrica será utilizada como orientação teórica para execução da análise prática. Neste sentido, a análise métrica de termos em vocabulários controlados faz uso da abordagem onométrica e técnicas de NLP para extração de métricas que possam traduzir as características de determinados vocabulários controlados. Logo, a análise de vocabulários divide-se em três tipos, aqui definidas como análise estrutural quantitativa, análise baseada em modelos de dados e análise métrica de termos.

Análise estrutural quantitativa objetiva extrair dados quantitativos das estruturas de descrição dos vocabulários controlados, expressos por meio de tags descritivas que representam as características de um vocabulário. Por exemplo "prefLabel", deste modo é possível verificar a composição de um vocabulário e suas características predominantes. A análise baseada em modelos de dados é uma análise especializada, que dependerá do modelo de dados ou formato em que estiver disponível o vocabulário controlado. Esta análise pode ser realizada com o uso de softwares ou linguagens de programação que processam esses modelos de dados e fornecem resultados sobre a inconsistência de dados.

A análise métrica de termos pode ser realizada com o auxílio de software ou um conjunto de ferramentas de bibliotecas utilizadas em linguagens de programação. Este tipo de análise permite identificar desde as características sintáticas dos termos até os tipos de relacionamentos que aderem a um perfil semântico. A análise de termos também pode ser utilizada para identificar padrões e tendências no uso da linguagem, quer seja especializada ou não.

A onometria pode ser implementada no processo de avaliação de um vocabulário controlado. A priori sua influência de uso é de fato muito mais teórica do que aplicada, por se tratar de uma descoberta teórica que precisa de mais estudos de caráter exploratório, com intuito de melhor compreender esse método de avaliação. Por isso é importante fundamentar-se também em definições mais consolidadas sobre o que é o processo de avaliação de forma geral e aproximar essa discussão aos registros encontrados na literatura

sobre os diversos tipos de avaliações ou análises similares de vocabulários controlados que podem contribuir para esta pesquisa.

4 AVALIAÇÃO DE VOCABULÁRIOS CONTROLADOS

O termo avaliação possui definições em diversas áreas, em uma perspectiva geral a definição de Contandriopoulos, Champagne, Denis e Pineault (1997) oferece uma noção geral sobre o assunto. Ao afirmar que avaliar consiste fundamentalmente em fazer um julgamento de valor a respeito de uma intervenção ou sobre qualquer um de seus componentes, com o objetivo de ajudar na tomada de decisões. Este julgamento pode ser resultado da aplicação de critérios e de normas (avaliação normativa) ou se elabora a partir de um procedimento científico (pesquisa avaliativa) (Contandriopoulos, Champagne, Denis E Pineault, 1997).

Contandriopoulos (2006) enfatiza que avaliar pode ser uma atividade que consiste em aplicar um julgamento de valor a uma intervenção, através de um dispositivo capaz de fornecer informações cientificamente válidas e socialmente legítimas, permitindo aos diferentes atores envolvidos, um julgamento capaz de ser traduzido em ação. Neste sentido Contandriopoulos (2006) destaca que o ato de avaliar produz um impacto direto na tomada de decisão.

Hadji (2001) é outro autor fora do escopo da Ciência da Informação que apresenta alguns tipos de avaliação no processo de aprendizagem e cita dois tipos de avaliação. A avaliação de referência normativa, onde ressalta que a avaliação livre de normas é utopia, sem possibilidade lógica. A norma não é subjugante nem libertadora, é um modelo valorizado pelos pares. E a avaliação criterial aprecia determinado comportamento situando-o em relação a um alvo – critério ou objetivo a ser atingido. A avaliação normativa é em parte criteriosa, porque situa alguns desempenhos com relação aos outros e se refere a critérios de conteúdo.

Na Ciência da Informação existem algumas indicações que envolvem a questão da avaliação da informação e da produção científica (Paim; Nehmy, 2008; Alvarez, Caregnato, 2017). Algumas abordagens contemporâneas sobre avaliação em Arquivologia e Ciência da Informação são apresentadas por Ferreira, Rockembach (2017) que destaca dois tipos: a análise do contexto e uma abordagem pós-moderna, que destaca a macro avaliação e a avaliação do fluxo da informação, com uma abordagem científica. Há outros tipos de avaliação que são apresentados ao longo texto, todavia, a maioria é restrito a determinado tipo de KOS.

O que é importante destacar é que autores como (Contandriopoulos, Champagne, Denis, Pineault, 1997; Contandriopoulos, 2006; Hadji, 2001) enfatizam o aspecto normativo do processo de avaliação. Neste sentido, as normas para a implementação de vocabulários controlados são importantes referências para o desenvolvimento de parâmetros de verificação. Outro ponto destacado é que a National Information Standards Organization (2017) destaca que a avaliação de vocabulários no contexto da organização ou projeto pode incluir a identificação de licenças, políticas de manutenção e adequação, o que indica que a manutenção e continuidade de um vocabulário controlado pode ser o resultado de um processo de avaliação.

A National Information Standards Organization (2010) não cita diretamente o processo de avaliação, mas ressalta a importância da manutenção dos vocabulários controlados enfatizando que esse processo deve permanecer durante a vida útil do produto. A National Information Standards Organization (2010) pontua mais claramente o processo de avaliação, apontando inclusive três tipos, que são: avaliação heurística, modelagem de afinidade e teste de usabilidade. Além disso, destaca os dois principais motivos para avaliar vocabulários, que consistem em determinar se o vocabulário controlado fornece resultados de pesquisa adequados (ou seja, alta relevância e *recall*) e se corresponde às expectativas dos usuários quanto aos termos contidos nos vocabulários (National Information Standards Organization, 2010).

De acordo com um relatório técnico publicado pela National Information Standards Organization (2017) sobre problemas no gerenciamento de vocabulários controlados, existe a possibilidade de um vocabulário se tornar "órfão", significando vocabulários cujas organizações de manutenção perderam financiamento ou não puderam fazer a transição da impressão para plataformas digitais e não há mais uma fonte autorizada de terminologia, política, documentação ou informações de licenciamento (National Information Standards Organization, 2017).

Um vocabulário órfão é apenas um tipo de problema encontrado, outros que podem emergir se referem ao processo de transição de um vocabulário controlado. Por fim, sugerem realizar algum tipo de reengenharia na construção ou conversão de vocabulários. Esse processo de reengenharia poderia ser substituído por avaliações periódicas dos vocabulários, adequando-os às demandas de ordem conceitual e tecnológica. Na Ciência da

Informação existem distintos tipos de avaliação, que incluem métodos com abordagens objetivas, subjetivas, automáticas ou manuais. Enfatizando a usabilidade, estrutura, relevância e a qualidade geral de um vocabulário.

Urdiciain (1998) aponta que existem diversos estudos sobre avaliação de sistemas principalmente com ênfase na recuperação da informação, destacando que o primeiro teste conhecido foi o *ASTIA-Uniterm test* em 1953. Svenonius (1986) questiona se experimentos de recuperação são a melhor abordagem para avaliar a eficácia de diferentes formas de controle de vocabulário. Por isso é importante considerar a importância de estabelecer tipos avaliações que consigam captar diferentes nuances dos vocabulários controlados

Park, Richards e Brenza (2019) explicam a importância da avaliação, inclusive para o avanço de padrões e modelos de representação de dados, como é o caso do BIBFRAME. Park, Richards e Brenza (2019) ressaltam que a avaliação de vocabulários controlados em toda a web é uma área que parece particularmente relevante para estudos posteriores, o BIBFRAME incentiva a descrição de relacionamentos com fontes de dados oficiais desenvolvidos e mantidos em uma multiplicidade de contextos. No entanto, muitos vocabulários controlados são desenvolvidos com tempo limitado e, portanto, não possuem governança estável para desenvolvimento e manutenção contínua (Park; Richards; Brenza, 2019).

Baker, Vandenbussche e Vatant (2013) destacam uma possível governança de vocabulários, enfatizando a necessidade de um plano que assegure que um recurso identificável por meio de uma URI seja acessível no presente e no futuro. Pois, quando não há planos de manutenção de um vocabulário, seus recursos podem se tornar inacessíveis ou desconectados de outros dados e fontes. Neste sentido, ressaltam que os provedores de dados e os designers de vocabulário não podem seguir o princípio de reutilização, a menos que possam encontrar o que precisam reutilizar (Baker, Vandenbussche e Vatant, 2013).

Também é importante destacar, que um vocabulário publicado e acessível significa que é facilmente analisável, pois existem vocabulários que possuem milhares de termos e relacionamentos, o que inviabiliza o uso de técnicas manuais. De acordo com Nayak, Dutta, Ajwani, Nicholson e Sala (2019), há uma necessidade de técnicas automatizadas, para avaliar a evolução de hierarquias de conhecimento muito grandes que capturam a subsunção semântica lógica. Por isso, analisar grandes volumes de dados tem se tornado um desafio,

interligando vários campos de estudo, que envolvem preservação, acesso, avaliação, governança, entre outros.

A análise de grandes volumes de dados é uma tendência que se consolida e implica diretamente na tomada de decisão. Os vocabulários controlados precisam ser observados também sob essa perspectiva, pois operam com unidades básicas do conhecimento em contínua expansão. Com base no resultado da avaliação, os responsáveis de atualizar o esquema organizacional poderiam tomar decisões informadas sobre os novos itens de dados inseridos, classes que precisam ser adequadas, que novas classes são necessárias, reorganização de esquemas, entre outros (Börner, Hardy, Herr, Holloway e Paley, 2007).

Zhang, Ogletree, Greenberg e Rowell (2015) afirmam que vocabulários controlados são sistemas semânticos úteis para organizar e acessar recursos e para apoiar a interoperabilidade semântica entre descrições de objetos e repositórios. Os vocabulários controlados permitem uma organização estruturada do conhecimento por meio de termos e esquemas de representação.

O desenvolvimento de um modelo de representação é uma das etapas contidas no processo de Organização do Conhecimento, por esta razão, existem diversos modelos de dados e metadados disponíveis no ambiente digital. Atrelado a este fator, o uso de ferramentas que possam dar sentido a estes modelos de representação é cada vez maior, principalmente no que diz respeito aos SOCs.

A *priori* é possível indicar algumas contribuições bibliográficas que podem auxiliar na construção de algumas diretrizes para avaliação de vocabulários controlados, do tipo tesauros. Todavia, foi necessário recorrer a verificação dos processos de avaliação de diferentes tipos de vocabulários controlados e SOCs, como a taxonomia e até mesmo as ontologias, ainda que esta última não seja um tipo de vocabulário controlado, é possível obter alguns parâmetros de avaliação utilizados em sistemas mais complexos.

No âmbito das taxonomias Aquino, Carlan e Brascher (2009) analisaram de acordo com critérios de categorização, controle terminológico, relacionamento entre termos e multidimensionalidade. Enquanto Cavalcante (2012) propõe a adoção de dois critérios para avaliação de taxonomias que são comunicabilidade e organização executáveis por meio da análise dos níveis hierárquicos, utilização de facetas e definição de inconsistências.

No âmbito dos tesauros Soler-Monreal e Gil-Leiva (2011) apresentam uma divisão em grupos principais que denominam de avaliação qualitativa intrínseca e avaliação intrínseca quantitativa ou estatística extrínseca. Para Soler-Monreal e Gil-Leiva (2011) a avaliação intrínseca pode ser realizada para analisar a estrutura, os campos temáticos ou facetas, notas de escopo, relações semânticas, grau de especificidade entre outros aspectos de um vocabulário controlado.

Por outro lado, a avaliação extrínseca estuda o impacto nos sistemas de informação que os utilizam tanto na indexação quanto na recuperação (Soler-Monreal e Gil-Leiva, 2011). Enquanto Quarati, Albertoni e Martino (2016) se propõem a avaliar tesauros em SKOS e afirmam em um dos tópicos de seu trabalho que "não se pode controlar o que não se pode medir", neste sentido utilizam o *Analytic Hierarchy Process* (AHP) com abordagem de análise de vocabulários em SKOS.

Embora não pertença aos vocabulários controlados, as ontologias são tipos de SOCS que podem ser utilizadas como referências externas de avaliação. Kim e Oh (2019) explicam que as metodologias de avaliação de ontologias podem se enquadrar em quatro categorias, que consistem em comparar ontologias, utilizar e verificar suas funcionalidades, ao avaliar a adequação e interconexão com os dados de origem e avaliando por meio de especialistas com base em critérios predefinidos. Uma das grandes dificuldades de avaliar uma ontologia é que sua concepção é um procedimento complexo. Na Ciência da Informação Ramalho (2010) propõe um método de desenvolvimento de ontologias à luz da teoria do conceito e a teoria da classificação facetada, utilizadas na elaboração de tesauros (Campos e Gomes, 2006; Gomes, 1990).

Independentemente do tipo e função do vocabulário controlado, a avaliação precisa ser incluída desde o processo de elaboração. Os variados tipos de avaliação operam sob uma perspectiva de uso e relevância dos vocabulários controlados, Lorenzon (2011) corrobora esta afirmação destacando que a maioria dos estudos sobre avaliação de tesauros ou de linguagens de indexação está relacionada com a análise do desempenho do serviço de indexação para medir o quanto o serviço atende ou não as solicitações dos seus usuários.

Há diversas formas de analisar um vocabulário controlado, por exemplo, Vállez, Pedraza-Jiménez, Codina, Blanco e Rovira (2015) analisaram por meio dos *logs* de consulta, para saber o quão bem adaptado era um vocabulário controlado à maneira atual de acessar

informações por meio de pesquisas por palavras-chave nos mecanismos de pesquisa. Assim como pela verificação de erros e problemas estruturais conforme destacam Manaf; Bechhofer; Stevens (2012), Mastora; Peponakis; Kapidakis (2017), Mader; Haslhofer; Isaac (2012), Suominen; Mader, (2014) e Suominen; Hyvönen (2012).

Kless e Milton (2010) destacam que os seguintes critérios também podem ser utilizados na avaliação de um vocabulário controlado, são eles: pureza, exaustividade, ausência de redundâncias e clareza conceitual, nível de pré-coordenação, correção sintática e estrutural, completeza e pureza estrutural, navegabilidade, qualidade de documentação, complexidade e consistência. Moreiro-González (2011), enfatiza que outros critérios podem ser incorporados no processo de elaboração de um vocabulário controlado: comunicabilidade, utilidade, motivação ou estimulação e compatibilidade.

A National Information Standards Organization (2010) destaca que os alguns métodos devem ser considerados quando se pretende medir a qualidade ou eficácia de um vocabulário controlado, referindo-se, portanto, a três tipos de avaliação. A avaliação heurística que pode ser informal e qualitativa ou formal e quantitativa com base em um conjunto de critérios que definem a Interação Humano-Computador (IHC) (Nielsen e Molich, 1990).

A modelagem de afinidade reúne uma amostra representativa de usuários para avaliar um vocabulário solicitando uma classificação dos termos, analisados contra a hierarquia de termos existentes para identificar o nível de similaridade. Por fim, o teste de usabilidade, que é uma espécie de avaliação holística do sistema de informação, que fornece informações sobre a eficácia do vocabulário controlado e propõe o estabelecimento de métricas de uso (Nielsen, 1993).

Soergel (2002) e Lancaster (2002) consideram que a avaliação possa responder a determinados critérios como englobar aspectos sintáticos, semânticos e conceituais dos termos, assim como os seus relacionamentos podem conter critérios adicionais aplicáveis aos vocabulários acessíveis eletronicamente, Além de razões de equivalência, reciprocidade, definição, flexibilidade, níveis de pré-coordenação com termos estruturados hierarquicamente e tamanho do grupo de termos.

Pinto (2008) considera que os fatores que determinam a qualidade de um vocabulário controlado são medidos por meio do arcabouço conceitual, desempenho

(performance), formato e sistema de auxílio. Essas tendências de análises deixaram em estado de latência a preocupação com a manutenção dos vocabulários controlados, considerando a vivacidade linguística contida nestes SOCs. Lopes (2000) ressalta que os vocabulários controlados não permanecem estáticos, ao longo dos anos, pois são incorporados novos termos e a tendência é aumentar a especificidade, ou ainda, provocar a substituição desses termos a partir de uma determinada data.

Currás (2010) destaca a importância da atualização de um vocabulário controlado ao longo da sua vida útil. A manutenção nem sempre está prevista durante o processo de elaboração, por isso, Martínez García (2009) enfatiza que o principal problema enfrentado pelos tesauros tradicionais é a obsolescência. Esta preocupação com a sustentabilidade dos vocabulários reacendeu, portanto, a valorização das garantias que envolvem sua concepção, entre elas a garantia organizacional.

Quando um vocabulário é publicado, há de se questionar não somente o seu desempenho atual, mas qual será seu nível de relevância a longo prazo e quais garantias asseguram sua continuidade. É nesta perspectiva que avaliação de vocabulários controlados ganha um ponto de reflexão teórica, aliada a instrumentalização metodológica. Neste sentido, Barité (2007) destaca a importância das garantias, que podem ser literárias, do usuário, cultural e organizacional. Martínez-Ávila e Budd (2017) também destacam a importância da garantia epistêmica, para o desenvolvimento de vocabulários controlados, a garantia epistêmica se limita às reivindicações ônticas de enunciados, além de avaliar as alegações de conhecimento.

A Garantia organizacional atrelada à usabilidade e desempenho dos vocabulários controlados oferece um caminho promissor na condução de pesquisas dedicadas à avaliação de vocabulários. A National Information Standards Organization (2017) ressalta que a questão da sustentabilidade dos vocabulários é protegida por compromissos organizacionais ou institucionais, políticas que deixam claro quem faz esses compromissos e o que eles significam, bem como um registro de manutenção e crescimento responsável. Outro fator que merece destaque é a documentação e o controle de versão dos vocabulários.

Um vocabulário sem esses compromissos não será sustentável ao longo do tempo e pode ser um investimento questionável para as organizações, pois com o tempo vocabulários insustentáveis tendem a se tornar órfãos (National Information Standards

Organization, 2017). Um vocabulário controlado que se torna órfão tende a desaparecer devido à ausência de documentação ou registros que contextualizam sua origem, objetivos e meios de preservação. Um vocabulário órfão significa sobretudo um enfraquecimento dos compromissos institucionais e da relevância destes tipos de SOCs.

Haas et al (2008) abordando sobre um vocabulário controlado da área médica explicita que a manutenção exige um comprometimento a longo prazo. Pois o desenvolvimento e a manutenção devem incluir meios de validar, avaliar e manter um vocabulário resultante; o papel das organizações e das partes interessadas no design de vocabulário se beneficia de lições aprendidas em outras pesquisas sobre vocabulários controlados. Zaharee (2012) cita um plano de manutenção como parte do desenvolvimento de vocabulários controlados. E ressalta que um vocabulário nunca é concluído, a menos que não seja mais usado.

A avaliação permite identificar erros de ortografia e sinônimos comuns que poderiam ser adicionados e determinar quais pesquisas não estão produzindo nenhum resultado e podem ser melhoradas (Ryan, 2014). Uma avaliação pode englobar diversos aspectos, mas há de se considerar que os dados que estão agrupados em vocabulários controlados possuem importantes aplicações que podem revelar suas principais características.

A avaliação por meio da qualidade dos dados também pode ser um importante fator de mudança na forma como os termos são percebidos, enquanto tipo de dados. A avaliação de vocabulários controlados é uma tarefa difícil e complexa, por isso, nesta pesquisa ela é composta por três etapas distintas, que consistem na análise estrutural quantitativa, baseada em modelos de dados e linguística via NLP. Porém, o processo de avaliação começa desde a seleção dos vocabulários controlados, tomando como base os requisitos gerais apresentados.

4.1 Análise preliminar

A análise dos vocabulários controlados revelou-se complexa, devido a vários motivos. O primeiro deles refere-se ao fato de utilizar a BARTOC como referência de vocabulários indexados para seleção prévia. Os filtros de buscas aplicados para o tipo de SOC

selecionado foram apenas de tesauros, para língua portuguesa e o assunto tem como base o Dewey Decimal Classification. A priori foram retornados 46 resultados, o que não significa que todos são tesauros brasileiros, porque há vários tesauros multilíngues que possuem termos em português.

Ao analisar previamente esse resultado foi constatado que muitos vocabulários não estavam no formato RDF ou SKOS, mas em formatos diversos, outros estavam inativos, foram descontinuados ou desatualizados. Para ser mais preciso apenas 39,13% (18 vocabulários) estavam no formato RDF ou SKOS, porém nem todos estavam definitivamente acessíveis para análise completa, porque alguns utilizavam softwares como o Tematres que permite apenas exportações parciais em SKOS, quando não há um SPARQL Endpoint ou a disponibilização do vocabulário para download.

Alguns vocabulários estavam em formatos de leitura de texto como *Portable Document Format* (PDF), 15,21% (7 vocabulários) Embora alguns tesauros também estivessem disponíveis em outros formatos, como foi o caso do Tesauro Brasileiro de Ciência da Informação (TBCI) que na versão em PDF foi publicado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e na Versão em SKOS publicado pela Universidade Estadual de Londrina (UEL), a BARTOC indexa apenas a versão em PDF, portanto será considerada apenas essa versão neste estudo. O VCGE também está disponível em RDF/SKOS e outros formatos. Também há casos em que estes tesauros estão desatualizados há mais de 20 anos, como o *European multilingual thesaurus on AIDS and HIV infection*, cuja versão mais recente é de 1999. O que possibilita inclusive classificá-lo como um vocabulário órfão .

Outros tesauros estavam em formatos diversos 17,39% (8 tesauros), não sendo possível identificar com clareza alguns deles pois estão embutidos em sistemas próprios ou de sistemas de busca como foi o caso *Thesaurus of Clinical Sign*s (Integrado ao sistema de busca *Orphanet*) e do *The Brazilian Thesaurus of Education* (Brased, integrado ao sistema Pergamum). Alguns disponibilizam para download no formato CSV ou JSON. Embora seja possível realizar análise tanto em CSV quanto JSON utilizando bibliotecas de programação em *Python* como *Numpy* e *Pandas* esses formatos serão ignorados nesta pesquisa.

Por fim, identificou-se que 26,08% (12 tesauros) estavam inativos, descontinuados ou simplesmente não localizados, pois o link de acesso a estes estão quebrados ou são redirecionados para outras páginas. Por não estarem acessíveis ou

atualizados esses tesauros podem ser considerados como órfãos e mesmo indexados na BARTOC com um breve resumo não há como obter informações mais completas sobre eles, exceto se possuírem registros na literatura científica. Logo, percebe-se que mais da metade dos tesauros localizados 60,86% (28 tesauros) não estão no formato SKOS ou RDF.

Dentre os vocabulários apresentados nesta categoria, foi possível verificar, que alguns estão disponíveis para download e outros estão parcialmente disponíveis. Isto quer dizer, que tesauros como o TBCI disponibilizado pela UEL e o Tesauro da UNESP estão disponíveis para exportação no formato SKOS, mas só é possível fazer isto de modo parcial, à medida que o usuário está navegando pelo tesauro. Também foi identificado que 5 destes tesauros são brasileiros e monolíngues. Neste sentido, também é possível constatar que o tesauro mais antigo nessa categoria é o AGROVOC, publicado em 1980. O que traduz a perenidade desse vocabulário até os dias atuais em padrões contemporâneos à realidade tecnológica.

Algumas limitações foram identificadas nesses vocabulários controlados, principalmente no que diz respeito à análise preliminar, pois em alguns casos não é possível verificar tesauro fora do ambiente de gestão. Uma opção viável para estes casos seria um ponto de consulta *SPARQL*, todavia não são todos os tesauros que possuem esse recurso disponível. Também nota-se que muitos vocabulários foram publicados após os anos 2000, o que permite inferir que há uma certa infraestrutura tecnológica básica que permite publicar vocabulários controlados em modelos de dados mais modernos.

Apenas com base nesta análise preliminar é possível realizar estudos que englobam a avaliação de tesauros em nível nacional, continental e mundial com intuito de identificar apenas vocabulários que estão em formatos de dados linkados, ou que estão órfãos. Com base apenas nos requisitos de disponibilidade, atualização, documentação e padrões reconhecidos, neste caso SKOS RDF e suas variações Turtle, N3, NQuad. Por fim, foram selecionados 10 vocabulários controlados, pois foram os únicos que estavam totalmente disponíveis para download e permitem análises completas.

Existem vocabulários controlados que possuem um número pequeno de termos e consequentemente o tamanho do arquivo está na casa dos megabytes, enquanto há vocabulários com milhares de termos cujo tamanho do arquivo ultrapassa os Gigabytes. O tamanho do arquivo pode ser um limitador para computadores com baixo poder de

processamento e pouca memória RAM. Por exemplo, na etapa de análise baseada em modelos de dados, o AGROVOC supera 800 MB em sua versão simplificada. Embora seja admirável a quantidade de termos listados nesse tesauro é necessária utilização de computadores com no mínimo 12GB de RAM para essa quantidade de dados, esta dificuldade aumenta se houver alguma função complexa de análise, por exemplo, comparações e cruzamento de dados. Diferente da análise estrutural quantitativa, que permite verificar basicamente qualquer arquivo RDF ou SKOS de forma simplificada.

Para melhor identificar os vocabulários controlados que estão em formatos de RDF ou SKOS, o quadro 07 apresenta os vocabulários controlados analisados, que atenderam um simples requisito, que se resume em estar disponível por completo para download. Isto significa que estes vocabulários podem ser reutilizados totalmente ou parcialmente, como foi o caso do LandVoc um sub-vocabulário do AGROVOC sobre conceitos relacionados à governança da terra licenciado sob uma atribuição Creative Commons (CC BY 3.0 IGO).

Quadro 07 - Vocabulários controlados em RDF/SKOS

Nº	**TESAURO**	**TIPO**	**ANO DE PUBLICAÇÃO**
1	AGROVOC	Multilíngue	1980
2	CAB Thesaurus	Multilíngue	1983
3	Thesaurus Multilingue da União Europeia (EuroVoc)	Multilíngue	1984
4	General Multilingual Environmental Thesaurus (GEMET)	Multilíngue	2001
5	Vocabulário Controlado do Governo Eletrônico (VCGE)	Monolíngue	2004
6	International Coastal Atlas Network Coastal Erosion Global Thesaurus	Multilíngue	2011
7	Europeana Fashion Vocabulary	Multilíngue	2012
8	Partage Plus Vocabulary	Multilíngue	2012
9	Resource Type Vocabulary	Multilíngue	2015
10	LandVoc	Multilíngue	2017

Fonte: Elaborado pelo autor

Para contextualizar esses tesauros no escopo de análise geral é preciso conhecer um pouco mais sobre cada um, começando pelo mais antigo destes, o AGROVOC um vocabulário controlado que cobre todas as áreas de interesse da Organização das Nações Unidas para Agricultura e Alimentação (*The Food and Agriculture Organization* - FAO), incluindo alimentação, nutrição, agricultura, pesca, silvicultura, meio ambiente, etc. O AGROVOC é publicado pela FAO e editado por uma comunidade de especialistas desde o início da década de 1980 e passou de catálogos impressos para tecnologias da web semântica. O AGROVOC consiste em mais de 39.000 conceitos disponíveis para download, acesso via Web Services ou SPARQL endpoint.

O CAB Thesaurus é o segundo vocabulário controlado mais antigo desta lista, que está em uso constante desde 1983. Ele faz parte dos bancos de dados CAB *Abstracts*™ e *Global Health* com abrangência terminológica de nomes de plantas, animais e microorganismos. O CABI é uma organização sem fins lucrativos de desenvolvimento e informação baseada em ciência.

O *Thesaurus Multilingue da União Europeia* (EuroVoc) é o terceiro mais antigo, o EuroVoc é um tesauro que foi originalmente criado para processar a informação documental produzida pelas instituições da UE em 23 línguas. Trata-se de um thesaurus pluridisciplinar que abrange domínios suficientemente amplos para abranger tanto o ponto de vista comunitário como o nacional, com uma certa ênfase na atividade parlamentar. O EuroVoc é gerenciado pelo Escritório de Publicações da União Européia, que avançou com o gerenciamento de tesauros baseado em ontologia e tecnologias de web semântica em conformidade com as recomendações do W3C, bem como as últimas tendências nos padrões de tesauros.

O *General Multilingual Environmental Thesaurus* (GEMET) foi desenvolvido desde 1995 como uma ferramenta de indexação, recuperação e controle para o *European Topic Centre on Catalog of Data Sources* (ETC/CDS) e a *European Environment Agency* (EEA), Copenhagen. A ideia básica para o desenvolvimento do GEMET foi usar o melhor dos excelentes tesauros multilíngues atualmente disponíveis, a fim de economizar tempo, energia e dinheiro por isso ele é resultado da fusão de pelo menos 8 tipos de vocabulários controlados distintos, o "*Umwelt Thesaurus" of Umweltbundesamt* (UBA), o *Thesaurus Italiano per l'Ambiente* (TIA), o *Multilingual Environment Thesaurus* (MET), o *EnVoc Thesaurus*, o *Thesaurus de Medio Ambiente*, o *Lexique environnement - Planète*, *Descriptors*

of relevant documents of the EEA, denominado "*Europe's Environment* e o *Eurovoc*. O GEMET começou a ser desenvolvido em 1995, mas foi publicado em 2001 e desde então é atualizado, sendo a última versão do ano de 2021.

O Vocabulário Controlado do Governo Eletrônico (VCGE) é um vocabulário brasileiro que começou com uma lista de termos no governo federal de âmbito geral depois passou para uma Lista de Categorias de Governo – LCG – em março de 2004, como uma lista que contempla todos os assuntos relacionados com a atuação de Governo e dois anos depois mudou para Lista de Assuntos de Governo – LAG – , com um foco em taxonomia de navegação, finalmente em 2010 surge como o VCGE – Vocabulário Controlado e Governo Eletrônico.

O *International Coastal Atlas Network Coastal Erosion Global Thesaurus* (ICANCEGT) é gerenciado por meio do NERC *Vocabulary Server* (NVS) e mantido pelo *British Oceanographic Data Centre no National Oceanography Centre* (NOC) em Liverpool e Southampton, e recebe financiamento do *Natural Environment Research Council* (NERC) no Reino Unido. Este tesauro foi provavelmente publicado no ano de 2011 segundo informações extraídas da própria BARTOC. Possui termos em 5 idiomas abrangendo as áreas de Ciências, Geologia, Hidrologia e Meteorologia. É possível acessá-lo via download direto, SPARQL endpoint ou APIs RESTful.

O *Europeana Fashion Vocabulary* (EFV) é baseado no *Art and Architecture Thesaurus* (AAT) do *Getty Research Institute* e abrange conceitos relacionados à moda. O EFV surgiu de um projeto cofinanciado pela Comissão Europeia que começou em 2012 e durou 36 meses.

O *Resource Type Vocabulary* é um tesauro mantido pela *Confederation of Open Access Repositories* (COAR), uma associação internacional com 157 membros e parceiros de todo o mundo representando bibliotecas, universidades, instituições de pesquisa, financiadores governamentais e outros. O COAR possui outros vocabulários, todos disponibilizados sob uma licença *Creative Commons*.

O *Linked Land Governance Thesaurus* (LandVoc) é um vocabulário controlado que abrange conceitos relacionados à governança fundiária, que foi desenvolvido como um subvocabulário do AGROVOC gerenciado sobre a curadoria da *Land Portal Foundation* e hospedado pela FAO. Embora esteja disponível em RDF, esse tesauro compartilha o mesmo SPARQL *Endpoint* do AGROVOC.

Por fim, o *Partage Plus Vocabulary* é um vocabulário controlado multilíngue estruturado para objetos de patrimônio cultural relacionados à produção estética referenciando e baseado no *Art & Architecture Thesaurus* do *Getty Research Institute*. O *Partage Plus Vocabulary* contém 705 conceitos representados por rótulos em 16 idiomas, esse vocabulário está disponível para reutilização gratuita. Sendo possível realizar o *download* ou navegar por uma interface de navegação. Não há informações precisas sobre a data de publicação deste tesauro, embora o Projeto Partage Plus tenha sido financiado pelo Programa de Apoio às Políticas de TIC da Comissão Europeia iniciado em 2012 com duração de duração de 24 meses.

De acordo com esse método de pré-seleção de vocabulários é possível identificar uma característica comum entre eles, todos possuem algum tipo de informação sobre sua origem e manutenção. Documentos que indicam o contexto de surgimento e uso do vocabulário que transmitem uma identidade e registram a história do vocabulário controlado ou os processos que culminaram no seu surgimento, assim como dos colaboradores e financiadores. Neste sentido, as análises seguintes são focadas nos três tipos definidos nesta pesquisa, a primeira delas é a análise estrutural quantitativa, explanada no tópico a seguir.

4.2 Análise estrutural quantitativa

A análise estrutural quantitativa é uma contagem de tags, focada em nas propriedades descritivas dos vocabulários controlados, pois nem sempre um tesauro possuirá um número equivalente de termos para conceitos, ou de termos para notas de definição. Também há possibilidade de um vocabulário controlado possuir muitos termos alternativos, devido ao fato de originar-se de um dicionário de sinônimos. Para iniciar esta análise foi criado um código escrito na Linguagem Python, cuja tarefa se resume a ler o arquivo em RDF, realizar uma contagem e apresentá-la, depois disso, essa contagem listada foi salva em arquivo CSV para elaboração dos gráficos.

O processo de avaliação de vocabulários controlados pode se tornar um trabalho árduo, primeiro por se tratar de programação, que para muitos bibliotecários parece algo difícil, segundo porque o RDF não é um modelo de dados tão conhecido no ambiente de programação, por isso há pouco suporte da comunidade de desenvolvedores que relatam *bugs* de execução. Porém há uma biblioteca em Python, específica para RDF, o que de certo modo facilitou o trabalho para quem possui conhecimentos de lógica de programação, essa

biblioteca se chama RDFLib, que é um pacote em *Python* para analisar e serializar RDF/XML, N3, NTriples, N-Quads, Turtle, TriX, Trig, JSON-LD e suporte a consultas SPARQL.

O fato de ser pouco conhecido no ambiente de programação faz do RDFLib uma biblioteca com uma extensa documentação, porém, de moderado nível de dificuldade, pois as análises feitas nesta pesquisa com uso dessa biblioteca é apenas uma prova superficial de como ela pode ser utilizada para criar e analisar vocabulários controlados. O ambiente de execução do código é o Jupyter notebook, que surgiu de um projeto de código aberto sem fins lucrativos, nascido do Projeto IPython em 2014, que oferece suporte à ciência de dados interativa e à computação científica em todas as linguagens de programação, o Jupyter é um software de código aberto e gratuito.

O *Jupyter* foi utilizado no ambiente de colaboração do *Google Research*, chamado *Collaboratory*, ou "*Colab*" do *Google Workspace* acessado com e-mail institucional, o *Google Colab* permite que qualquer pessoa escreva e execute um em código python por meio do navegador e é recomendado para aprendizado de máquina, análise de dados e educação tecnológica. O *Google Colab* é um serviço hospedado que não requer configuração para uso e fornece acesso a recursos de computação, incluindo GPUs. O *Google Colab* foi selecionado como ambiente de execução porque é um ambiente que foi projetado para este fim e não possui as limitações de um computador pessoal, que muitas vezes precisa compartilhar a memória RAM com outras tarefas enquanto o código é executado, o ambiente possuía 12.7 GB de RAM e 107.7 GB de armazenamento de disco.

Para realizar a contagem das propriedades descritivas dos vocabulários controlados utilizou-se o pacote de bibliotecas em python denominado *RDFLib*, que se beneficia dos XMLs Namespaces de cada vocabulário para identificar cada tag e realizar a contagem. Neste sentido, o código foi dividido em 7 etapas que consistem em:

1 Criar um Graph(), que é uma coleção tripla do subject (s), predicate (p) e object (o);

2 Carregar o arquivo RDF no gráfico;

3 Criar um dicionário vazio para armazenar as contagens de tags;

4 Verificar por todos as triplas subject (s), predicate (p) e object (o) no gráfico;

5 Obter a tag do predicado (p);

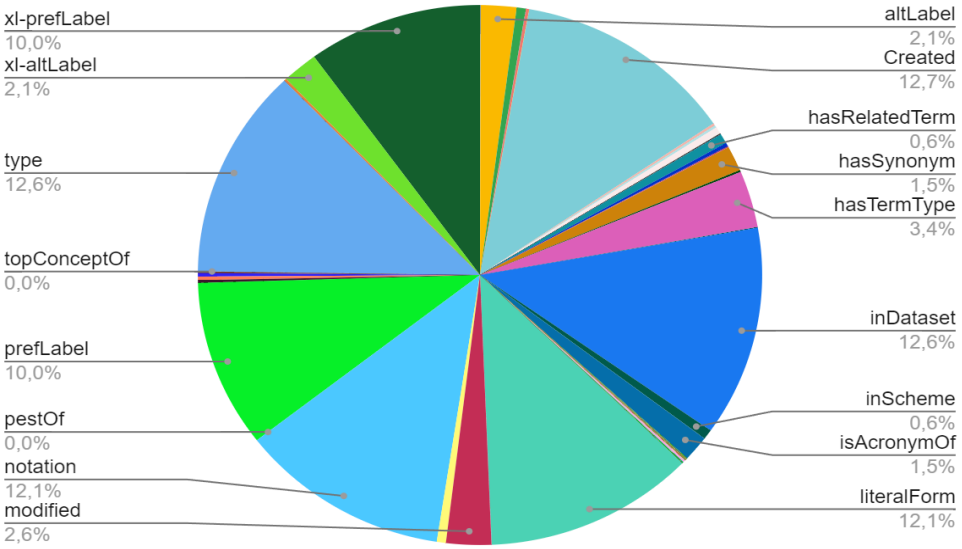
6 Adicionar a contagem das tags no dicionário;

7 Imprimir as contagens de tags.

Após a estruturação do código a análise foi realizada inserindo o link que apontava diretamente para o arquivo em RDF, como foi o caso do AGROVOC e EUROVOC, devido ao tamanho do arquivo ou realizando o download na página do vocabulário controlado e fazendo o upload do arquivo no ambiente de execução. Para validar essa contagem automática, um vocabulário de teste foi utilizado, nesse caso o International Coastal Atlas Network Coastal Erosion Global Thesaurus, por ser o menor em tamanho de arquivo, portanto, mais fácil de identificar um erro. Além desta contagem automática, foi realizada uma contagem unitária por meio do filtro de contagem do NotePad++ com o uso da *tag* prefLabel, resultando em 4486 correspondências no arquivo, mas sabendo que essa tag é duplicada na descrição de item, por exemplo, "<prefLabel>Termos<prefLabel>", é necessário dividir por 2, que resultam em 2.243 registros. Para *tags* únicas, que não precisam ser divididas por 2, como é o caso de hasTopConcept, obteve-se o resultado de 7 registros . Ambos os resultados foram iguais no código em *python* apresentado, o que oferece uma certa segurança na apresentação dos dados.

De posse dessas informações, a seguir serão apresentados as análises individuais de cada tesauro, começando pelo AGROVOC. A análise estrutural quantitativa desse vocabulário controlado identificou 8.016.469 de registros divididos em 129 propriedades oriundas de 17 XML-Namespaces diferentes, como DC, SKOS, RDF, FOAF, etc. Para ilustrar melhor essa representação, a figura 01 apresenta a composição do AGROVOC segundo as suas propriedades mais relevantes segundo de acordo com a proporção do vocabulário controlado.

Figura 01 - Representação estrutural quantitativa do AGROVOC



Fonte: Dados da pesquisa

O destaque do AGROVOC é direcionado pela tag do Dublin Core (DC) denominada dct:created que se destina ao registro da data de criação de um recurso, com 1.019.731 de registros. A tag rdf:type que é é uma instância de rdf:Property usada para indicar que um recurso é uma instância de uma classe é a segunda mais expressiva deste vocabulário controlado com 1.008.227 registros. A tag inDataset pertence ao XML-NS do Vocabulary of Interlinked Datasets (VoID) e representa um conjunto de triplos RDF que são publicados, mantidos ou agregados por um único provedor. A quarta tag mais expressiva é skosxl:literalForm, que é usada para fornecer a forma literal de uma tag skosxl:Label, conforme ilustrado na figura 02.

Figura 02 - Tag skosxl:literalForm

```
<rdf:Description rdf:about="http://aims.fao.org/aos/agrovoc/xl_fr_2a7d8f77">
    <skosxl:literalForm xml:lang="fr">culture de plantes aquatiques</skosxl:literalForm>
</rdf:Description>
```

Fonte: Dados da pesquisa

O uso de tags skos-xl em alguns casos pode interferir em uma análise estatística, pois é provável que um mesmo termo esteja em diferentes posições, seja ela skosxl:literalForm e skos:prefLabel ao mesmo tempo, como identificado na figura 03. Porém

skosxl:literalForm também pode ser uma subpropriedade de skos:prefLabel e uma instância de owl:Class segundo Miles e Bechhofer (2009).

Figura 03 - Tags skosxl:literalForm e skos:prefLabel

```
C:\Users\janai\Downloads\agrovoc_lod.rdf (2 ocorrências)
  Linha 13483584:     <skosxl:literalForm xml:lang="fr">culture de plantes aquatiques</skosxl:literalForm>
  Linha 16599734:     <skos:prefLabel xml:lang="fr">culture de plantes aquatiques</skos:prefLabel>
```

Fonte: Dados da pesquisa

Outra propriedade expressiva é a tag representada por skos:notation, definida como uma cadeia de caracteres utilizada para identificar um conceito, a notação difere de um rótulo lexical porque não é representada em linguagem natural conforme a **figura 04**. Diferente de skos:note, que é representada em linguagem natural.

Figura 04 - Tag skos:notation

```
<rdf:Description rdf:about="http://aims.fao.org/aos/agrovoc/xl_fr_2a7d8f77">
    <skos:notation rdf:datatype="http://aims.fao.org/aos/agrovoc/AgrovocCode">a1a5dac2</skos:notation>
</rdf:Description>
```

Fonte: Dados da pesquisa

Se o uso de tags skosxl:prefLabel fosse somado com as tags prefLabels independente do XML Schema, os termos preferidos *prefLabel,* representam 20% do vocabulário com 1.605.324 registros, por isso foi adicionado uma observação na contagem, porque a tag *skosxl:prefLabel* foi identificada com uma extensão utilizada para instâncias de classe, conforme a **figura 05,** com propriedades análogas as propriedades skos:prefLabel, porque ela não possui as mesmas características de uma prefLabel, que possui termos em Linguagem Natural.

Figura 05 - Tags skosxl: prefLabel

```
<skosxl:prefLabel rdf:resource="http://aims.fao.org/aos/agrovoc/xl_tr_ae72e971"/>
<skosxl:prefLabel rdf:resource="http://aims.fao.org/aos/agrovoc/xl_en_ebf613d2"/>
<skosxl:prefLabel rdf:resource="http://aims.fao.org/aos/agrovoc/xl_la_61c9564d"/>
<skosxl:prefLabel rdf:resource="http://aims.fao.org/aos/agrovoc/xl_es_b72a73ed"/>
<skosxl:prefLabel rdf:resource="http://aims.fao.org/aos/agrovoc/xl_fr_741eb019"/>
<skosxl:prefLabel rdf:resource="http://aims.fao.org/aos/agrovoc/xl_pt_c23e15ad"/>
```

Fonte: Dados da pesquisa

A tag prefLabel possui 802.662 registros, entretanto, há de se ressaltar que este é um vocabulário controlado multilíngue com termos em 41 idiomas distintos em constante evolução. Por isso, há uma pequena divergência no resultado total apresentado pelo portal de navegação do AGROVOC utilizando o SKOSMOS do arquivo disponível para download. Ao clonar a tabela com os termos divididos por idioma foi realizada a soma total dos 41 idiomas que representam 802.285 registros, menos da metade do total apresentado. Segundo um documento produzido pela FAO (2022) o AGROVOC em 2021 consistia em mais de 39.000 conceitos e 925 000 termos em até 41 idiomas, ou seja, já se passaram mais de 2 anos até o ano de 2023, que esses dados foram apresentados e nitidamente esses números aumentaram.

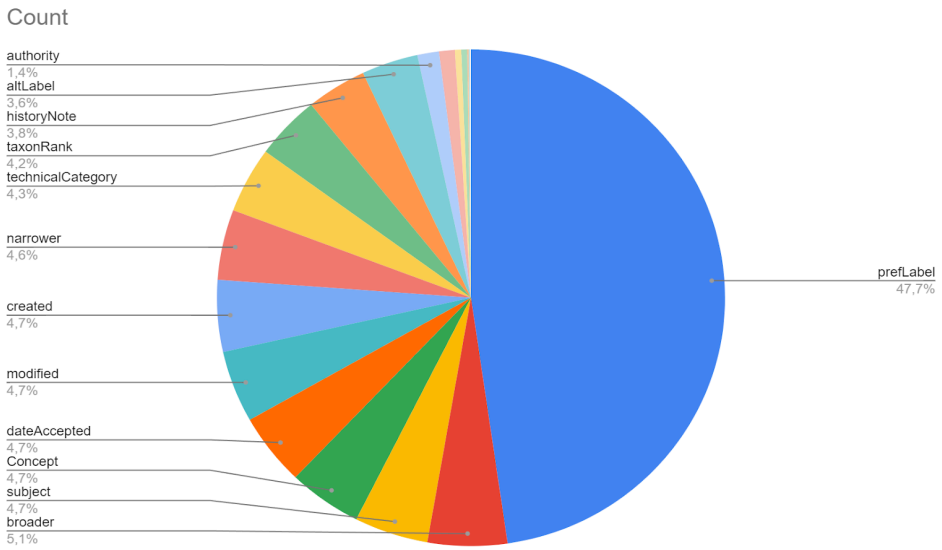
A *tag modified* é uma propriedade *Dublin Cor*e que se refere a em que o recurso foi alterado. Esta tag possui 206.868 registros que indicam uma possível taxa de alteração ou atualização do vocabulário desde que foi disponibilizado neste modelo de dados. A tag altLabel representa apenas 2,1% do AGROVOC mesmo possuindo 164.499,00 registros a tag skosxl:altLabel segue a mesma lógica das tag skosxl:prefLabel. As tags hasTermType (272.192 registros), isAcronymOf (121.577 registros) e hasSynonym (120.179 registros) são partes da *agrontology* que fornece propriedades específicas de domínio para enriquecer a descrição de conceitos, por meio de descritores *Vocabulary of a Friend* (VOAF), para vinculá-los para AGROVOC.

Não foi possível representar graficamente todo o vocabulário porque seria 129 fatias no gráfico, o que o tornaria ilegível, por isso não é representado nenhuma tag com menos de 50.000 registros. Porém, há tags com apenas 1 registro, como propertyNumber, que pertence ao VOAF estabelece o número de propriedades definidas no namespace do vocabulário. Ou seja, mesmo reconhecendo a importância desse elemento de descrição, ele possui pouca expressividade diante do tamanho do AGROVOC. Outro ponto que vale destacar é a dimensão que esse vocabulário possui, em tamanho, número de registros e abrangência de termos. Por isso é iminente que uso de técnicas de análise automatizada façam parte dos estudos de Terminologia e Organização do conhecimento.

Seguindo a ordem temporal apresentada, o *CAB Thesaurus* é o próximo vocabulário controlado a ser analisado. O CAB Thesaurus possui 186.417 registros divididos em 23 propriedades de descrição por meio de 6 XML-NS diferentes. **A figura 06** apresenta a composição desse vocabulário com as 14 propriedades mais relevantes diante da proporção

do tesauro, ou seja que possuíam mais de 2.000 registros que representam um pouco mais de 1% deste vocabulário controlado.

Figura 06 - Representação estrutural quantitativa do CAB Thesaurus



Fonte: Dados da pesquisa

A tag com maior destaque no CAB Thesaurus é prefLabel com 88.912 registros que representam quase metade deste tesauro. Porém, há uma excentricidade, porque mesmo se tratando de um tesauro multilingue que apresenta termos em inglês e outros 10 idiomas, a expressão de Nehru, quando afirma que a ciência não pertence a um país específico (ÍNDIA, 1968). Logo, um termo receberia a mesma designação em diferentes línguas, e parece que é o que ocorre com o CAB Thesaurus, conforme a figura 07.

Figura 07 - Repetição de termos do CAB Thesaurus

```
<skos:Concept rdf:about="http://id.cabi.org/cabt/4036">
<skos:prefLabel xml:lang="en-GB">Abacarus</skos:prefLabel>
<txn:taxonRank>Genus</txn:taxonRank>
<skos:broader rdf:resource="http://id.cabi.org/cabt/Eriophyidae" />
<skos:narrower rdf:resource="http://id.cabi.org/cabt/Abacarus%20gossypii" />
<skos:narrower rdf:resource="http://id.cabi.org/cabt/Abacarus%20hystrix" />
<dcterms:subject>Organism Names</dcterms:subject>
<cabi:technicalCategory>Scientific Name - Organisms</cabi:technicalCategory>
<skos:prefLabel xml:lang="da">Abacarus</skos:prefLabel>
<skos:prefLabel xml:lang="de">Abacarus</skos:prefLabel>
<skos:prefLabel xml:lang="fr">Abacarus</skos:prefLabel>
<skos:prefLabel xml:lang="it">Abacarus</skos:prefLabel>
<skos:prefLabel xml:lang="nl">Abacarus</skos:prefLabel>
<skos:prefLabel xml:lang="no">Abacarus</skos:prefLabel>
<skos:prefLabel xml:lang="pt">Abacarus</skos:prefLabel>
<skos:prefLabel xml:lang="es">Abacarus</skos:prefLabel>
<skos:prefLabel xml:lang="fi">Abacarus</skos:prefLabel>
<skos:prefLabel xml:lang="sv">Abacarus</skos:prefLabel>
```

Fonte: Dados da pesquisa

Os termos preferidos se repetem ao longo de todo o vocabulário, o que significa dizer que, podemos dividir o número de registros por 11 para obter o número exato de termos únicos para esta tag. No entanto, essa divisão seria imprecisa, pois o resultado seria 8.082,909090909091, o que significa dizer que não há uma distribuição igual por idioma. Ao aplicar um filtro de busca pelo idioma principal do tesauro, com expressão <skos:prefLabel xml:lang="en-GB">, obteve-se 8.725 termos preferidos. Entende-se que este seja o número real de tags prefLabel. Na página do CAB Thesaurus o número estimado de termos superiores seria 3,1 milhões, enquanto os nomes de plantas, animais e microorganismos seriam de 287.800. No entanto, o total de todos os registros de tags somam apenas 186.417, se o CAB thesaurus possuísse por volta de 3 milhões de conceitos ele seria maior que o AGROVOC, em linhas e tamanho de arquivo, porém o CAB Thesaurus possui apenas 11,6 MB, enquanto AGROVOC mais de 830 MB na sua versão em NT e 1,2 GB em RDF.

A segunda tag mais relevante é broader com 9.430 registros, utilizada para declarar que um conceito é mais amplo, ou seja, mais geral. E em terceiro lugar a tag *subject* do DC namespace com 8.769 registros, que representam um assunto do recurso descrito com um valor literal. As tags Concept, dateAccepted e type possuem o mesmo número de registros 8.725. O que faz total sentido para o uso da tag concept, tendo em vista que esta tag representa um conceito por meio de uma URI, que normalmente está associada a um termo, como é possível constatar na figura 08.

Figura 08 - URI do termo abacarus do CAB Thesaurus

Fonte: Dados da pesquisa

A tag dateAccepted refere-se à data de aceitação de um recurso descrito com o DC namespace, ela pode divergir de outras duas propriedades de tempo, como as tags created com 8.709 registros e modified com 8.688 registros, que representam as datas de criação e modificação respectivamente, como é possível verificar na figura 09.

Figura 09 - Tags de datas Dublin Core do CAB Thesaurus

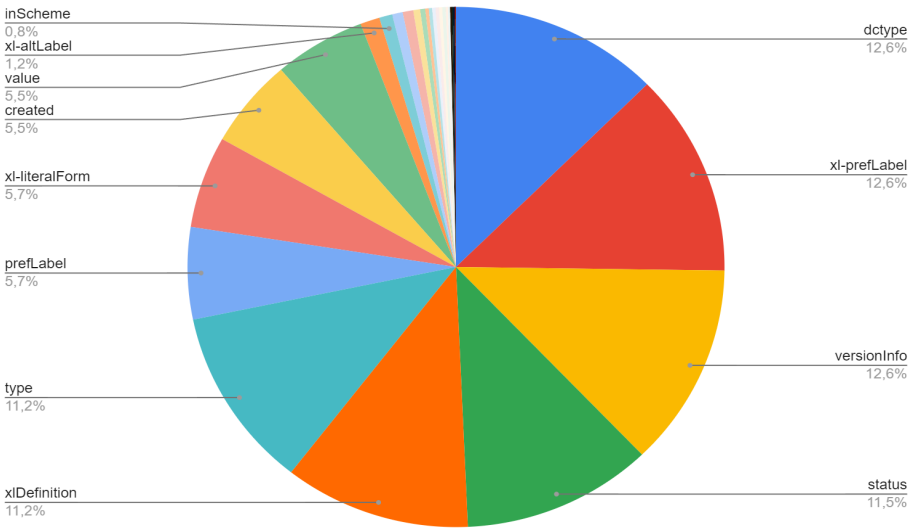
```
<dcterms:created>2004-08-03</dcterms:created>
<dcterms:dateAccepted>2008-08-19</dcterms:dateAccepted>
<dcterms:modified>2012-05-01</dcterms:modified>
```

Fonte: Dados da pesquisa

A tag *narrower* possui 8534 registros, com menos conceitos específicos declarados, diferente da tag broader. A propriedade *technicalCategory* (7.980 registros) é uma tag que pertence ao XML-NS da própria CABI, que é responsável pelo CAB-Thesaurus. Enquanto o taxonRank possui 7.744 registros, esta tag pertence a uma ontologia de conceitos taxonômicos denominada Taxonconcept. A historyNote (7164 registros) é uma tag que também está associada aos registros de tempo utilizando declarações de versionamento do editor do vocabulário. A tag altLabel possui 6.708 registros, enquanto authority que também pertence a Taxonconcept tem 2.565 registros.

O Thesaurus Multilingue da União Europeia (EuroVoc) é o terceiro vocabulário a ser analisado e possui 3.662.023 registros divididos em 37 propriedades de descrição com 6 namespaces diferentes. Para ilustrar a composição do EuroVoc, a figura 10 representa as principais tags com maior índice de registros nesse vocabulário controlado.

Figura 10 - Representação estrutural quantitativa do Eurovoc



Fonte: Dados da pesquisa

A *tag type* do DC namespace possui 461.726 registros, sendo umas das propriedades mais utilizadas nesse vocabulário, seguido das tags xl-prefLabel pertencente a SKOSXL namespace com 461.598 registros, *versionInfo* da OWL namespace com 461.597 registros, a *status* da EUROVOC Ontologies com 420.227 registros, xlDefinition com 411.242 registros e RDF type com 411242 registros. Somente essas 6 tags representam 70% das propriedades de descrição desse vocabulário. A tag prefLabel possui 209.234 registros divididos em 23 idiomas diferentes, enquanto a tag xl-literalForm também possui 209.234, não por mera coincidência, conforme apresentado na figura 11.

Figura 11 - Repetições de tags no Eurovoc

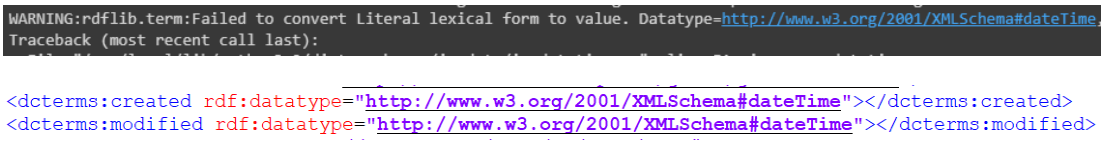
```
C:\Users\janai\Downloads\Vocabularios RDF\eurovoc.rdf (4 ocorrências)
  Linha     3791:    <prefLabel xmlns="http://www.w3.org/2004/02/skos/core#" xml:lang="en">short-term financing</prefLabel>
  Linha    13932:    <literalForm xmlns="http://www.w3.org/2008/05/skos-xl#" xml:lang="en">short-term financing</literalForm>

C:\Users\janai\Downloads\Vocabularios RDF\eurovoc.rdf (4 ocorrências)
  Linha     3799:    <prefLabel xmlns="http://www.w3.org/2004/02/skos/core#" xml:lang="it">finanziamento a breve termine</prefLabel>
  Linha     4158:    <altLabel xmlns="http://www.w3.org/2004/02/skos/core#" xml:lang="it">finanziamento a breve termine in pool</altLabel>
  Linha    13950:    <literalForm xmlns="http://www.w3.org/2008/05/skos-xl#" xml:lang="it">finanziamento a breve termine</literalForm>
  Linha   134134:    <literalForm xmlns="http://www.w3.org/2008/05/skos-xl#" xml:lang="it">finanziamento a breve termine in pool</literalForm>
```

Fonte: Dados da pesquisa

Percebe-se que o uso do mesmo termo em tags diferentes é comum, principalmente as tags que suportam caracteres em linguagem natural, embora não seja uma regra a distribuição exata de termos por tags, a alLabel por exemplo, possui apenas 23.308, e também repete um termo xl-literalForm, o que significa que nem todos os termos de uma determinada tag é repetido. A tag created possui 202.008 registros de datas, tag value possui 202.008 registros, a xl-altLabel tem 42.952, que não são necessariamente literais, mas podem ser URIs também e a tag inScheme possui apenas 28.848, representando menos de 1% da composição do Eurovoc.

Seguindo a listagem, o General Multilingual Environmental Thesaurus (GEMET) é o próximo tesauro a ser analisado, ele possui 323.635 registros, divididos entre 34 tags e 13 xml-ns diferentes. Durante a análise desse vocabulário foi identificado um erro, na conversão das tags que utilizavam que utilizavam as propriedades dateTime do XML Schema conforme a figura 12.

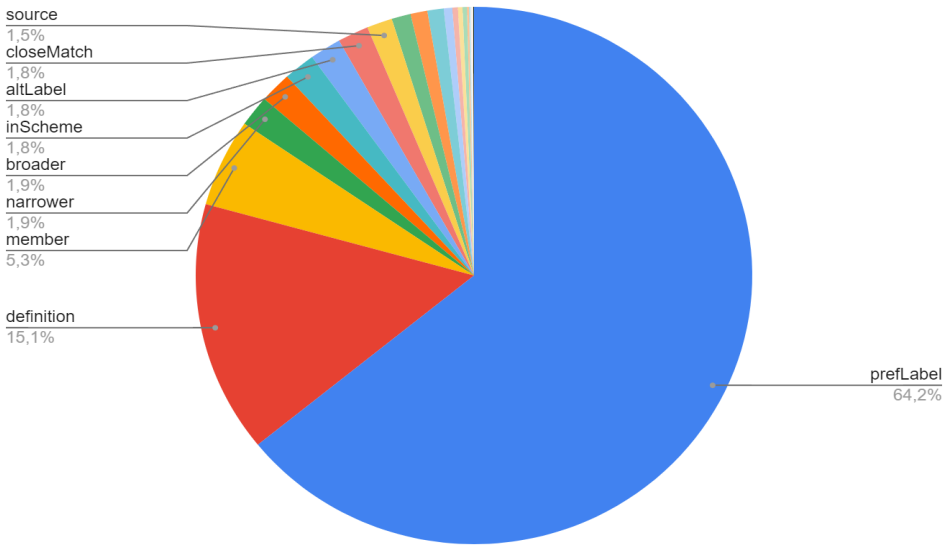
Figura 12 - Erros de dateTime do XML Schema



Fonte: Dados da pesquisa

Este erro afetava as tags created e modified, por isso, foi necessário realizar uma exceção de contagem para essas duas tags, além do fato de não conter nenhum valor literal de data. Logo, o número total de registros excluindo essas duas tags vazias, seria de 306.538 registros, e esse valor será considerado na **figura 13**, que representa a composição desse vocabulário controlado.

Figura 13 - Representação estrutural quantitativa do GEMET



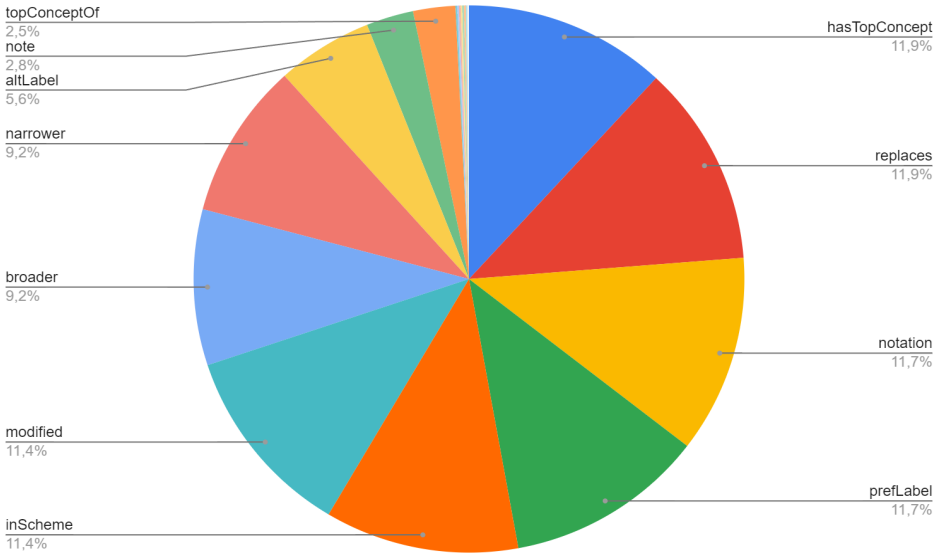
Fonte: Dados da pesquisa

PrefLabel é tag de maior destaque nesse vocabulário, representando mais da metade dele com 196.661 registros, isto se deve ao fato deste tesauro representar termos em 37 línguas diferentes, respeitando a grafia de cada uma delas, o que implica dizer que o número total de termos dividido por idioma abrangeria pelo menos 5.000 termos em cada língua, podendo variar para mais ou para menos, pois não há uma distribuição igualitária por idioma. Outra tag de participação relevante é definition 46.361 registros, porém as definições associadas a um determinado termo estão em todos os idiomas, por isso não há uma equivalência entre prefLabel e notation.

A tag member é a terceira mais expressiva com 16.178 registros, assim como as outras duas ela pertence ao skos-ns, seguida das tags narrower (5.689 registros), broader (5685 registros), inScheme (5.651 registros), altLabel (5.598 registros), closeMatch (5.522 registros) e source (4.547 registros). SKOS é o namespace predominante deste vocabulário, embora contenha uma variedade considerável de namespaces.

O VCGE é o menor vocabulário em número de totais de tags desta lista, com 1.010 registros. É o único monolíngue representado em língua portuguesa do Brasil, com 19 tipos distintos de tags utilizando 8 namespaces diferentes. A figura 14 representa a composição desse vocabulário controlado.

Figura 14 - Representação estrutural quantitativa do VCGE

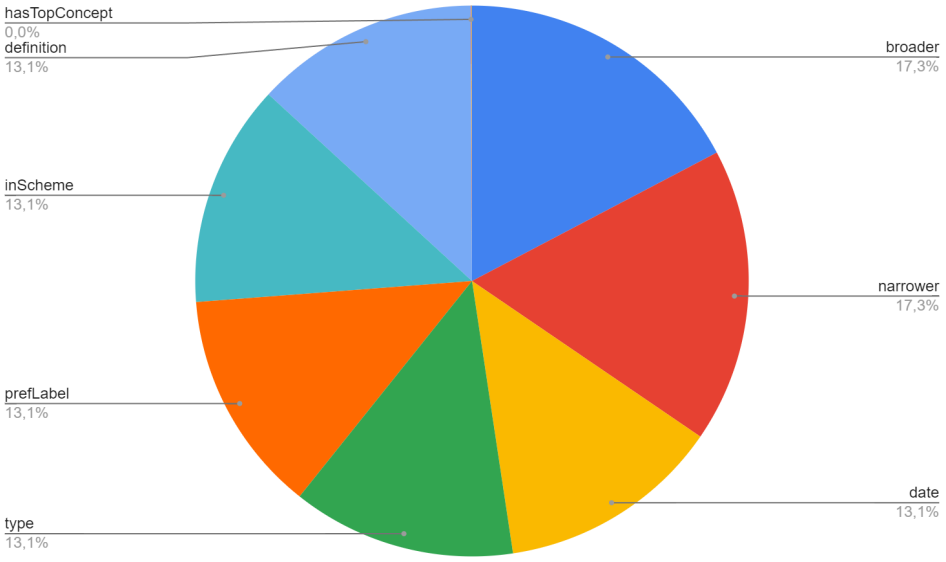


Fonte: Dados da pesquisa

HasTopConcept é a tag de maior participação nesse vocabulário com 120 registros, esta tag foi utilizada para representar a ordem numérica dos conceitos de 0 a 119, porque a tag 0 está apenas informando que se trata do vocabulário padrão do Governo Eletrônico. A tag replace do DC Namespace que representa recurso relacionado que é suplantado, deslocado ou substituído pelo recurso descrito também possui 120 registros. A tag notation, possui 118 registros, assim como prefLabel. As tags inScheme e modified também possui o mesmo número de registros que totalizam 115. O mesmo acontece com as broader e narrower para indicar termos gerais e específicos. A tag altLabel possui 57 registros, note 28 registros e topConceptOf 25 registros, o principal namespace utilizado nesse vocabulário é o SKOS.

O International Coastal Atlas Network Coastal Erosion Global Thesaurus (ICANCEGT) possui 17.155 registros dividido em 15 propriedades de descrição utilizando 19 namespaces diferentes, o que é no mínimo curioso, pois nem todos os namespaces são utilizados na descrição, pois 7 das 15 propriedades são do SKOS-NS, ou seja, restariam apenas 8 propriedades para 18 namespaces. O ICANCEGT é um tesauro focado em relacionamento de termos, pois as tags com maiores registros são broader e narrower com 2.964 registros cada uma.

Figura 15 - Representação estrutural quantitativa do ICANCEGT



Fonte: Dados da pesquisa

As tags prefLabel, date e type possuem 2.243 registros cada uma, enquanto as tags inScheme e definition apresentam 2.242 registros. A tag hasTopConcept possui apenas 7 registros, o restante das tags possuem apenas 1 registro, inclusive a tag altLabel. O ICANCEGT possui poucas tags de descrição que são distribuídas de forma muito equilibrada entre termos preferidos e como eles podem ser considerados gerais e específicos.

O Europeana Fashion Vocabulary possui 49.427 registros dividido entre 27 tipos de tags diferentes, com 38 XML namespaces, repetindo o mesmo caso do ICANCEGT. O Europeana Fashion Vocabulary também apresentou erros durante a contagem de tags, conforme a figura xx. Este erro indica que a URI analisada não era válida no momento em que o arquivo passou pelo *parse* da biblioteca do RDFLib, porém isso não interferiu na contagem.

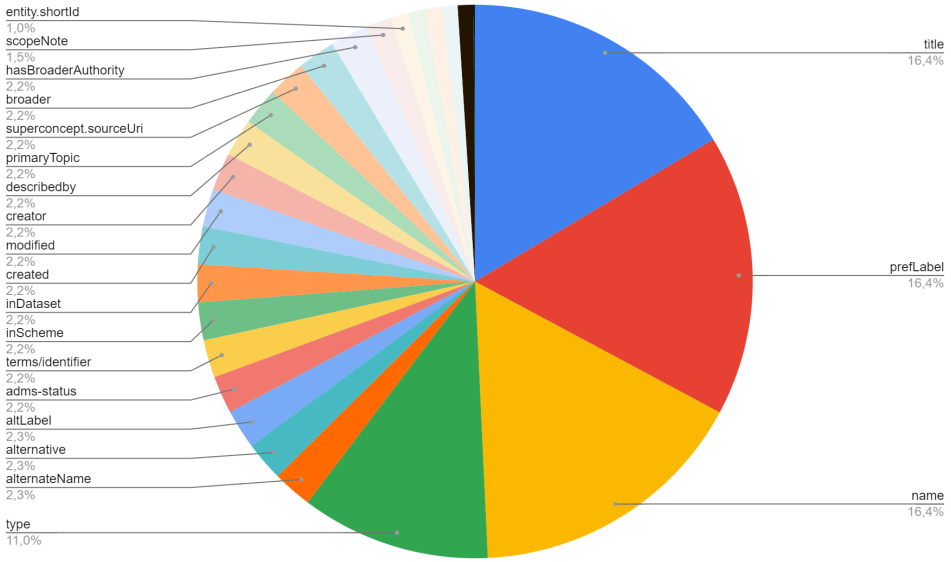
Figura 16 - Erros de *parse* do Europeana Fashion

Fonte: Dados da pesquisa

O Europeana Fashion Vocabulary é um tesauro baseado no AAT do Getty Research, por isso, o erro de URI associado ao AAT não gera nenhuma surpresa. No entanto,

ao clicar no link da possível falha, a página final de redirecionamento está em perfeito funcionamento, logo não se trata apenas de simples erro de contagem, mas pode indicar um problema estrutural, o que será analisado mais adiante, a figura 17 apresenta a composição desse vocabulário controlado por tipo de tag.

Figura 17 - Representação estrutural quantitativa do Europeana Fashion



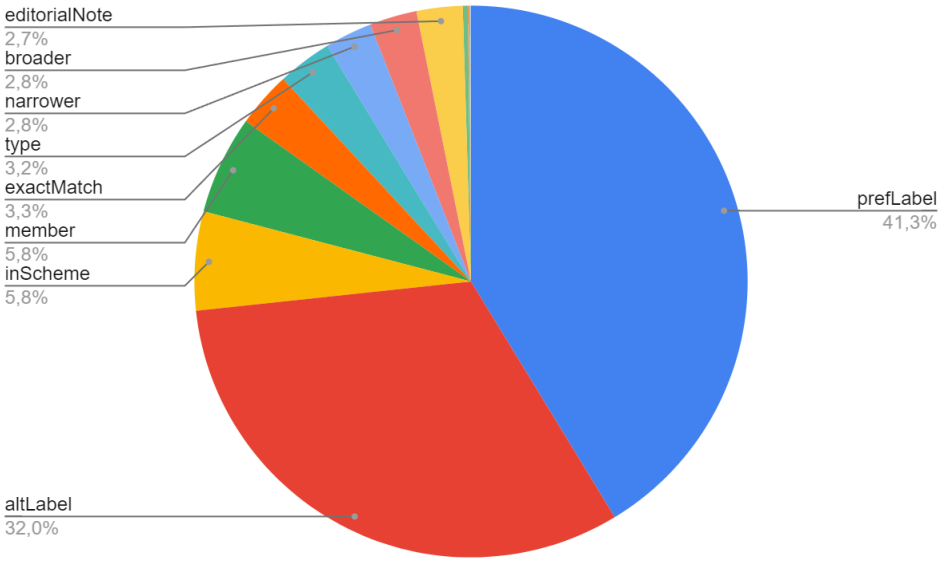
Fonte: Dados da pesquisa

As tags com maior expressividade nesse vocabulário são title, prefLabel e name, que possuem 8.117 registros cada, a soma dessas três tags representam quase metade desse vocabulário, enquanto a outra metade foi distribuída de forma razoavelmente proporcional entre os outros tipos de tags. A tag type possui 5.445, enquanto cada uma tags alternateName, alternative e altLabel possuem 1136 registros. As tags adms-status, terms/identifier, inScheme, inDataset, created, modified, creator, describedby, primaryTopic e superconcept.sourceUri também registraram 1.089 ocorrências cada uma delas. As tags broader e hasBroaderAuthority também possuem o mesmo número de registros 1.085, por fim a tag scopeNote obteve 730 e entity.shortId 499 registros.

O *Partage Plus* é um projeto que produziu diferentes tipos de representação, que varia de objetos, especialidades, lista de autoridades, técnicas e materiais de arte, porém, foi analisado apenas o *Partage Plus Vocabulary* (PPV) que possui 26.774 registros distribuídos

em 14 propriedades de descrição, 4 namespaces em 16 idiomas diferentes. A figura 18 representa a composição deste vocabulário controlado.

Figura 18 - Representação estrutural quantitativa do Partage Plus

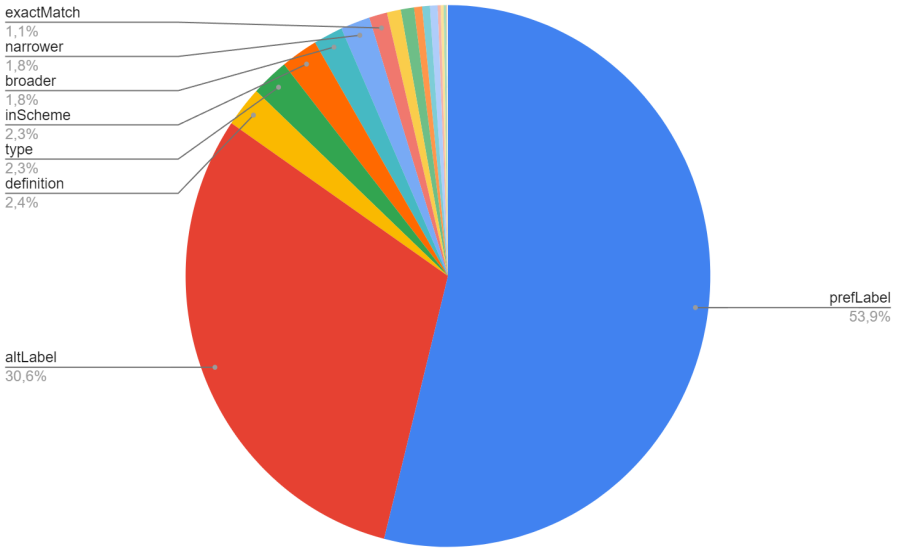


Fonte: Dados da pesquisa

A tag com maior representatividade é a prefLabel, que sozinha representa quase metade desse vocabulário controlado com 11.057 registros. Seguida da tag altLabel com 8.568 registros, somadas essas duas tags representam mais de 70% Partage Plus, que possui predominante tags do SKOS namespaces. As tags inScheme (1552 registros), member (1544 registros), exactMatch (871 registros), type (852 registros), narrower (741 registros), broader (740 registros) e editorialNote (721 registros).

O Resource Type Vocabulary (RTV) possui 4.431 registros distribuídos em 21 propriedades de descrição. A **figura 19** representa a composição desse vocabulário controlado, cuja maioria dos registros é representada pela tag prefLabel (2.389), seguida da tag altLabel (1.357).

Figura 19 - Representação estrutural quantitativa do Resource Type Vocabulary

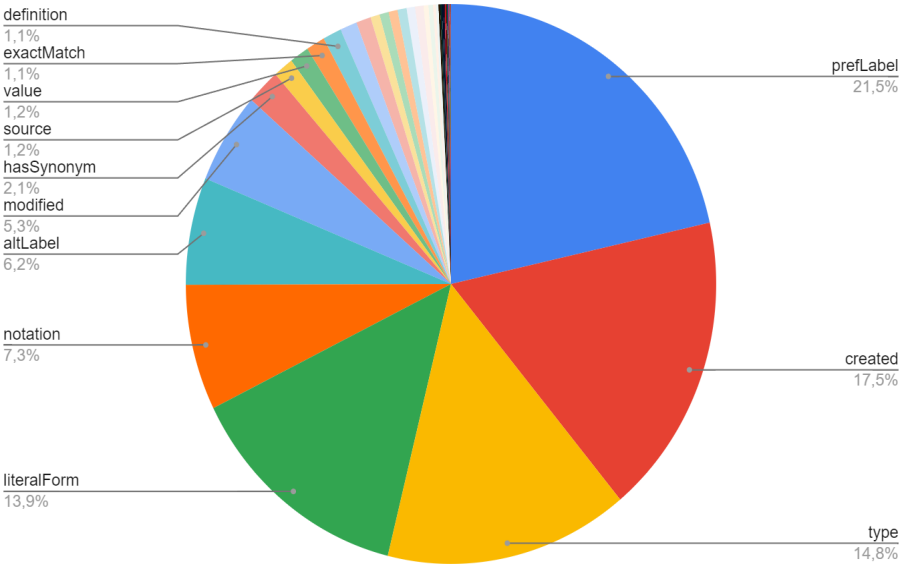


Fonte: Dados da pesquisa

A tag definition possui 107 registros, seguida das tags , type com 102 registros, inScheme (101 registros), broader (80 registros), narrower (80 registros) e exactMatch (49 registros). A maioria das tags utilizadas nesse vocabulário controlado pertencem ao SKOS namespaces, o que transmite uma certa simplicidade e objetividade a este vocabulário, característica também perceptível no Partage Plus Vocabulary.

O LandVoc é o último vocabulário controlado a ser analisado nesta etapa, ele possui 54.846 registros, divididos em 59 propriedades de descrição, utilizando 11 namespaces diferentes. A Figura 20 representa a composição do LandVoc a partir da contagem de tags que representam as propriedades do vocabulário controlado.

Figura 20 - Representação estrutural quantitativa do LandVoc



Fonte: Dados da pesquisa

Por se tratar de um sub-vocabulário do AGROVOC, o LandVoc possui muitas URIs que apontam para AGROVOC, inclusive conceitos e termos, além de possuir de forma simplificada as mesmas características de sua matriz terminológica. A tags com maiores expressividades numéricas são prefLabel (11.780), created (9.603), type (8.143), literalForm (7.597), notation (3.986), altLabel (3.400), modified (2.889), hasSynonym (1.129), source (674), value (645) exactMatch (625) e definition (624) registros.

Ao realizar análises estatísticas sobre a composição de cada vocabulário controlado é possível confirmar a afirmação de Gilreath (1993) ao explicar que ao mensurar numericamente um objeto, melhora-se a compreensão sobre ele. Não obstante, a análise quantitativa estrutural permite entender que os vocabulários controlados não são organizados de forma idêntica, apesar de possuírem muitos elementos em comum. Essa compreensão é necessária, pois não é possível analisar manualmente palavra por palavra ou relacionamentos entre esses termos em vocabulários controlados que possuem registros na casa dos milhares.

Apesar de todos os vocabulários possuírem documentação própria que remetem minimamente a história e campo temático do vocabulário controlado, essa não é uma realidade para todos os tipos de vocabulário controlados, o que pode dificultar ainda mais a compreensão destes. Outro fator de sucesso para esta análise rudimentar é fato estarem disponíveis em modelos de dados similares ou iguais, mesmo que não sejam tão populares

assim, o que ascende a necessidade de também disponibilizar em modelos ou formatos populares para outros tipos de análise utilizando ferramentas de análises mais consolidadas e com maior suporte de uso.

Considerando apenas os requisitos de disponibilidade, atualizações, suporte, documentação e modelos de dados foi possível selecionar uma pequena quantidade de vocabulários controlados. Tais requisitos não são demasiadamente complexos, apenas indicam que o processo de manutenção do mesmo está em funcionamento, ou que ele não se tornou órfão. Pois até a análise quantitativa exige o mínimo de informações básicas para contextualizar a avaliação geral, por exemplo, se algum vocabulário controlado possuísse um número expressivo de tags altLabel, isso poderia indicar que ele possui alguma origem com um dicionário de sinônimos, como foi o caso do VCGE, porém essa informação consta apenas na documentação do VCGE.

Nesta etapa de análise foi possível conhecer basicamente a composição de cada tesauro analisado, aproximando a compreensão da sua estrutura por meio das tags de descrição oriundas de diferentes tipos de namespaces. Porém, esta análise ainda não reflete o impacto que tais estruturas provocam ao estarem atreladas a um determinado modelo de dados, por isso, na etapa a seguir, serão identificados os principais problemas associados ao RDF e SKOS para a representação de vocabulários controlados.

4.3 Análise baseada em modelos de dados

A Análise baseada em modelos de dados fundamenta-se nas pesquisas desenvolvidas por Mader; Haslhofer; Isaac (2012), Suominen; Mader, (2014) e Suominen; Hyvönen (2012), que identificaram 26 problemas de qualidade analisados de forma automática e presentes em vocabulários SKOS registrados na literatura científica. Neste sentido, Suominen e Mader (2014) dividem os problemas em três categorias, que consistem em problemas de etiquetagem e documentação, problemas estruturais e problemas específicos de dados vinculados.

Os problemas de etiquetagem e documentação identificados por Suominen e Mader (2014) foram, as etiquetas vazias, tags de idioma omitido ou inválido, cobertura incompleta de idioma, conceitos não documentados, ausência de idiomas comuns, etiquetas ausentes, etiquetas sobrepostas, referências de notação ambígua, caracteres não imprimíveis em etiquetas, conceitos órfãos e clusters de conceito desconectados.

Os problemas estruturais identificados foram as relações hierárquicas cíclicas, relações associativas sem valor, conceitos únicos transitivamente relacionados, conceitos superiores omitidos, conceitos principais com conceitos mais amplos, redundância hierárquica, conceitos relacionados unidirecionalmente e conceitos relacionados reflexivamente. Sousa (2019) apresenta uma definição traduzida para cada tipo de problema que pode ser utilizada como recurso de consulta.

Os problemas específicos de dados vinculados identificados foram a ausência de links externos, recursos SKOS não definidos e violações do esquema de URI HTTP. Mader; Haslhofer; Isaac (2012), Suominen; Mader, (2014) e Suominen; Hyvönen (2012) dedicam-se a estudar esses tipos problemas ao vocabulários controlados publicados em SKOS e desde essas publicização destes trabalhos já se passaram mais de 10 anos, o que consequentemente inviabiliza a reprodução dos testes feitos, assim como o uso das ferramentas utilizadas na época. Por exemplo, Suominen; Mader (2014) utilizaram três tipos de ferramentas disponíveis para analisar 24 vocabulários SKOS, conforme a figura 21.

Figura 21 - Ferramentas de análise

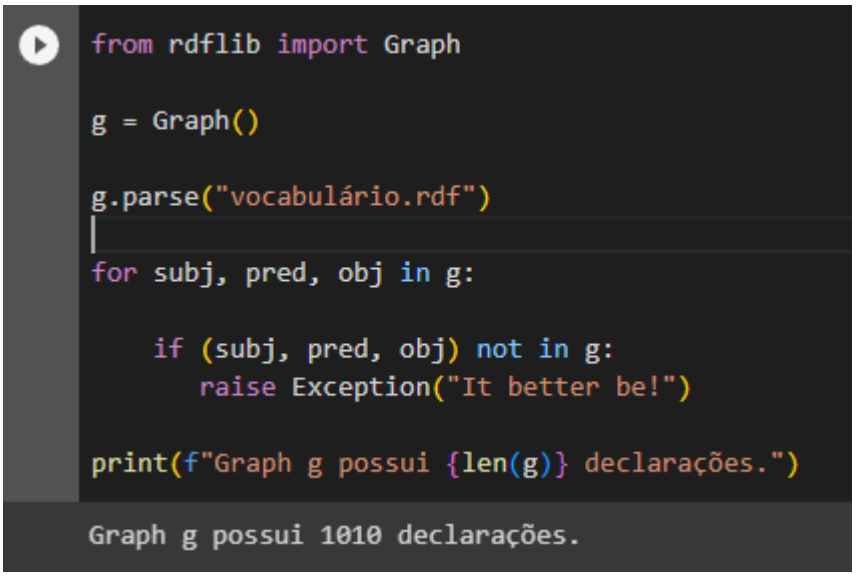
qSKOS	Skosify	verificador PoolParty
Versão 0.9.5 Licença GPLv3	0.6	(versão online atual) Proprietário
Linguagem de Implementação Java Inference Support Regras específicas do SKOS, RDFS Web Interface sem API sim (Java)	Regras específicas do MIT Python SKOS, RDFS sim sim (REST	(desconhecido) SKOS, RDFS, OWL sim não
URL da página inicial https://github.com/cmader/qSKOS/	online) http://code.google.com/p/skosify/	http://demo.semantic-web.at:8080/SkosServiços/cheque

Fonte: adaptado de Suominen e Mader (2014).

Até a presente data, essas ferramentas encontram-se indisponíveis para execução de análises remotas, porém o código fonte do qSKOS encontra-se disponível em um repositório de códigos de programação, entretanto, por ser desenvolvido em Linguagem Java, existe um certo grau de dificuldade de uso por não programadores. Embora essas ferramentas possuam comprovações de uso na literatura científica, existem também outras opções de análise que podem ser utilizadas com o uso da linguagem *Python*, como a biblioteca RDF Lib, que pode ser utilizada como analisadora e serializada de modelos de dados RDF/XML, N3, NTtriples, N-Quads, Turtle, TriX, JSON-LD, HexTuples, RDFa e Microdata. Além de possibilitar a implementação de consultas SPARQLs.

Para esta análise foi utilizada a biblioteca RDF Lib como recurso de análise dos vocabulários controlados publicados em RDF ou SKOS. A interface principal do RDFLib é um arquivo Graph. Os grafos RDFLib são contêineres não classificados para explorar as triplas RDF. Na página de documentação do RDF Lib são apresentados alguns exemplos e os principais recursos que podem ser utilizados, conforme apresentado na figura 22.

Figura 22 - Exemplo de código em Python com RDF Lib

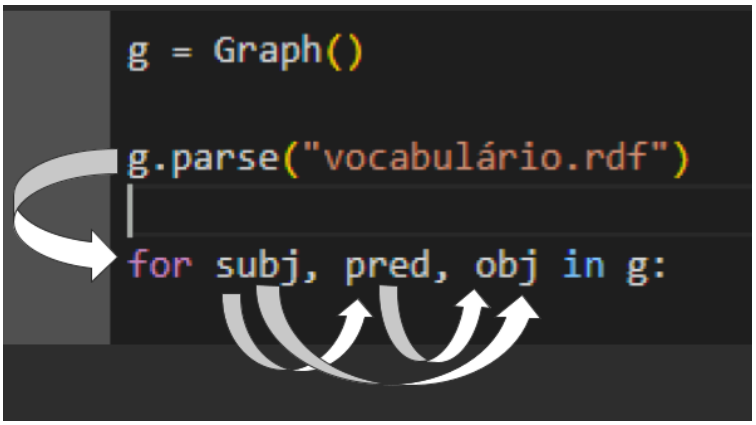


Fonte: Adaptado de RDF Lib (2023)

O código apresentado descreve de forma primária como essa biblioteca realiza suas operações, que consistem em criar um gráfico "Graph ()", ler um arquivo em RDF por meio da função parse, analisar as tripla subject, predicate e object, suportando exceções e apresentar como resultado por meio da função print o número de declarações existentes no vocabulário controlado analisado. A partir deste exemplo e da documentação disponível sobre essa biblioteca foi possível aprimorar esse código para encontrar possíveis problemas existentes em vocabulário controlados.

A identificação dos erros apresentados nesta pesquisa possuem como fundamentado os trabalhos de Mader; Haslhofer; Isaac (2012), Suominen; Mader, (2014) e Suominen; Hyvönen (2012), nas recomendações da W3C sobre RDF e na documentação da biblioteca RDF Lib em *Python*. Usando a própria regra de *parse* do tripla RDF é possível verificar se algum dos componentes da tripla está faltando, duplicada ou não está linkada uma à outra, conforme ilustrado na figura 23.

Figura 23 - Lógica de análise



Fonte: Adaptado de RDF Lib (2023)

Utilizando essa lógica de verificação foi possível identificar 15 tipos de problemas que podem ser encontrados em vocabulários controlados disponíveis em SKOS ou RDF, que consistem em verificar Erros de Sintaxe, Triplas Duplicadas, Nós em Branco, Namespaces Indefinidos, Sujeitos ausentes, Predicados ausentes, Objetos ausentes, Types Inconsistentes, Sujeitos duplicados, Objetos duplicados, Classes duplicadas, Tags de idioma Inconsistentes, Propriedades Duplicadas, Classes perdidas e URIs Inválidas.

A primeira verificação se baseia na identificação de erros de sintaxe serializando o vocabulário controlado para Notation3 (N3), uma sintaxe RDF legível, alternativa, compacta e que permite maior expressividade. Segundo Berners-Lee e Connolly (2011) é uma linguagem de asserção e lógica, um superconjunto de RDF, adicionando fórmulas (literais que são gráficos em si), variáveis, implicação lógica e predicados funcionais, além de fornecer uma alternativa de sintaxe textual para RDF/XML. Para ilustrar as outras etapas de identificação de erros foi elaborada a tabela 06, alocando os clusters de erros de acordo com as categorias sugeridas por Suominen; Mader (2014).

A segunda verificação consiste em identificar triplas duplicadas, ou seja se cada tripla contendo subject, predicate e object se repete no vocabulário, não se trata de uma contagem individual de cada elemento da tripla, apenas da tripla completa. A terceira verificação com Nós em Branco, utilizando um recurso da própria biblioteca rdflib, chamado Blank Node (Nó em Branco) também chamado de BNode, que representa um recurso para o qual uma URI ou literal não é declarada, um Bnode também pode ser chamado de recurso anônimo. A verificação de Namespaces Indefinidos consiste em identificar problemas relacionados aos namespaces dos modelos de dados declarados em cada vocabulário, embora

a rdflib possua o recurso NamespaceManager, não foi possível explorá-lo para esta finalidade devido aos constantes erros para executar o código, por isso foi necessário o usar uma verificação utilizada em arquivos xml, considerando os arquivos declarados em RDF, para os que estavam em N3 ou NT foi necessário serializar e converter para RDF.

Para identificar os Sujeitos, Predicados e Objetos ausentes, foi necessário verificar individualmente cada elemento da tripla procurando por elementos ausentes. Os Types Inconsistentes foram identificados utilizando o recurso RDF.type e RDFS.Resource da rdflib. A verificação de Sujeitos e Objetos duplicados, também refere-se a uma busca individual de cada elemento da tripla, adicionando uma contagem de repetições. A identificação de tags de idiomas consideradas inconsistentes foi realizada utilizando o recurso da rdflib chamada Literais que são valores de atributos em RDF, utilizando o langString com uma tag de idioma.

Também foram verificadas as classes duplicadas utilizando o recurso RDF.type e RDFS.Class da rdflib. As Propriedades Duplicadas foram verificadas percorrendo cada tripla considerando apenas o subject e o predicate, criando uma lista e verificando repetições. As Classes perdidas foram verificadas com o uso do recurso RDF.type e RDFS.Class da biblioteca rdflib, com intuito de identificar as classes ausentes. Por fim, as URIs Inválidas foram verificadas utilizando o recurso de broken links para verificar URIs quebradas. O quadro 08 representa uma categorização dos tipos de erros que serão verificados em cada vocabulário controlado.

Quadro 08 – Categorização dos erros identificados

CATEGORIA	ERROS RDFLib	Definição
Problemas de etiquetagem	Tags de idioma inconsistentes	Identifica tags de idioma não padronizadas
	Erros de Sintaxe	Identifica erros de declaração sintática
	Triplas Duplicadas	Identifica triplas repetidas
	Nós em Branco	Identifica blank nodes
	Sujeitos ausentes	Identifica sujeitos ausentes ou desconexos com predicates ou objects
	Predicados ausentes	Identifica predicados ausentes ou desconexos com subjects ou objects
	Objetos ausentes	Identifica objects ausentes ou desconexos com subjects ou predicates
	Types Inconsistentes	Identifica Types não padronizadas
	Sujeitos duplicados	Identifica somente subjects repetidos

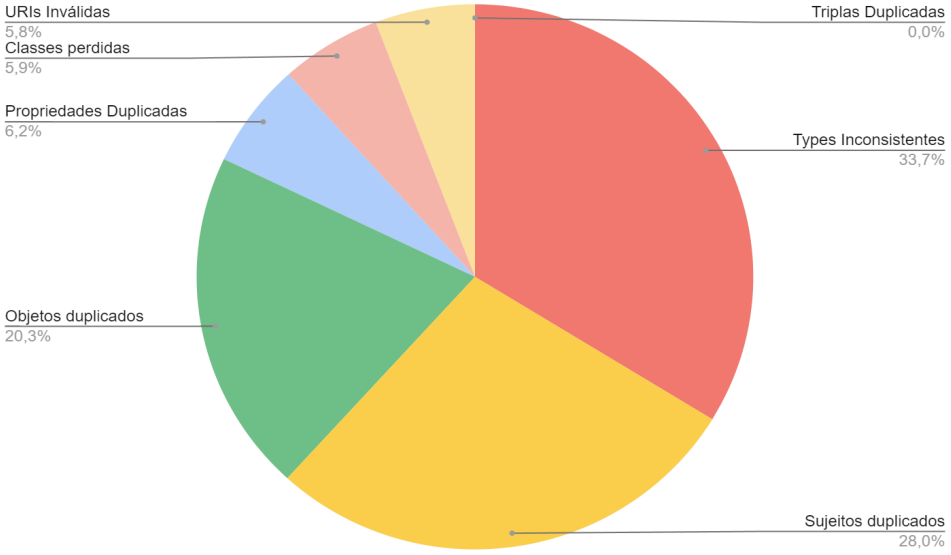
Questões Estruturais	Objetos duplicados	Identifica somente objects repetidos
	Classes duplicadas	Identifica somente classes repetidas
	Propriedades Duplicadas	Identifica propriedades RDF repetidas
	Classes ausentes	Identifica classes RDF ausentes
Problemas de dados vinculados	Namespaces indefinidos	Identifica Namespaces não declarados
	URIs inválidas	Identifica URIs quebradas

Fonte: elaborado pelo autor

Os erros apresentados predominantemente referem-se às questões estruturais associadas aos modelos de dados, o que se justifica pela proposta desse tipo de análise, que focaliza as características estruturais, diferente da abordagem de Suominen; Mader (2014), que incluem muito mais problemas de documentação e problemas estruturais distintos enfatizando problemas conceituais e de relacionamentos. Os problemas de etiquetagem e documentação relacionados às tags de idiomas são muito mais comuns em vocabulários controlados multilíngues. Enquanto os problemas de dados vinculados são os tipos de erros mais comuns encontrados em vocabulários SKOS e RDF.

Seguindo a ordem de análise da etapa anterior, começamos com o AGROVOC, que dos 15 tipos de problemas verificados, foram constatados somente 9 tipos de erros com 357 Triplas Duplicadas, 66 Nós em Branco, 33 Namespaces Indefinidos, 5784759 Types Inconsistentes, 4802763 Sujeitos duplicados, 3485815 Objetos duplicados, 0 Tags de idioma Inconsistentes, 1064965 Propriedades Duplicadas, 1006532 Classes perdidas, 995428 URIs Inválidas. Embora os números pareçam expressivos, é necessário considerar que o arquivo analisado possui milhões de linhas contendo declarações em RDF.

Figura 24 - Análise BMD AGROVOC



Fonte: elaborado pelo autor

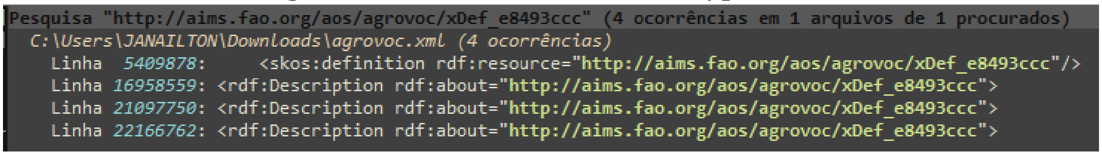
O tipo de erro com o maior número de ocorrências são de Types Inconsistentes, porém é importante considerar que as ocorrências de Types Inconsistentes não necessariamente representam uma falha, mas uma possível erro de posicionamento de tag, pois foi constatado que em alguns casos o recurso não foi definido como rdf:type por exemplo, mas como resource e acrescentou-se uma extensão skos:definition, que é um recurso de documentação, enquanto o rdf:resource é um recurso de classe, o que pode ser interpretado como um erro, conforme apresentado na figura 25. Embora tal declaração não prejudique a interpretação das tags, o correto seria declarar como rdf:type.

Figura 25 - Erros Types BMD AGROVOC

Fonte: elaborado pelo autor

Utilizando a interface de navegação do AGROVOC e comparando com a análise do arquivo RDF é possível constatar as indicações de erros durante a navegação online e permite compreender de forma pontual, cada erro localizado. Por exemplo, esse mesmo recurso declarado como skos:definition possui outras 4 ocorrências no arquivo analisado como rdf:description, conforme a figura 27. Trata-se de um problema estrutural que pode ser corrigido com a substituição de uma tag mais apropriada, que skos:definition.

Figura 26 - Linhas com Erros de Types BMD

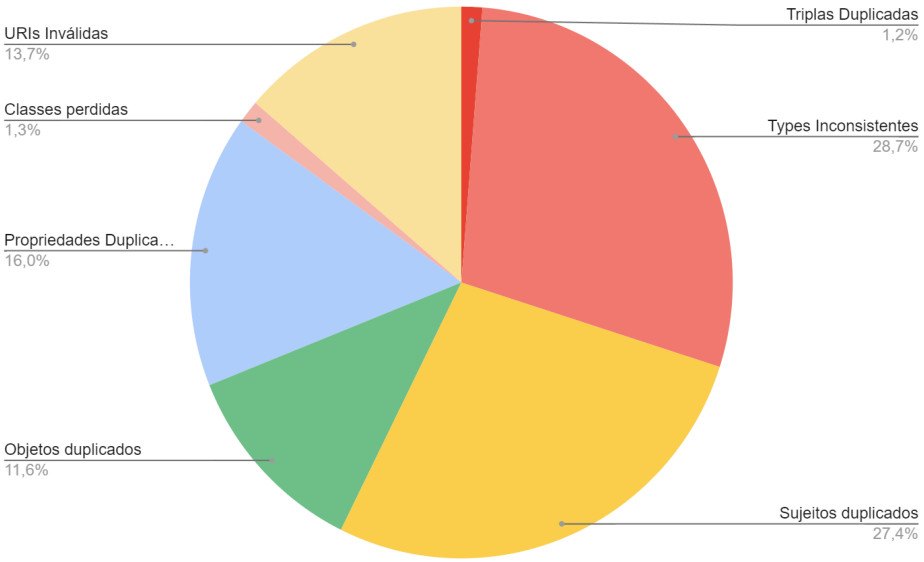


Fonte: elaborado pelo autor

Outros dois problemas também relacionados às questões estruturais foram a identificação de Propriedades Duplicadas devido ao uso das *tags* skos:prefLabel e skosxl:prefLabel, que podem ser correspondentes entre si e Classes perdidas, que estão associada às *tags* skos-xl#Label. Foram identificados também dois tipos de problemas de dados vinculados, que consistem em Namespaces perdidos e URIs Inválidas. A ausência de namespaces é indicada pela sugestão de possíveis namespaces que podem ser utilizados ou que não foram declarados no arquivo RDF, por exemplo o CSVW *Namespace Vocabulary Terms*, indicado em uma das ocorrências. Por fim, as URIs inválidas são problemas comuns e esperados em ambientes web.

O CAB Thesaurus apresentou sete tipos de erros, que consistem em 8078 Triplas Duplicadas, 5 Namespaces Indefinidos, 186417 Types Inconsistentes, 177692 Sujeitos duplicados, 75412 Objetos duplicados, 103995 Propriedades Duplicadas, 8726 Classes ausentes, URIs Inválidas 89051.

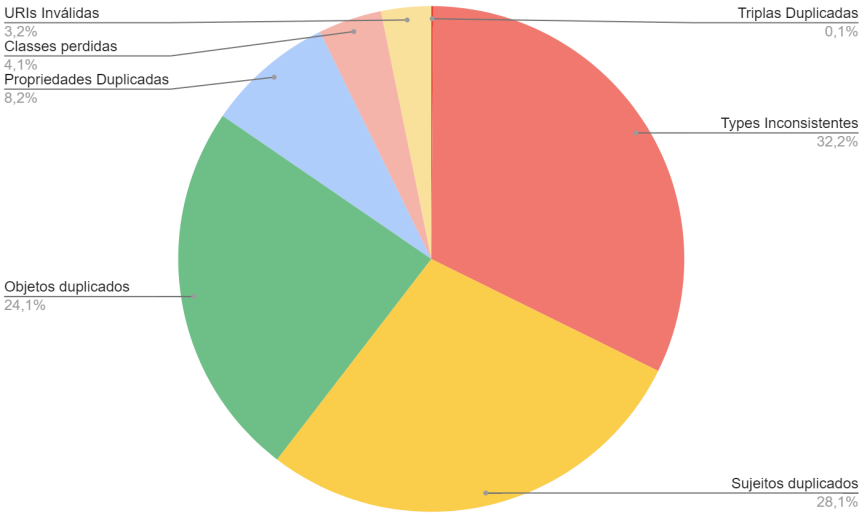
Figura 27 - Erros BMD CAB Thesaurus



Fonte: elaborado pelo autor

Seguindo o processo de análise, o EuroVoc apresentou 8 tipos de problemas, com 13045 Triplas Duplicadas, 6 Namespaces Indefinidos, 3662023 Types Inconsistentes, 3200297 Sujeitos duplicados, 2746934 Objetos duplicados, 929647 Propriedades Duplicadas, 461731 Classes perdidas e 364033 URIs Inválidas. Destaque-se que a maioria dos problemas identificados referem-se às questões estruturais, conforme apresentado na figura 28.

Figura 28 - Análise BMD EUROVOC

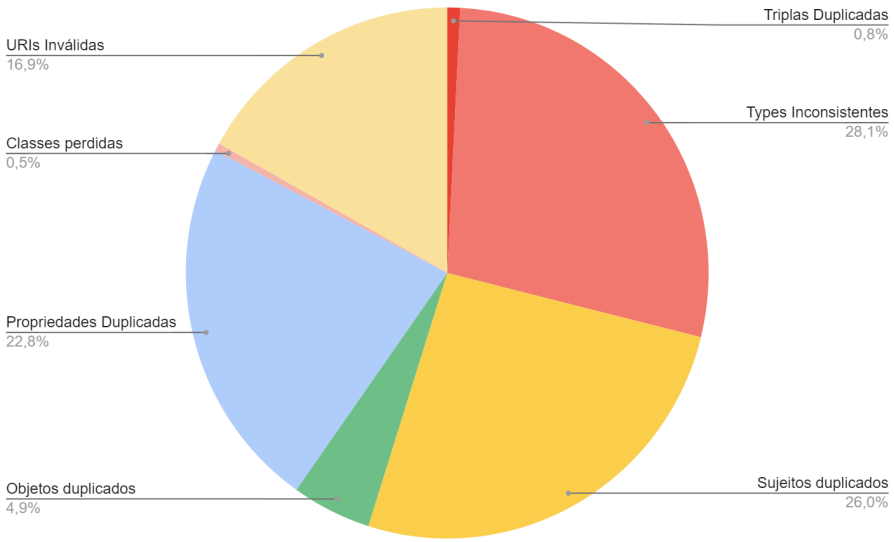


Fonte: elaborado pelo autor

O EuroVoc possui uma ampla distribuição dos 8 tipos de erros, o que pode ser considerado tomando como base o tamanho desse vocabulário controlado, que possui milhares de termos e relacionamentos entre si. Esses erros começam a se tornar comuns à medida que a análise avança com outros vocabulários controlados, com exemplos semelhantes de erros, conforme já demonstrado no AGROVOC.

O GEMET também apresentou 7 tipos de erros, que consistem em 9622 Triplas Duplicadas, 6 Namespaces Indefinidos, 343717 Types Inconsistentes, 317896 Sujeitos duplicados, 60020 Objetos duplicados, 279311 Propriedades Duplicadas, 5746 Classes perdidas e 206712 URIs Inválidas, conforme apresentado na figura 29.

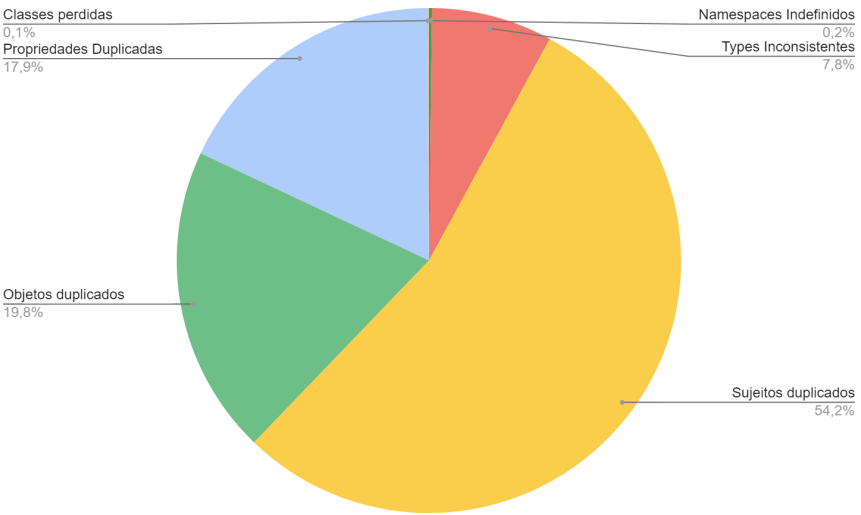
Figura 29 - Análise BMD GEMET



Fonte: elaborado pelo autor

O VCGE apresentou 7 tipos de erros, com 3 Namespaces Indefinidos, 128 Types Inconsistentes, 894 Sujeitos duplicados, 326 Objetos duplicados, 295 Propriedades Duplicadas e 2 Classes perdidas, conforme apresentado na figura 30.

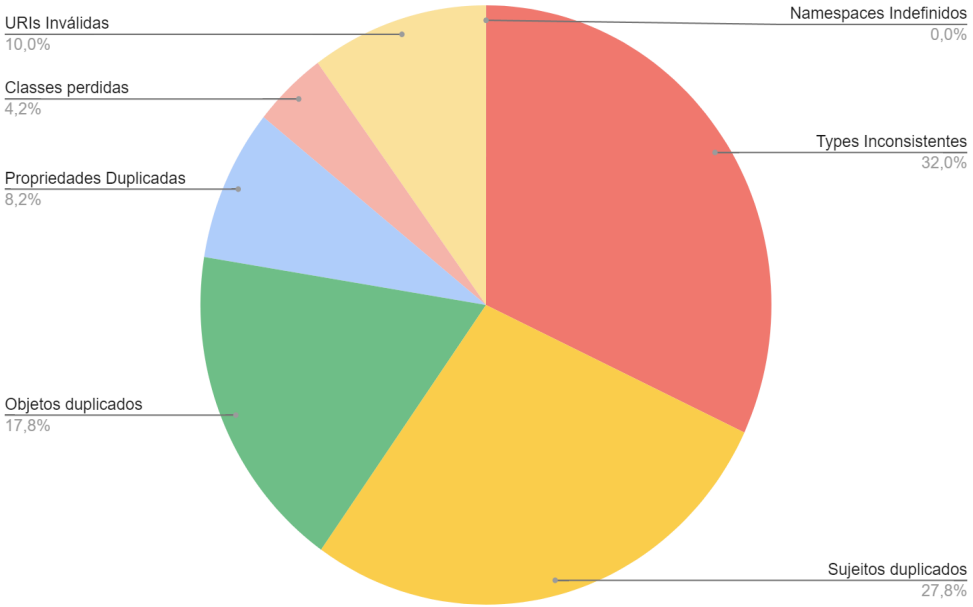
Figura 30 - Análise BMD VCGE



Fonte: elaborado pelo autor

O ICANCEGT também registrou 7 erros distintos, que consistem em 4 Namespaces Indefinidos, 17155 Types Inconsistentes, 14912 Sujeitos duplicados, 9527 Objetos duplicados, 4408 Propriedades Duplicadas, 2245 Classes perdidas e 5381 URIs Inválidas. Conforme apresentado na figura 31.

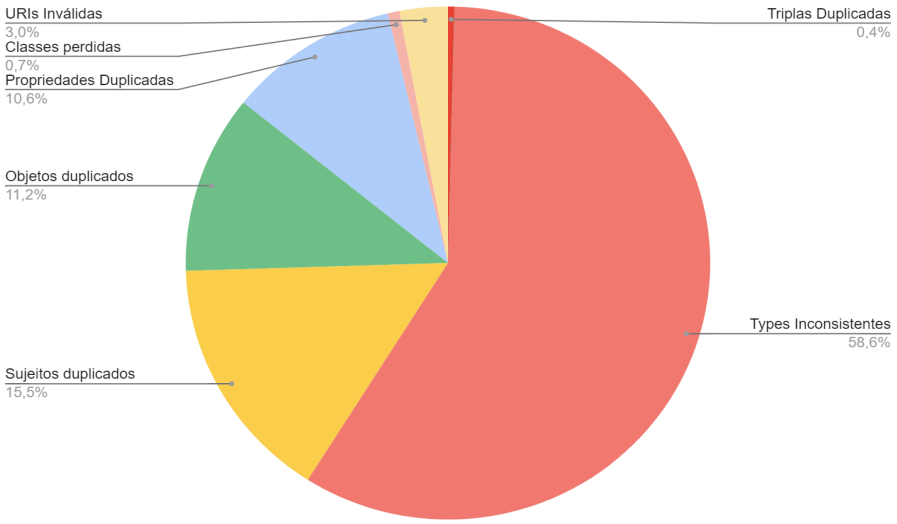
Figura 31 - Análise BMD ICANCEGT



Fonte: elaborado pelo autor

O Europeana Fashion Vocabulary registrou 9 tipos de erros que se dividem em 1 Erros de Sintaxe, 1185 Triplas Duplicadas, 11 Namespaces Indefinidos, 178106 Types Inconsistentes, 47249 Sujeitos duplicados, 34180 Objetos duplicados, 32080 Propriedades Duplicadas, 2183 Classes perdidas e 9050 URIs Inválidas, conforme apresentado na figura 32.

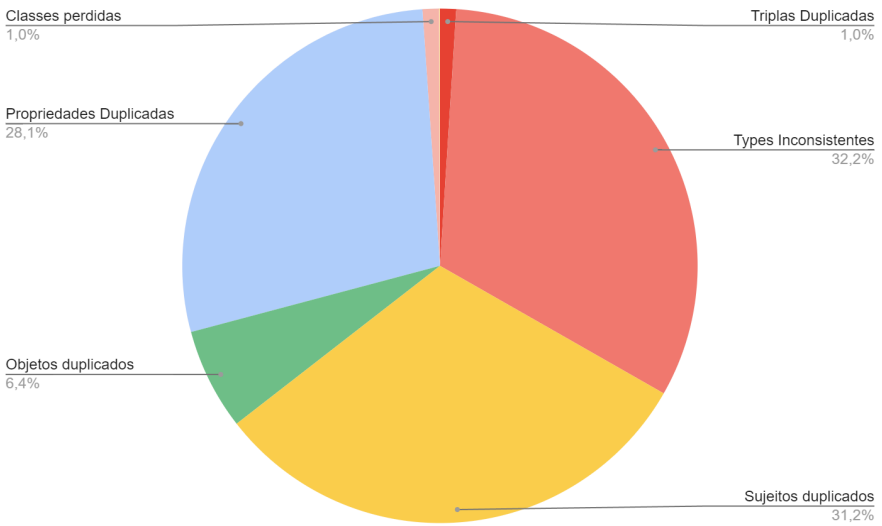
Figura 32 - Análise BMD Europeana Fashion Vocabulary



Fonte: elaborado pelo autor

O Partage Plus Vocabulary também apresentou 8 tipos de erros que consistem em 835 Triplas Duplicadas, 2 Namespaces Indefinidos, 26774 Types Inconsistentes, 25922 Sujeitos duplicados, 5293 Objetos duplicados, 23292 Propriedades Duplicadas, 855 Classes perdidas, 49 URIs Inválidas, conforme apresentado na figura 33.

Figura 33 - Análise BMD Partage Plus Vocabulary

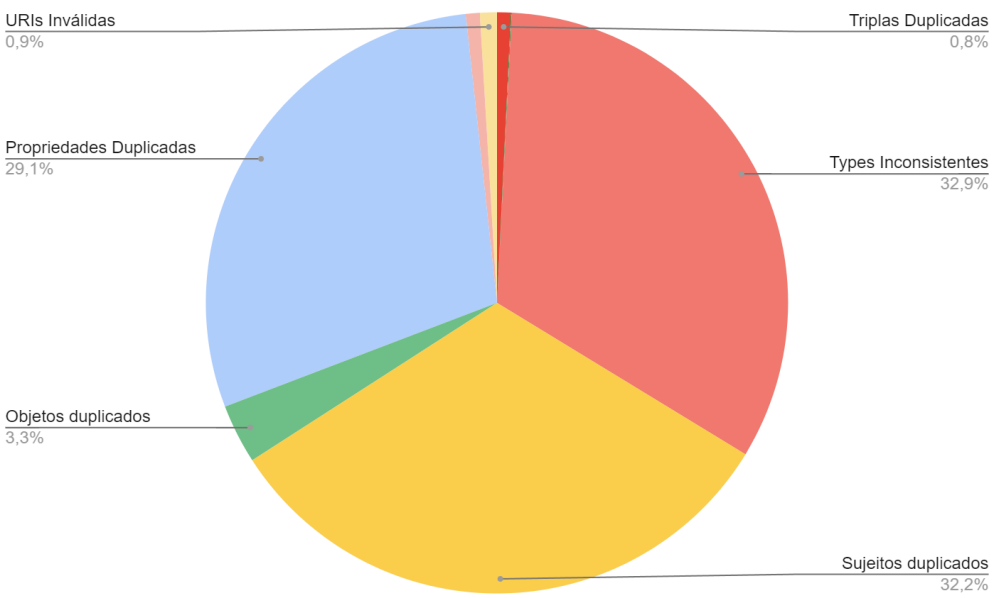


Fonte: elaborado pelo autor

O Resource Type Vocabulary (RTV) registrou 8 ocorrências de erros que consistem em 103 Triplas Duplicadas, 5 Namespaces Indefinidos, 4431 Types Inconsistentes, 4329 Sujeitos duplicados, 444 Objetos duplicados, 3916 Propriedades Duplicadas, 104

Classes perdidas e 127 URIs Inválidas. A figura 34 apresenta uma ilustração percentual de cada problema distribuído em um gráfico.

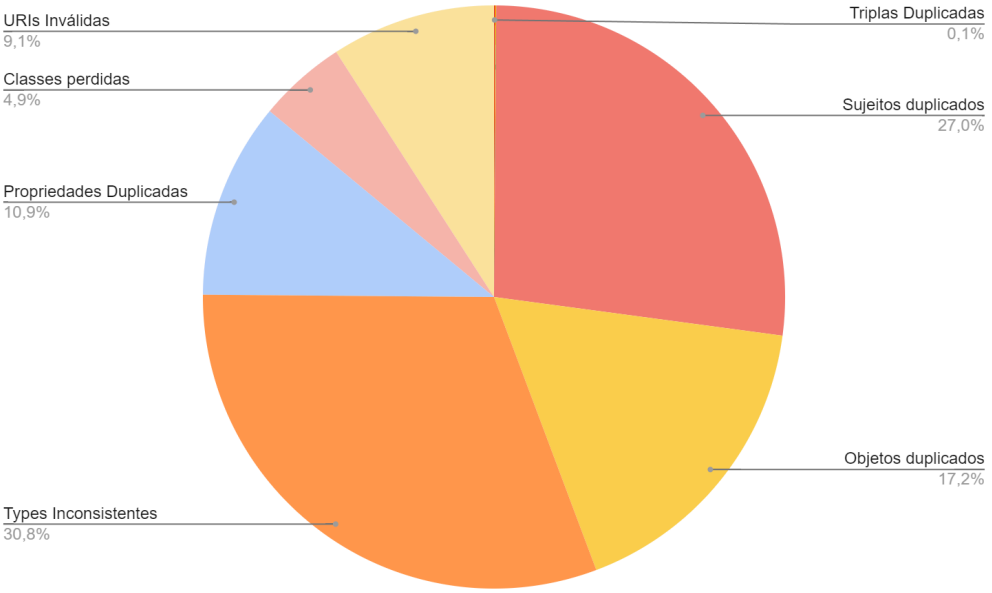
Figura 34 -Análise BMD RTV



Fonte: elaborado pelo autor

Por fim, o LandVoc também apresentou 9 tipos de erros distribuídos entre 109 Triplas Duplicadas, 85 Nós em Branco, 9 Namespaces Indefinidos, 44960 Sujeitos duplicados, 28593 Objetos duplicados, 51364 Types Inconsistentes, 18106 Propriedades Duplicadas, 8131 Classes perdidas, 15194 URIs Inválidas.

Figura 35 - Análise BMD LandVoc



Fonte: elaborado pelo autor

Os erros aqui descritos identificados nos vocabulários controlados analisados, são indícios que podem ser utilizados para melhorar sua qualidade. Por isso é necessário investigar com mais profundidade cada tipo de erro. A partir dessas indicações também é possível projetar frameworks que sinalizem erros mapeados que podem ser comuns em ambientes vinculados a partir desses modelos de dados. Para completar essa avaliação, a próxima análise se baseia no processamento de termos que foram extraídos de cada vocabulário controlado, como será demonstrado a seguir.

4.4 Análise métrica de termos

A análise métrica de termos reúne técnicas de NLP com uso de uma biblioteca de programação intitulada a NLTK, amplamente difundida com Python em ambientes acadêmicos. A realização desta etapa foi dividida em seis partes, que consistiram em verificar o número de tokens, o tamanho do vocabulário, a diversidade lexical, a polaridade de sentimento, subjetividade e o grau de entropia do vocabulário controlado. A NLTK possui uma limitação de 23 idiomas, o que significa que vocabulários controlados um número maior que esse farão parte dos termos ignorados durante o processamento.

Para calcular o número de tokens em cada vocabulário controlado, os objetos foram divididos e contados como palavras individuais chamadas de tokens. O tamanho do vocabulário foi calculado considerando apenas as palavras únicas e não repetidas. A diversidade lexical é a razão entre o número de palavras únicas e o número total de palavras no texto, quanto maior é a diversidade lexical, mais rico e variado é o texto.

Para identificar a polaridade de sentimento e a subjetividade do vocabulário controlado utilizou-se a biblioteca TextBlob do Python para processar dados textuais. Pois ela oferece uma API para tarefas comuns de NLP, como marcação de partes da fala, extração de frases nominais, análise de sentimento, classificação, entre outros, para analisar os dados de texto. O sentimento é uma medida da polaridade e da subjetividade do texto, que variam de -1 a 1. A polaridade indica se o texto é positivo, negativo ou neutro, com base em conjunto de palavras indexadas nesta biblioteca que indica a subjetividade do texto com base em fatos ou opiniões indexadas nessa biblioteca.

Por fim, para calcular a entropia do vocabulário foi utilizada a biblioteca math, utilizando uma medida da incerteza ou da informação contida no texto, com base na fórmula da entropia de Shannon, que soma o produto da probabilidade de cada caractere pelo logaritmo dessa probabilidade. O cálculo de entropia do vocabulário foi realizado de acordo com Bird, Klein e Loper (2009) quando explanam sobre entropia e ganho de informação para medir a desorganização de um conjunto de valores de entrada, para isso, usam a seguinte fórmula:

NLTK $H = -\Sigma l \text{ in labels } P(l) \times \log 2 P(l).$

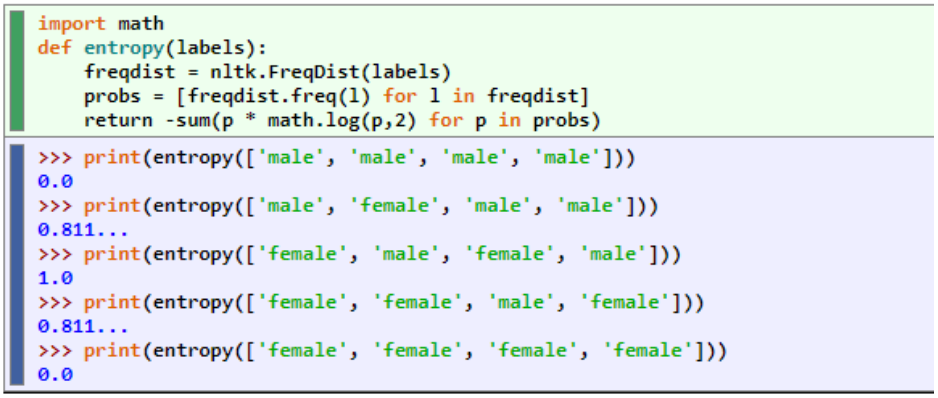
A entropia é maior quando o texto tem mais diversidade e menor quando o texto tem mais repetição, logo, ela depende do tipo e do tamanho do texto. Usando a fórmula original da entropia de Shannon, a entropia é uma medida da aleatoriedade e imprevisibilidade do texto e pode ser usada para analisar a complexidade ou o conteúdo da informação do texto.

Shannon $H = -\sum p(i) * \log(p(i), 2)$

Pinto (2017) afirma que quanto menor a entropia, maior o peso da informação, menor será a taxa de perda da fonte e mais completa a informação se apresentará. Embora o uso dessa fórmula pareça complexa e de difícil entendimento a priori, é necessário esclarecer

que há exemplificações sobre o uso do cálculo de entropia no livro elaborado por Bird, Klein e Loper (2009) que explana sobre os recursos disponíveis com NLTK, conforme apresentado na figura 36.

Figura 36: Exemplo de cálculo de entropia da informação



```
import math
def entropy(labels):
    freqdist = nltk.FreqDist(labels)
    probs = [freqdist.freq(l) for l in freqdist]
    return -sum(p * math.log(p,2) for p in probs)
```

```
>>> print(entropy(['male', 'male', 'male', 'male']))
0.0
>>> print(entropy(['male', 'female', 'male', 'male']))
0.811...
>>> print(entropy(['female', 'male', 'female', 'male']))
1.0
>>> print(entropy(['female', 'female', 'male', 'female']))
0.811...
>>> print(entropy(['female', 'female', 'female', 'female']))
0.0
```

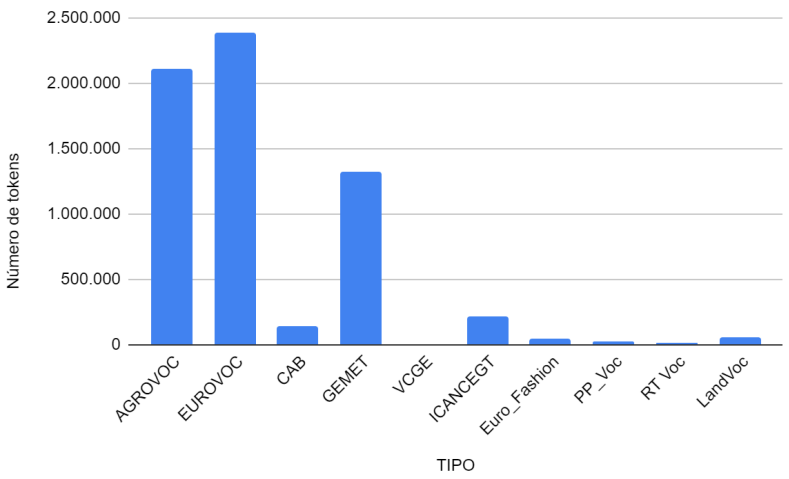
Example 4.3 (code_entropy.py): **Figure 4.3**: Calculating the Entropy of a List of Labels

Fonte: Bird, Klein e Loper (2009)

Assim como a entropia da informação, as outras cinco etapas também exploram os recursos apresentados por Bird, Klein e Loper (2009), por isso é possível aplicar seu uso aos vocabulários controlados disponíveis em RDF ou SKOS. Ou seja, a fórmula base de cálculo dos recursos textuais já estão definidas pelas bibliotecas em python, cabendo ao programador designar as funções desejadas, considerando as possibilidades de uso de cada biblioteca em python.

Apresentar índices de desempenho para vocabulários reserva uma certa dificuldade, por isso cada parte da análise métrica é apresentada de modo comparativo com outros vocabulários controlados, no intuito de identificar médias de desempenho para entre si. Para ilustrar o número de tokens a figura 37 apresenta os vocabulários os mais expressivos numericamente por meio da contagem de tokens.

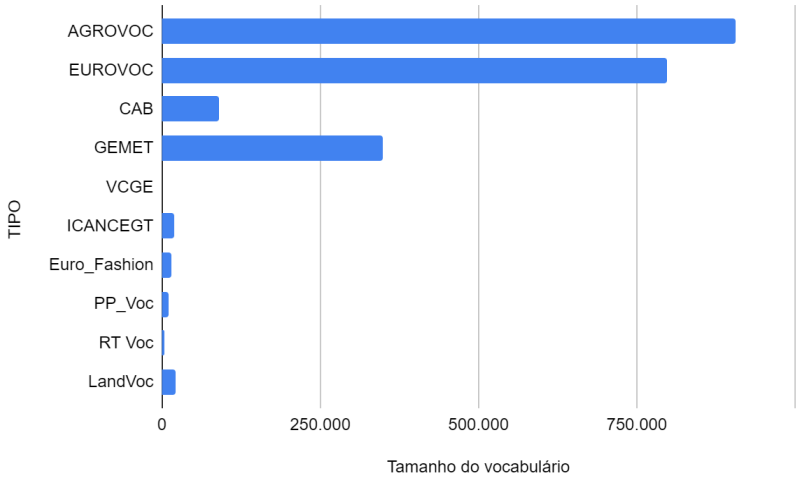
Figura 37: Número de tokens



Fonte: Dados da pesquisa (2023)

É possível constatar que o vocabulário controlado com o maior número de tokens, isto é considerando todas as palavras, inclusive repetições é o Eurovoc, com mais de 2 milhões, embora o AGROVOC e o GEMET também possuam um número expressivo também. Enquanto, o VCGE, como já foi declarado anteriormente, é o menor vocabulário identificado. Outra medida apresentada trata-se do tamanho de um vocabulário controlado, considerando apenas a unicidade de cada palavra, conforme apresentado na figura 38.

Figura 38: Tamanho dos vocabulário

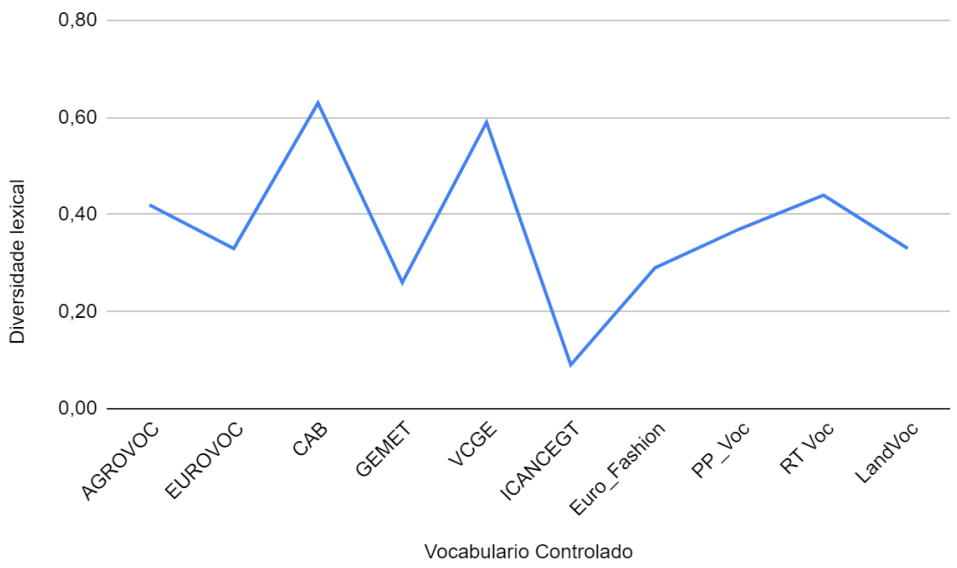


Fonte: dados da pesquisa (2023)

Neste caso o AGROVOC é o maior vocabulário controlado considerando o número de palavras únicas com quase 1 milhão de termos, seguido do Eurovoc, com mais 750 mil termos e do GEMET com mais de 250 mil termos. Assim, como o cálculo do número de tokens, o tamanho do vocabulário apresenta uma dimensão real do tamanho de um vocabulário controlado, que muitas vezes não é divulgado pela instituição mantenedora e quando divulgado não há orientações de como verificar esses dados.

Para estipular um índice de diversidade lexical de cada vocabulário controlado foi adotada uma escala centesimal de 0 a 1, onde os vocabulários controlados que alcançaram pontuações mais próximas de 0 possuiriam menor diversidade lexical e os que estivessem mais próximos de 1 teriam maior diversidade lexical, conforme ilustrado na figura 39.

Figura 39: Diversidade lexical

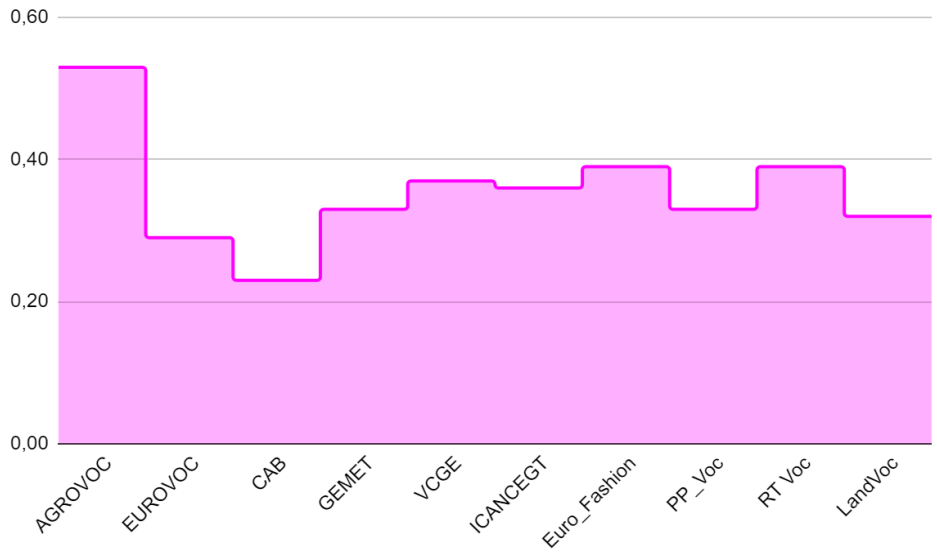


Fonte: dados da pesquisa (2023)

Dentre os vocabulários controlados analisados, o CAB tesauros e o VCGE são os que possuem maior diversidade lexical, acima de 0,50, vocabulários controlados maiores como o AGROVOC e o EUROVOC não tenham alcançado a melhor pontuação mesmo possuindo um repertório maior de termos, assim como o GEMET, um vocabulários controlado numericamente expressivo, mas que registrou pouca diversidade léxica, embora o ICANCEGT tenha apresentado o pior desempenho.

Para analisar o nível de subjetividade de cada vocabulário controlado também foi adotada uma escala centesimal de 0 a 1, onde os vocabulários controlados com pontuações mais próximas de 0 são considerados menos subjetivos e os mais próximos de 1 são considerados mais subjetivos. Conforme apresentado na figura 40.

Figura 40: Nível de subjetividade

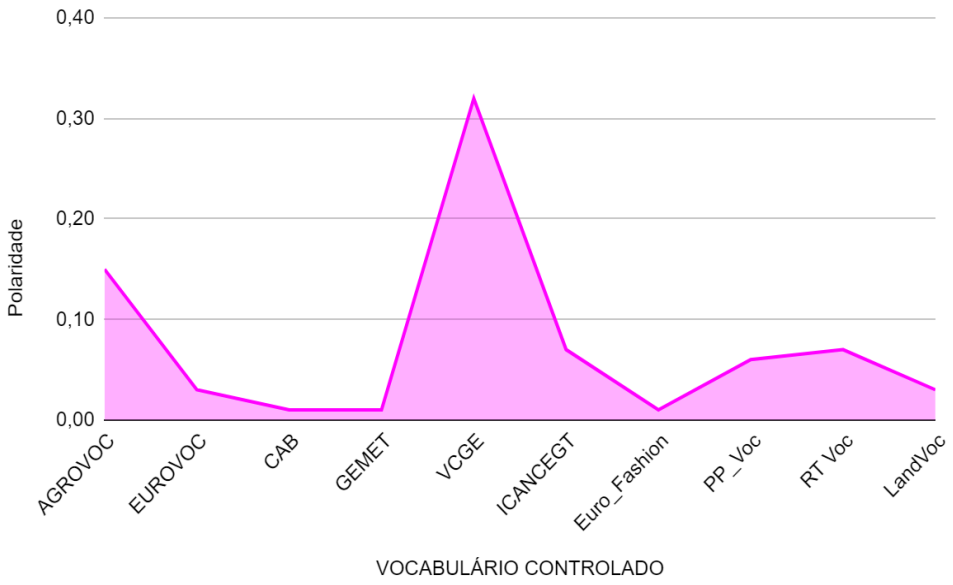


Fonte: Dados da pesquisa (2023)

Dos vocabulários controlados analisados, apenas o AGROVOC apresentou um índice elevado de subjetividade, embora ou EUROVOC seja o segundo maior vocabulário controlado analisado, ele apresentou quase a metade da pontuação do AGROVOC, o que demonstra um desempenho bem melhor dado o número de tokens e o tamanho do EUROVOC, como foi visto anteriormente. No entanto, o CAB Tesauro foi o que apresentou a menor pontuação, portanto pode ser considerado como aquele que possui o menor grau de subjetividade. Os outros vocabulários apresentaram uma média razoável, menor que 0,40.

Para analisar o grau de polaridade de cada vocabulário controlado também foi adotada uma escala centesimal de -1 a 1, pois este tipo de análise considera opções positivas e negativas de análise textual, ou até mesmo neutro. A figura 41 apresenta uma ilustração dos resultados obtidos na verificação de cada vocabulário controlado.

Figura 41: Grau de polaridade

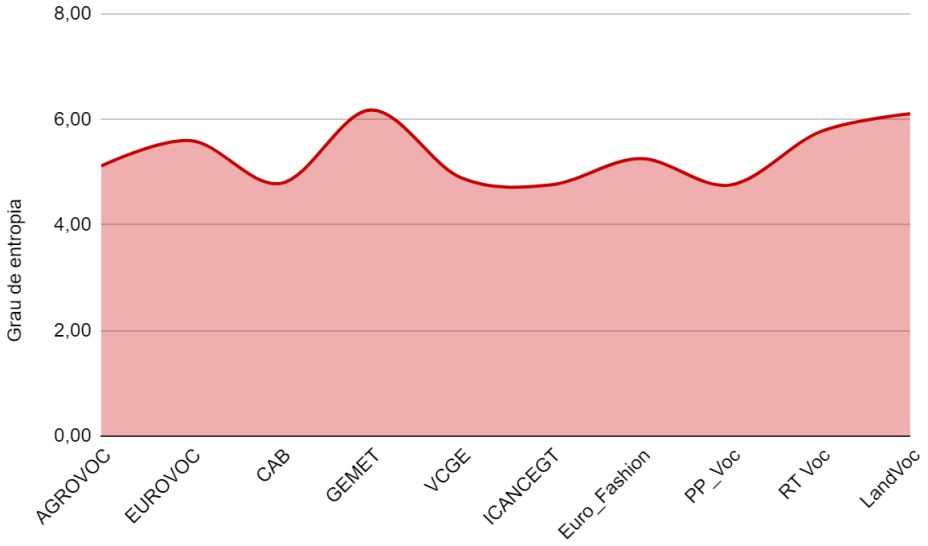


Fonte: Dados da pesquisa (2023)

Dentre os vocabulários controlados analisados, nenhum registrou índice negativo, não obstante, o VCGE foi o vocabulário controlado que apresentou maior grau de positividade, seguido do AGROVOC, enquanto o CAB tesauros, o GEMET e o Europeana Fashion apresentaram resultados muito próximo de zero, o que pode indicar um certo grau de neutralidade desses vocabulários.

Para verificar o grau de entropia foi considerada uma escala de 0 a 10, considerando o vocabulário controlado com menor entropia mais objetivo será o vocabulário, quanto maior a entropia, maior o grau de desordem do vocabulário controlado. Para ilustrar essa análise a figura 42 apresenta o grau de entropia resultante de cada vocabulário controlado.

Figura 42: Grau de entropia



Fonte: Dados da pesquisa (2023)

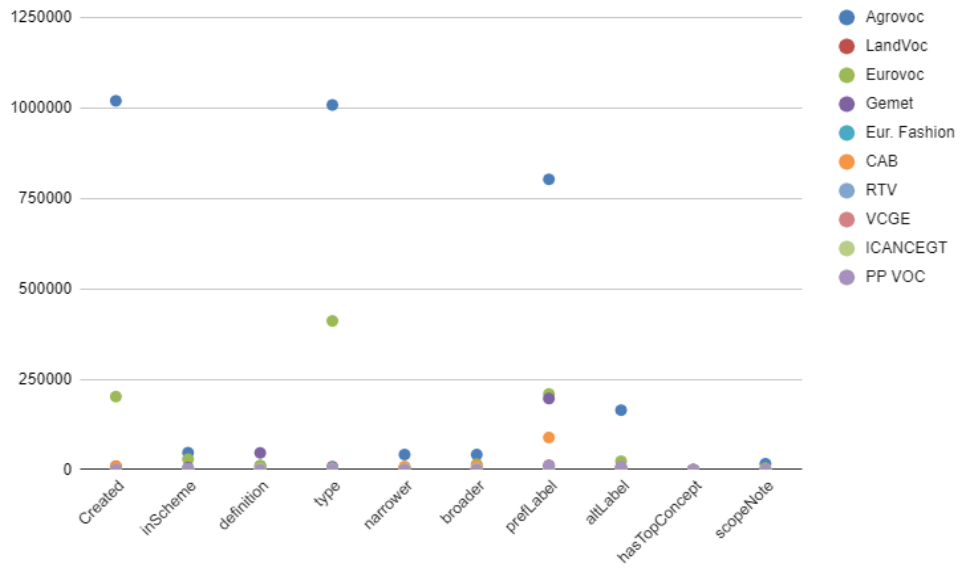
Percebe-se que todos os vocabulários controlados possuem um grau de entropia relativamente estável entre 4 e 6, embora apresentem grande diferenças entre números de termos, extensão e níveis diferentes de complexidade de relacionamentos entre si. O que pode indicar uma certa simetria encontrada nos vocabulários controlados, oferecendo inclusive, parâmetros para pesquisas futuras que envolvam alinhamento de vocabulários controlados com base no grau de entropia.

Esta última análise permite concluir a etapa final de avaliação dos vocabulários controlados selecionados para a pesquisa, assim como fornece insumos para a síntese dos resultados e conclusão da pesquisa. Por meio da análise M.T é possível vislumbrar o uso da onometria com técnicas já consolidadas de NLP para a avaliação de termos dos vocabulários controlados. Isto fornece subsídios para uma avaliação formal de vocabulários controlados, principalmente aqueles que possuem uma grande quantidade de termos, cujo esforço de análise seria manualmente inviável.

5 RESULTADOS E DISCUSSÕES

Para consolidar os resultados de cada tipo de análise demonstra-se a seguir uma comparação geral entre os vocabulários controlados. Começando com a Análise Estrutural Quantitativo, onde se percebe que alguns dos vocabulários controlados analisados possuem um comportamento peculiar, diferente dos outros, conforme apresentado na figura 43. Que apresenta as dez tags mais utilizadas entre os vocabulários controlados analisados, ou seja, que estão presentes em todos ou em 70% do total.

Figura 43 - Tags mais utilizadas



Fonte: Dados da pesquisa (2023)

As tags mais expressivas do AGROVOC e Eurovoc foram created e type, que basicamente é uma tag de registro DC e de definição de um tipo RDF. Tal expressividade pode ser explicada devido ao tamanho do vocabulário controlado, que possui muitos termos termos registrados e modificados ao longo dos anos, o que eleva a criação desses tipos de dados. Estes dados também demonstram que esses vocabulários estão crescendo ao longo do tempo. Ao contrário dos outros vocabulários controlados que aparentemente possuem um comportamento similar em suas estruturas. Outro ponto de destaque é o fato de algumas tags que estão presentes em todos os vocabulários controlados analisados conforme apresentado no quadro 09, excluindo as tags created e type.

Quadro 09 – Propriedades mais comuns

VOC/TAG	narrower	broader	prefLabel	altLabel
AGROVOC	41.781	41.781	802.662	164.499
LandVoc	303	303	11.780	3.400
EUROVOC	4.759	15.130	209.234	23.308
GEMET	5.689	5.685	196.661	5.598
Eur. Fashion	493	1.085	8.117	1.136
CAB	8.534	9.430	88.912	6.708
RTV	80	80	2.389	1.357
VCGE	93	93	118	57
ICANCEGT	2.964	2.964	2.243	1
PP VOC	741	740	11.057	8.568

Fonte: Dados da pesquisa (2023)

As tags apresentadas no quadro acima não são recorrentes por acaso, pois elas equivalem aos termos preferidos, alternativos e relacionamentos amplos e estreitos. Não obstante, refere-se a uma estrutura equivalente a representação de um tesauro em SKOS, que também é o namespace mais utilizado. Outro ponto de destaque é que em 70% dos tesauros o número de relacionamentos amplos e estreitos são idênticos ou muito próximos. E o número de termos alternativos é sempre menor que o número de termos preferidos, caso contrário seriam um dicionário de sinônimos, inclusive acende um alerta para identificar as características dos distintos tipos de vocabulários controlados.

Por meio da Análise Baseada em Modelos de Dados foi possível constatar os 8 tipos de problemas mais comuns que podemos encontrar em vocabulários controlados, conforme apresentado no quadro 10. Além disso, esses problemas são caracterizados como questões estruturais e de dados vinculados, o que implica dizer que são problemas relacionados ao modelo de dados adotado e não necessariamente ao conteúdo do vocabulário controlado.

Quadro 10 – Erros mais comuns identificados

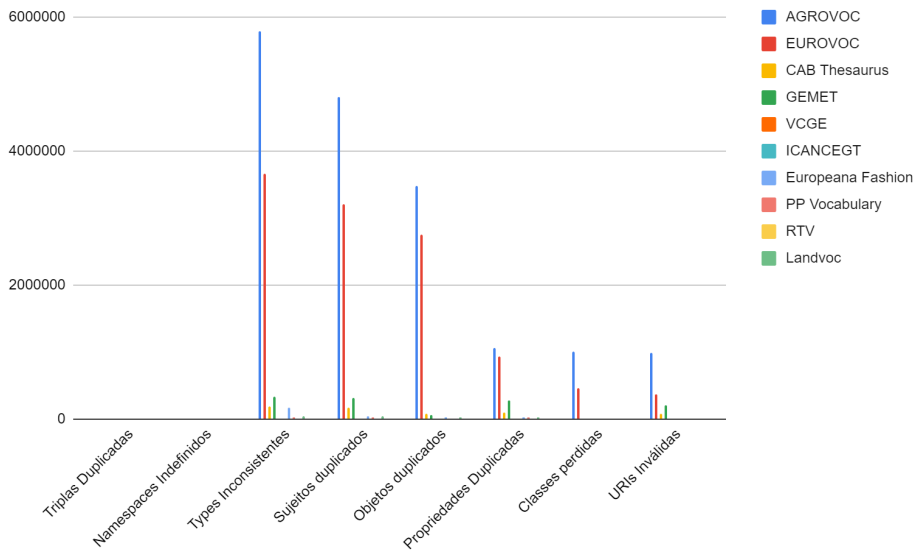
Vocabulário	Triplas Duplicadas	Namespaces Indefinidos	Types Inconsistentes	Sujeitos duplicados	Objetos duplicados	Propriedades Duplicadas	Classes perdidas	URIs Inválidas
AGROVOC	357	12	5784759	4802763	3485815	1064965	1006532	995428
EUROVOC	13045	6	3662023	3200297	2746934	929647	461731	364033
CAB Thesaurus	8078	5	186417	177692	75412	103995	8726	89051
GEMET	9622	6	343717	317896	60020	279311	5746	206712

VCGE	0	11	128	894	326	295	2	0
ICANCEGT	0	4	17155	14912	9527	4408	2245	5381
Europeana Fashion	1185	11	178106	47249	34180	32080	2183	9050
PP Vocabulary	835	2	26774	25922	5293	23292	855	49
RTV	103	30	4431	4329	444	3916	104	127
Landvoc	109	9	51364	44960	28593	18106	8131	15194

Fonte: Dados da pesquisa (2023)

Considerando a tabela acima e a recorrência de determinados problemas é possível ilustrar o problema que mais se destaca em todos os vocabulários controlados conforme apresentado na figura 44. Nota-se que o erro de propriedades ausentes é o que mais se destaca em 6 vocabulários. Seguido de types e namespaces inconsistentes.

Figura 44 - Análise BMD erros comuns totais



Fonte: Dados da pesquisa (2023)

O AGROVOC e o EUROVOC são os vocabulários controlados que mais apresentam tipos de problemas, muitos desses problemas se devem ao tamanho desses vocabulários controlados, que possuem conjuntos de termos na ordem dos milhares, além do tempo de existência deles, o que pode indicar muitas alterações de infraestrutura tecnológica,

que pode ocorrer por meio de migrações entre sistemas de informação ao longo do tempo, além de possíveis conversões de formatos e padrões distintos que entraram em vigor por determinado tempo.

Por meio da análise Métrica de Termo foi possível construir um quadro dos resultados obtidos nesta etapa da análise, diferente dos quadros anteriores que apresentam os erros e tipos mais comuns, desta vez foi possível apresentar todas as métricas extraídas dos termos de cada vocabulário controlado.

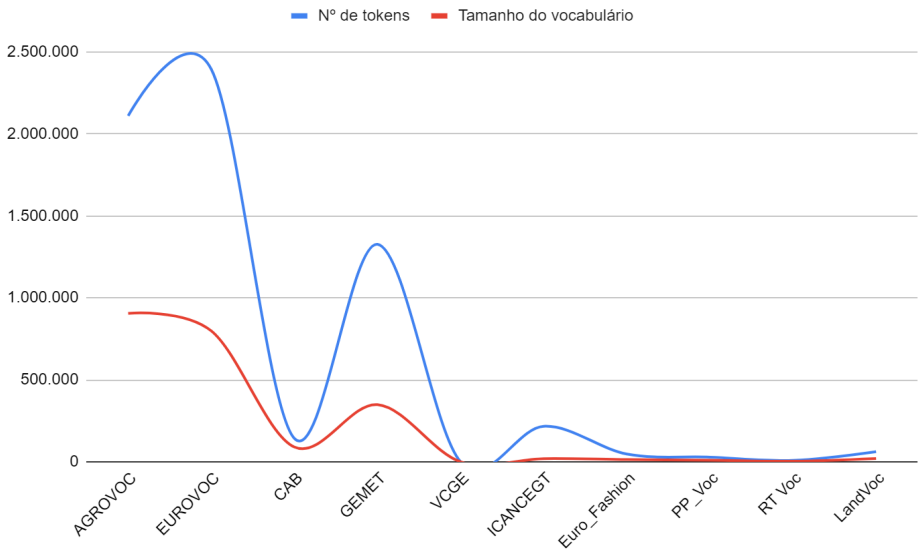
Quadro 11 – Métricas de termos totais

VOCABULÁRIO	Nº de tokens	Tamanho do vocabulário	Diversidade lexical	Subjetividade	Polaridade	Grau de entropia
AGROVOC	2.109.765	905.261	0,42	0,53	0,15	5,12
EUROVOC	2.393.804	798.711	0,33	0,29	0,03	5,60
CAB	143.294	90.751	0,63	0,23	0,01	4,79
GEMET	1.326.800	348.705	0,26	0,33	0,01	6,18
VCGE	646	386	0,59	0,37	0,32	4,90
ICANCEGT	217.141	19.957	0,09	0,36	0,07	4,76
Euro_Fashion	47.870	14.230	0,29	0,39	0,01	5,26
PP_Voc	28.995	10.736	0,37	0,33	0,06	4,76
RT Voc	9.602	4.234	0,44	0,39	0,07	5,77
LandVoc	62.461	20.992	0,33	0,32	0,03	6,11

Fonte: Dados da pesquisa (2023)

A análise Métrica de Termos foi sintetizada de duas formas, devida a escala de grandeza numérica, pois o número de tokens e o tamanho dos vocabulários controlados possuem variação de 0 a mais de 2 milhões de tokens, conforme apresentado na figura 45.

Figura 45 - Tokens e tamanho do vocabulário

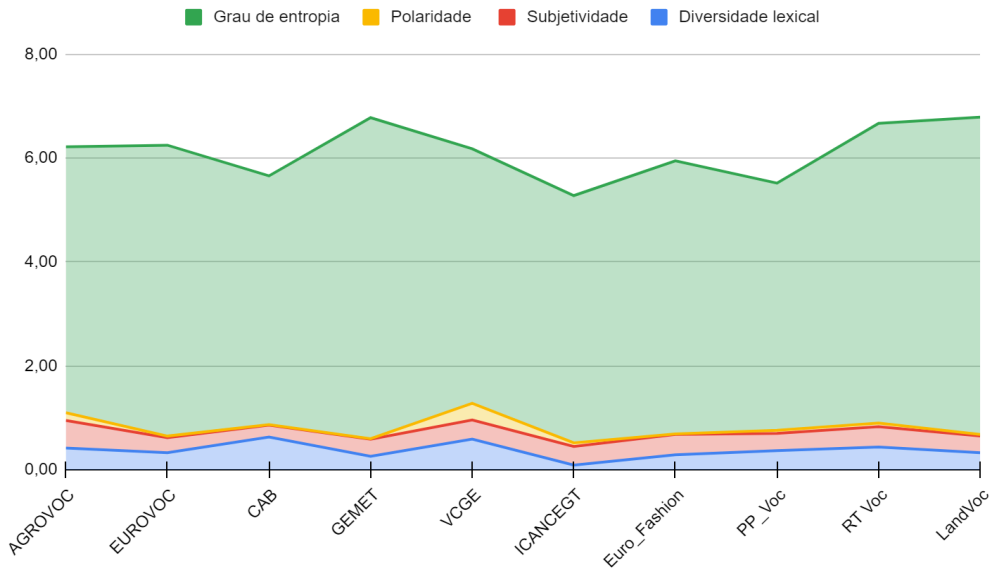


Fonte: Dados da pesquisa (2023)

O Eurovoc e o AGROVOC são os maiores vocabulários analisados, como já era esperado, porém o Eurovoc apresentou um maior número de tokens, possivelmente isso se deve ao fato de ser um tesauro multidisciplinar, com um nível de abrangência maior, enquanto o AGROVOC é um vocabulário especializado em alimentação e agricultura. As duas métricas apresentaram uma certa cadência, o que pode indicar uma afinidade entre essas duas medidas, assim como uma consideração de análise em conjunto dessas métricas.

A segunda síntese dos resultados da análise Métrica de Termos possui uma escala de grandeza menor, que varia de 0 a 10, embora a diversidade lexical, a subjetividade e polaridade possuam variações de 0 a 1. No entanto, para efeito de comparação entre essas métricas optou-se por uma ilustração agrupada, conforme a apresentada na figura 46.

Figura 46 - Graus entre vocabulários controlados



Fonte: Dados da pesquisa (2023)

A diversidade lexical, a subjetividade e polaridade possuem comportamento semelhante em todos os vocabulários controlados analisados, embora o VCGE, o AGROVOC e o CAB Tesauros tenham demonstrado uma elevação maior dessas características. No entanto, o mérito dessas características fica apenas com a diversidade lexical, ou seja, esses três vocabulários controlados foram os que apresentaram maior riqueza lexical, comparado aos outros.

Por outro lado, a entropia dos termos dos vocabulário controlados não parece se alinhar a nenhuma das medidas anteriores, porém quando comparada entre si, o GEMET é o vocabulário controlado com maior grau de entropia, alinhado ao pensamento de Bird, Klein e Loper (2009) e Pinto (2017) esse é vocabulário controlado pode ser considerado o menos conciso ou mais diversificado, seguindo do LandVoc, o Partage Plus e o ICANCEGT apresentaram um desempenho igual, portanto podem ser considerados os mais concisos e com menor risco de aleatoriedade.

De modo sintético, a partir dos resultados é possível fazer observações distintas sobre os três tipos de análises, a primeira diz respeito da composição e estrutura de cada vocabulários controlado, a análise estrutural quantitativa que se mostrou útil para mapear a composição estrutural de um vocabulário controlado, incluindo a possibilidade de classificar os tipos de vocabulários controlados considerando o seu arranjo descritivo e lógica de

estruturação. Por exemplo, seria possível determinar se um vocabulário controlado é uma taxonomia ou tesauro com base em tipos de tags mais expressivas.

A análise baseada em modelos de dados se mostrou complexa, pois mapear possíveis erros também pode incorrer na incompreensão dos tipos de erros, ou de um mesmo tipo de erro ser o motivo de outros erros, questionando inclusive se os resultados apresentados de fatos refletem integralmente o erro apresentado. Principalmente de vocabulários controlados muito grandes como o AGROVOC e EUROVOC, que necessitam de investigações detalhadas sobre cada tipo de erro devido a sua extensão. Por isso é importante ressaltar que os resultados obtidos nesse tipo de análise podem ser aprofundados em pesquisas futuras.

A análise Métrica de Termos estimula a possibilidade explorar os termos de um vocabulário controlado sobre o processamento de termos pós-seleção e validação de especialistas, identificando variações de comportamento textual que podem ser aplicados em termos.

A realização dos três tipos distintos de análises permitiu observar que a avaliação de vocabulários controlados pode ser realizada por meio de etapas previamente estabelecidas. As etapas de avaliação podem contemplar diversas características ou variáveis que compõem um vocabulário controlado. Mensurar ou criar métricas a partir de um objeto exige uma visão muito mais basilar do que estilos complexos de análises de difícil compreensão.

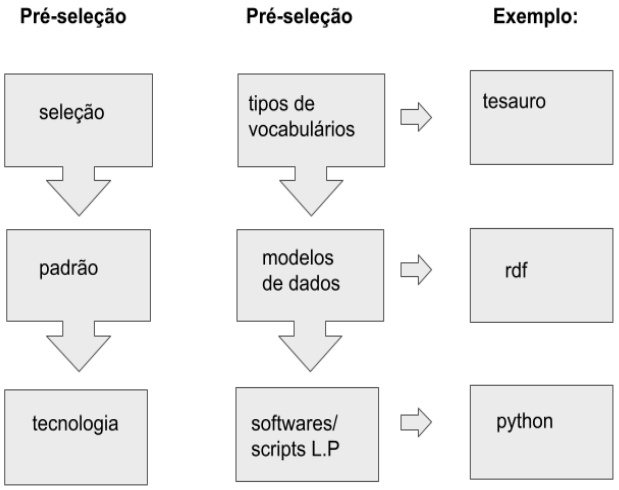
Embora o uso de ferramentas sofisticadas de análise façam parte do processo, é necessário relembrar que a lógica por trás de cada tipo de análise é baseada em contagem e operações matemáticas com uma sintaxe transliterada para uma linguagem de programação. Não obstante, realizar operações simples com números possui um nível de dificuldade, quando se trata de texto este nível aumenta, pois cada palavra é transformada em token que possui um valor e a partir do valor de cada token é possível extrair métricas.

A partir dos resultados obtidos foi possível verificar a possibilidade de avaliar um vocabulário controlado por meio de três formas distintas, a primeira é pode ser feita com uma contagem estatística. Uma análise estatística pode ser uma das principais formas de extrair informações, porque por meio dela é possível verificar a totalidade de um determinado objeto avaliado, assim como as partes que o compõem. Este tipo de análise pode ser fundamental para identificar o tamanho de um vocabulário controlado, sua estrutura de distribuição, alocação de termos e relacionamentos.

Não há uma pretensão de estabelecer regras para avaliação de vocabulários controlados, mas indicar de forma contributiva alguns princípios que podem ser opcionalmente adotados ou não, ficando inclusive a disposição para acréscimos de procedimentos ou etapas que melhor contemple o tipo de vocabulário controlado analisado. Por isso os três tipos de análises apresentadas oferecem perspectivas distintas que contribuem para o processo geral de avaliação de vocabulários controlados.

Antes de iniciar o processo de avaliação é necessária uma seleção de um tipo de vocabulário controlado, por exemplo, tesauros. A avaliação pode ser feita com um ou mais vocabulários controlados. Também é necessário reconhecer a forma que o vocabulário controlado foi publicado, se adere a algum tipo de padrão ou não. Superando essa etapa de seleção considera-se as possíveis ferramentas que possam realizar análises, que podem variar entre softwares específicos ou scripts personalizados com auxílio de uma Linguagem de Programação (L.P). Para ilustrar essas etapas a figura 47 apresenta um possível fluxo.

Figura 47 -Princípio de seleção e análise



Fonte: Elaborado pelo autor

Embora pareça em alguns casos como uma análise complexa, o processo em si não precisa ser de difícil compreensão, caso contrário, se tornaria difícil de reproduzir ou replicar, questionando portanto a viabilidade do processo. Outro ponto a ser considerado é o fato de utilizar instrumentos de representação que são objetos de estudo da Ciência da Informação, modelos de dados que são utilizados e reconhecidos na área e tecnologias em uso em ambientes acadêmicos, como é o caso da L.P python. Portanto é possível indicar alguns

princípios que podem ser ajustados para diferentes tipos de avaliação de vocabulários controlados.

O uso de uma linguagem de programação pode ter se destacado nesta pesquisa, porém esse tipo de análise não se restringe apenas a esses procedimentos ou ferramentas adotadas, por isso é possível sintetizar etapas que podem ser utilizadas em vocabulários controlados de forma geral, para implementação prática e objetiva, que consistem em:

1- Definir o tipo de vocabulário controlado que deseja analisar, pois será necessário limitar o escopo de avaliação.

2- Verificar a extensão do arquivo que armazena os dados do vocabulário controlado, por exemplo: .rdf, .xml, .json, .csv, .txt, entre outros. Procure por softwares ou L.P compatíveis com a extensão do vocabulário controlado escolhido, caso não possua uma extensão conhecida realize uma conversão para formatos populares ou que possuam maior suporte para resolução de problemas. Utilize softwares de conversão ou scripts para este fim.

3- Escolher uma ferramenta de software ou Linguagem de Programação popular para realizar suas análises, pois oferecerá maior suporte na resolução de possíveis intercorrências durante as análises.

4- Adotar alguma técnica de análise, por exemplo, NLP, porque é provável que já exista uma pacote de soluções em uso e com suporte para resolução de problemas.

5- Sintetizar os resultados como tabelas de dados, para facilitar a exportação e análise dos dados.

Cada tipo de software ou Linguagem de Programação escolhida apresentará uma interface de navegação distinta, por isso não há um modo pré-definido de navegação das etapas, mas por meio destes cinco passos. Para uma verificação analítica de cada vocabulário é necessário coletar o maior número possível de informações sobre o modelo de dados em que o vocabulário controlado foi disponibilizado e o software ou L.P escolhida, priorize sistemas ou pacotes nativos para o modelo escolhido, por exemplo a biblioteca rdfLib foi desenvolvida para a manipulação de arquivos RDF em python.

A análise estatística, nomeada neste trabalho como Análise Estrutural Quantitativa (AEQ), personifica a distribuição de propriedades e as principais características de um vocabulário controlado, determinando sua extensão Para realizar uma análise

estatística (AEQ) em vocabulários controlados considera-se os seguintes princípios que consistem em:

1- Realizar contagens básicas, com intuito de saber o número total de itens que contém no arquivo analisado.

2- Distribuir a contagem de acordo com as categorias, propriedades ou agrupamentos identificados no arquivo.

3- Validar a contagem por distribuição e verificar se ela é igual a contagem anterior.

Seguindo estas premissas além da contagem, obtém-se uma validação dos resultados obtidos. Ao sintetizar os resultados da contagem em tabelas é possível utilizar ferramentas de visualização de dados para ilustração dos resultados. A análise estatística AEQ é uma forma preliminar que indica se é possível prosseguir com análises mais complexas ou não, por isso podemos considerar como uma análise de viabilidade que sobretudo demonstra aplicabilidade e a composição do vocabulário analisado.

A análise Baseada em Modelos de Dados (BMD) fundamenta-se na AEQ, porém especializa-se no modelo de dados em que o vocabulário controlado foi publicado, por isso ela pode exigir uma pesquisa extensa sobre a documentação dos modelos de dados analisados, por exemplo o RDF e o SKOS são recomendações da W3C, logo essa organização organização internacional que desenvolve padrões para Web disponibiliza uma série de recursos sobre diversos tipos de modelos de dados, logo torna-se uma base de consulta obrigatória para este tipo de análise.

Para realizar uma análise BMD é necessário mapear todas as propriedades que compõem um determinado modelo de dados, assim como duas funções e divisões categóricas, identificando o papel que cada propriedade cumpre no processo de representação de um vocabulário controlado. Outro ponto que merece destaque é o fato de que muitos vocabulários controlados possuem outros esquemas de metadados embutidos, como o Dublin Core e muitos outros. Por isso, em muitos casos talvez seja necessário ampliar o conhecimento sobre outras tecnologias para completar sua análise.

A análise BMD objetiva a identificação de problemas associados a um determinado modelo de dados. A W3C já mapeou alguns problemas que podem ocorrer com diversos tipos de padrões para uso na web, assim como modelos de dados, com o SKOS e o RDF não é diferente, no entanto, não significa que necessariamente os problemas encontrados nos vocabulários controlados serão os mesmos. Porém, eles podem oferecer pistas para

identificar possíveis falhas na construção de um vocabulário controlado vinculado, como também o alinhamento às boas práticas para a publicação de dados vinculados. Os princípios que podem ser adotados em uma análise BMD consistem em:

1- Identificar o modelo de dados utilizado na representação do vocabulário escolhido e buscar informações de documentação técnica oficial na página do projeto ou instituição mantenedora do modelo de dados.

2- Buscar na literatura científica, publicações que enfatizam análises com os modelos de dados escolhidos, assim como tecnologias adotadas na pesquisa.

3- Procurar por problemas e erros conhecidos para o modelo de dados escolhido ou modelos similares, registrados na literatura ou nas documentações oficiais.

Seguindo essas 3 premissas e considerando os princípios apresentados anteriormente é possível construir um caminho para análise especializada, que supera a análise estatística e busca informações mais precisas sobre a estrutura na qual foi construído um vocabulário controlado. A união dessas duas etapas é suficiente para fornecer uma avaliação de vocabulário controlado, partindo de uma análise geral para uma análise especializada.

Por fim, o último tipo de análise, que se baseia na Métrica de Termos (MT) se mostrou a mais complexa, pois não se trata apenas de uma contagem, mas da transformação de palavras em números, para que sejam aplicadas fórmulas pré-definidas de cálculo textual, ou seja cada palavra se transforma em um token. Este tipo de análise objetiva apenas em extrair métricas de um conjunto de termos de um vocabulário controlado, com base na onometria, apresentada por Gilreath (1994), pois não se trata apenas de compilar técnicas de análise linguística, mas de contribuir para uma análise que possa ser incluída no escopo da Ciência da Informação. Por isso, a análise MT por de ser aplicada em três etapas que consistem em:

1- Extrair uma lista de termos de um vocabulário controlado, pois os termos já foram validados por especialistas.

2- Identificar técnicas de NLP com viabilidade de aplicação de processamento de texto em listas de palavras, ou que fundamentam medidas personalizadas.

3- Compilar os resultados de acordo com a escala numérica adotada, por exemplo, existem parâmetros de resultados que variam entre 0 e 1, enquanto outros podem variar de 0 a 100 ou mais.

Embora esse tipo de análise envolva o processamento de termos, ela não pode ser considerada uma análise semântica de termos, pois envolveria questões muito mais profundas de relacionamentos entre termos e contextualização de uso. No entanto a análise MT oferece suporte para implementações de análises mais complexas e sofisticada com o objetivo de encontrar problemas semânticos que poderiam estar atrelados aos vocabulários controlados

Os resultados obtidos e apresentados oferecem mais informações do que foi descrito neste trabalho, porém é possível tecer de forma resumida algumas considerações conclusivas dessa pesquisa, como apresentado a seguir.

6 CONCLUSÃO

Este trabalho encerra-se com algumas ponderações observadas ao longo da pesquisa, dentre elas podemos ressaltar que um vocabulário controlado é o resultado de um esforço conjunto de diversos especialistas que buscam representar uma determinada quantidade de conceitos de um domínio do conhecimento, utilizando termos significativos. Um vocabulário controlado possui uma função muito específica, que é auxiliar o processo de recuperação da informação, todavia, a sua continuidade e gerenciamento, impõe determinados desafios que podem ser desde a sua manutenção, atualização, suporte, estabelecimento de políticas de controle de termos, políticas de acesso, gestão dos dados que são inseridos no vocabulário, entre outros.

Outro aspecto que precisa ser destacado sobre os vocabulários controlados refere-se à manutenção, porque isso também define a sua continuidade. Embora não seja o foco desta pesquisa, reitera-se a importância deste tema para pesquisas futuras na área, pois isso diz respeito ao acesso futuro de um vocabulário controlado. Pois, diversos vocabulários podem ter surgido nos últimos anos e depois de um tempo simplesmente desapareceram, seja por falta de manutenção, porque o domínio de acesso expirou ou porque os dados foram perdidos, entre outros fatores. Por isso é muito importante que na etapa de planejamento, uma instituição, setor ou profissional que está desenvolvendo um vocabulário controlado inclua também um plano de manutenção para esse tipo de SOC.

Ao longo dos últimos anos inúmeras iniciativas têm sido propostas com o intuito de desenvolver e aperfeiçoar métodos e ferramentas que possam potencializar o uso de linguagens e vocabulários controlados em ambientes computacionais, buscando favorecer um melhor atendimento às demandas informacionais contemporâneas, originando um crescente número de pesquisas relacionadas ao desenvolvimento e uso de SOCs. A National Information Standards Organization (2017) enfatiza que muitos projetos atualmente estão sendo financiados para considerar os problemas de “grandes dados”, particularmente dados de pesquisa científica, todos dependem de vocabulários de metadados estáveis.

As normas que orientam a elaboração de vocabulários controlados são importantes instrumentos de padronização dos vocabulários controlados e podem ser utilizadas para nortear os métodos de avaliação dos mesmos. Assim como o uso de técnicas e ferramentas computacionais se apresentam como impulsionadores de análises automáticas em

massa. Esta pesquisa permitiu apresentar uma espécie de roteiro prático que pode ser aplicado no processo de avaliação de vocabulários controlados.

Dentre as dificuldade encontradas durante a pesquisa, ressalta-se a dificuldade de suporte no desenvolvimento dos códigos e erros constante no processamento dos arquivos de maior tamanho, porque mesmo em um ambiente específico e controlado para realização de análises de dados, dependendo da tarefa toda a memória RAM é ocupada, o que causava um travamento e reinício do processamento do arquivo. Ou seja, em computadores com memória RAM inferior a 12GB alguns tipos de vocabulários podem não ser processados ou o tempo de resposta de cada etapa do será muito longa, ao ponto de exceder o *timeout* de conexão do ambiente caso não haja outro tipo de interação ou resposta.

Por isso, que de certo modo a literatura científica diz que há um desafio na manipulação de grandes volumes de dados em ambientes limitados. Mas isto só é percebido quando os pesquisadores de fato lidam com esses dados. Outro ponto que merece destaque é o reconhecimento dos limites que os recursos computacionais oferecem, pois nem sempre se encontra suporte para o processamento de texto em todos os idiomas, por isso é importante que haja publicações de dados em idiomas nativos para ampliar os tipos de análises disponíveis.

Todavia, nem tudo se resume em dificuldades, pois a inclusão de registros históricos sobre conflitos terminológicos foi uma descoberta realizada durante a pesquisa bibliográfica e redação do referencial teórico. Apesar de exaustiva, esta etapa permitiu refinar a busca em documentos que agregaram substancialmente o dorso teórico da pesquisa. Assim como a constatação de que a onometria não possuía registros na literatura nacional e oferece um diálogo moderado com a terminologia, organização do conhecimento e avaliação de termos, que por consequência pode ser aplicado em vocabulários controlados.

Outro fator relevante para esta pesquisa reside no conhecimento dos modelos de dados sobre os quais estão disponibilizados os vocabulários controlados, porque a leitura das recomendações técnicas da W3C permitiram reconhecer os recursos necessários para verificação de cada etapa da análise, assim como o conhecimento da linguagem de programação python, que sobretudo oferece a infraestrutura para o desenvolvimento de códigos personalizados com base em bibliotecas que executam tarefas de alta complexidade.

Uma ação não mencionada anteriormente foi a leitura da documentação das bibliotecas rdfLib e NLTK, porque muito da lógica empregada na análise partiram das indicações constantes na própria documentação técnica. Embora não seja de fácil compreensão, porque pode exigir conhecimentos prévios é indispensável a consulta da documentação da biblioteca durante o processo de desenvolvimento.

Sobre os vocabulários controlados é possível concluir que o processo de avaliação pode ser realizado por meio de uma ou mais etapas, mas neste caso optou-se por três tipos para alcançar um maior nível de possibilidades e contextos de uso. Com a análise Estrutural Quantitativo foi possível perceber que o processo de contagem das tags permitiu um mapeamento da composição de cada vocabulário controlado, o que muitas vezes não é perceptível pelo desenvolvedor e muito menos para o usuário, a partir desse tipo de análise é possível constar a dimensionamento e o tamanho que esses aglomerados de termos conceituais possuem.

A partir da análise Baseada em Modelos de Dados foi possível verificar a estrutura de cada vocabulário controlado e como o crescimento deles podem implicar no surgimento de mais erros. Que podem ser desde repetições a URIs que não estão mais ativas. Por isso é possível afirmar que quanto menor o vocabulário controlado menor a possibilidade de problemas e erros. No entanto, o surgimento de erros não pode ser um impedimento para a evolução de um vocabulário controlado, pois é necessário lidar com os erros e buscar soluções que possam corrigi-los.

A análise métrica de termos foi uma das etapas mais difíceis, pois os cinco tipos de métricas apresentados foram resultados de tentativa e erro de dezenas recursos que não apresentaram resultados satisfatórios durante o desenvolvimento do código, ou seja foram as únicas opções que restaram que garantia o mínimo de viabilidade, tendo em vista a limitação idiomática da biblioteca e a ausência de padrões de referência que pudessem nortear os resultados apresentados.

Por isso é importante ressaltar que os resultados apresentados nas análises Baseada em Modelos de Dados e métrica de termos possuem limitações, não obstantes podem ser passíveis de erros embora não tenha sido determinada formalmente a margem de precisão de dos resultados. No entanto é possível afirmar que somente foram apresentados tipos de erros e métricas que se mostram minimamente viáveis, ou seja foi possível constatar onde estava o erro identificado, como foi demonstrado no AGROVOC, assim como na análise

métrica foi aplicada somente o processamento das palavras extraídas, sem nenhuma contaminação de caracteres além dos idiomas não contemplados, que simplesmente são ignorados durante o processamento automático.

Os princípios de avaliação de vocabulários controlados apresentados nos resultados dessa pesquisa foram particularmente motivados pela onometria, que pode ser entendida como um princípio de avaliação de termos, a elaboração desses princípios foram redigidos com base na trajetória de análise realizada em cada etapa. O quadro 12 apresenta de forma resumida os princípios que podem ser aplicados em cada tipo de análise.

Quadro 12 – Princípios de avaliação

Requisitos	**AEQ**	**BMD**	**MT**
Tipo de vocabulário controlado	Contagem inicial	Identificação de modelos de dados	Extração de lista de termos
Extensão do arquivo	Distribuição da contagem	Registros na literatura científica	Técnicas de NLP
Categorias, propriedades ou agrupamentos	Validação da contagem	Mapeamento de erros e problemas	Compilação de resultados

Fonte: Elaborado pelo autor

Os princípios apresentados podem ser utilizados integralmente ou parcialmente, isso dependerá da ênfase de análise que se pretende explorar no processo de avaliação de um vocabulário controlado. Também é possível selecionar princípios distintos de cada tipo de análise para desenvolver formas alternativas de avaliação. Por isso, não há uma pretensão de estabelecer regras de avaliação, mas de contribuir com princípios identificados no decurso desta pesquisa.

De modo geral a avaliação dos vocabulários controlados permitiu uma aproximação desse tipo de SOC, sob uma ótica exploratória. Revelando dificuldades teóricas e práticas no desenvolvimento e na avaliação de vocabulários controlados. Diferente das análises específicas realizadas em taxonomias e tesauros, esse tipo de avaliação pode ser implementado em qualquer tipo de vocabulário controlado codificado em RDF ou SKOS.

Observa-se que todos os objetivos foram alcançados e sobre avaliação de vocabulários controlados se mostrou sobretudo, viável.

Os dados resultantes dessa análise foram disponibilizados em uma plataforma de hospedagem de código-fonte e arquivos com controle de versão, para acesso aos arquivos dos resultados apresentados e do plano de dados[8], tendo em vista que estes vocabulários controlados são atualizados e podem possuir versões distintas futuramente. O que pode dificultar a verificação futura dos dados resultantes dessa pesquisa. Por isso espera-se que este trabalho possa oferecer uma pequena contribuição para a Ciência da Informação e todos os campos de estudos que possuem alguma relação com esta pesquisa.

Há um grande desafio na análise de dados que são disponibilizados em quantidades cada vez maiores, limitando o uso por falta de conhecimento, técnica ou hardwares com poder de processamento mais robustos. O fato é que mesmo com essas dificuldades é possível vislumbrar um futuro em que esses tipos de análise não serão tão complexas ou pertencentes a um seleto grupo de pesquisadores.

[8] Ver Apêndice B

REFERÊNCIAS

ALMEIDA, M. B.; SOUZA, R.R.; FONSECA F., The semantic in the Semantic Web: a critical evaluation. **Knowledge Organization**. v.38, n. 3, 2011. p 187-203.

ALONSO-ARROYO, A.; FUJITA, M. S. L.; GIL-LEIVA, I.; PANDIELLA, A. Protocolo verbal: análisis de la producción científica, 1941-2013. **Informação & Sociedade**: Estudos; v. 26, n. 2, 2016. Disponível em: http://hdl.handle.net/20.500.11959/brapci/93158. Acesso em: 26 out. 2023.

ALVAREZ, G. R.; CAREGNATO, S. E. A ciência da informação e sua contribuição para a avaliação do conhecimento científico. **BIBLOS**, [S.l.], v. 31, n. 1, p. 09-26, ago. 2017. Disponível em: https://doi.org/10.14295/biblos.v31i1.5987. Acesso em: 17 jun. 2020.

ALVES, J. B. da M. **Teoria geral de sistemas: em busca da interdisciplinaridade.** Florianópolis: Instituto Stela, 2012. 179p.

AQUINO, I. J.; CARLAN, E.; BRASCHER, M. B. Princípios classificatórios para a construção de taxonomias. **Ponto de acesso**, Salvador, v. 3, n. 3, p. 196-215, dez. 2009. Disponível em: https://portalseer.ufba.br/index.php/revistaici/article/view/3626. Acesso em: 15 mai. 2020.

AUBERT, F. H. **Introdução à metodologia da pesquisa terminólogica bilíngue**. 2. ed. São Paulo: FFLCH/CITRAT, 2001.

BABINI, M. Do conceito à palavra: os dicionários onomasiológicos. **Cien. Culto**, São Paulo, v. 58, n. 2, pág. 38-41, junho de 2006 . Disponível em: http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252006000200015&lng=en&nrm=iso. Acesso em: 03 de fev. 2022.

BAKER, T.; BECHHOFER, S.; ISAAC, A.; MILES, A.; SCHREIBER, G.; SUMMERS, E. Key Choices in the Design of Simple Knowledge Organization System (SKOS). **Journal of Web Semantics**, v.20, 2013. Disponível em: https://www.sciencedirect.com/science/article/pii/S1570826813000176?via%3Dihub. Acesso em: 10 mar. 2021

BAKER, T.; VANDENBUSSCHE, P. Y.; VATANT, B. Requirements for vocabulary preservation and governance. **Library Hi Tech**, v. 31, n. 4, p. 657-668, 2013. Disponível em: https://doi-org.ez31.periodicos.capes.gov.br/10.1108/LHT-03-2013-0027. Acesso em: 15 mai. 2020.

BARBOSA, Alice Príncipe. **Teoria e prática dos sistemas de classificação bibliográfica**. Rio de Janeiro: Instituto Brasileiro de Bibliografia e Documentação, 1969.

BARDlN, L. **Análise de conteúdo**. Lisboa: Edições 70, 1977, 225p.

BARITÉ, M. La garantía literaria: vigencia y proyección teóricometodológica. Encontro Nacional de Pesquisa em Ciência da Informação, 8, **Anais**... Salvador: ENANCIB, 2007. Disponível em: http://www.enancib.ppgci.ufba.br/artigos/GT2--068.pdf. Acesso em: 15 mai. 2020.

BARITÉ, M. Sistemas de Organização do Conhecimento: uma tipologia atualizada. **Informação & Informação**, [S.l.], v. 16, n. 2, p. 122-139, dez. 2011. ISSN 1981-8920. Disponível em: https://www.uel.br/revistas/uel/index.php/informacao/article/view/9952. Acesso em: 17 fev. 2022.

BARROS, L. A. Aspectos epistemológicos e perspectivas científicas da terminologia. **Ciência e Cultura**, São Paulo, v. 58, n. 2, p. 22-26, abr./jun. 2006.

BARROS, L. A. **Curso Básico de terminologia**. São Paulo: Edusp, 2004.

BERGMAN, M. K. **An Intrepid Guide to Ontologies**. 2007. Disponível em: www.mkbergman.com//?p=374. Acesso em 26 de abril de 2021.

BERNERS-LEE, Tim; CONNOLLY, Dan. **Notation3 (N3)**: a readable RDF syntax. 2011. Disponível em: http://www.w3.org/TeamSubmission/2011/SUBM-n3-20110328/. Acesso em: 12 dez. 2021.

BEVILACQUA, C. R.; FINATTO, M. J. B. Lexicografia e terminografia: alguns contrapontos fundamentais. **Alfa**, São Paulo, v. 50, n. 2: 43-54, 2006. Disponível em: http://www.ufrgs.br/textecc/textquim/arquivos/03-Bevilacqua-Finatto.pdf. Acesso em: 12 dez. 2021.

BIDERMAN, M. T. C. A ciência da lexicografia. **Alfa**, São Paulo, v. 28, 1984.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural Language Processing with Python**: analyzing text with the Natural Language Toolkit. O'Reilly Media: California, 2009. Disponível em: https://www.nltk.org/book/. Acesso em: 22 abr 2021.

BÖRNER, K.; HARDY, E.; HERR, B.; HOLLOWAY, T.; PALEY, W. B. Taxonomy visualization in support of the semi-automatic validation and optimization of organizational schemas. **Journal of Informetrics**, v.1, n.3, 2007, p. 214-225. Disponível em: https://doi.org/10.1016/j.joi.2007.03.002. Acesso em: 05 mai. 2020.

BRICKLEY, D.; GUHA, R.V. **RDF Schema 1.1**. W3C Recommendation, 2014. Disponível em: https://www.w3.org/TR/2014/REC-rdf-schema-20140225/. Acesso em: 12 dez. 2021.

CABRÉ CASTELLVÍ, M. T. Teorias da terminologia: descrição, prescrição e explicação. Tradução de Diego Napoleão Viana Azevedo. **Cad. Trad**., Florianópolis, v. 39, nº 3, p. 507-558, set-dez, 2019. Disponível em: https://doi.org/10.5007/2175-7968.2019v39n3p507. Acesso em: 12 dez. 2021.

CABRÉ, M. T. A Terminologia, uma disciplina em evolução: passado, presente e alguns elementos de futuro. **Debate Terminológico**. n. 01 .2005. Disponível em:https://seer.ufrgs.br/riterm/article/view/21286. Acesso em: 05 mai. 2020.

CABRÉ, M. T. La terminología hoy:concepciones, tendencias y aplicaciones. **Ciência da Informação** - v. 24, número 3, 1995.

CABRÉ, M. T. **La terminología: Representación y comunicación.** Uma propuesta de base comunicativa y otros artículos. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. 1999.

CAMBRIA E.; WHITE, B. Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]", **IEEE Computational Intelligence Magazine**, v. 9, n. 2, p. 48-57, mai., 2014. Disponível em: 10.1109/MCI.2014.2307227. Acesso em: 15 mai. 2020.

CAMPOS, M. L. A.; GOMES, H. E. Metodologia de elaboração de tesauro conceitual: a categorização como princípio norteador. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 11, n. 3, set/dez. 2006. Disponível em: http://www.eci.ufmg.br/pcionline/index.php/pci/article/viewFile/273. Acesso em: 15 mai. 2020.

CATARINO, M. E. Simple Knowledge Organization System: construindo sistemas de organização do conhecimento no contexto da Web Semântica. **Informação & Tecnologia (ITEC)**: Marília/João Pessoa, 1(1): p.17-28, jan./jun., 2014.

CAVALCANTE, R. da S. **Critérios para a avaliação de taxonomias navegacionais em sítios de comércio eletrônico**. 2012. 88 f., il. Dissertação (Mestrado em Ciência da Informação) - Universidade de Brasília, Brasília, 2012.

CLARK, Kendall Grant; FEIGENBAUM, Lee; TORRES, Elias . **SPARQL Protocol for RDF**. W3C Recommendation, 2008. Disponível em: https://www.w3.org/TR/2008/REC-rdf-sparql-protocol-20080115/. Acesso em: 12 nov. 2022.

CYGANIAK, Richard; WOOD, David; LANTHALER, Markus. **RDF 1.1 Concepts and Abstract Syntax**. W3C Recommendation, 2014. Disponível em: https://www.w3.org/TR/rdf11-concepts/. Acesso: 16 fev. 2022.

CONTANDRIOPOULOS, A. P. Avaliando a institucionalização da avaliação. **Ciência & Saúde Coletiva**, v.11, n.3, p.705-711, 2006.

CONTANDRIOPOULOS, A. P.; CHAMPAGNE, F., DENIS, J. L.; PINEAULT, R. **A avaliação na área da saúde**: conceitos e métodos. In:_____. HARTZ, ZMA., (Org.). Avaliação em Saúde: dos modelos conceituais à prática na análise da implantação de programas [online]. Rio de Janeiro: Editora FIOCRUZ, 1997. 132 p.

COSTA, M. A. F.; COSTA, M. de F. B. **Metodologia da pesquisa**: conceitos e técnicas. Rio de Janeiro: Interciência, 2001.

COUTO, H. H. do. Onomasiologia e semasiologia revisitadas pela ecolinguística. **Revista de Estudos da Linguagem**, [Sl], v. 20, n. 2, pág. 183-210, dez. 2012. ISSN 2237-2083. Disponível em: http://periodicos.letras.ufmg.br/index.php/relin/article/view/2748/2703. Acesso em: 16 fev. 2022.

CUNHA, Murilo Bastos da; CAVALCANTI, Cordélia Robalinho de Oliveira. Dicionário de biblioteconomia e arquivologia. Brasília: Briquet de Lemos, 2008. 451p.

CURRÁS, E. **Ontologias, taxonomia e tesauros em teoria de sistemas e sistemática**. Tradução de Jaime Robredo. Brasília: Thesaurus, 2010. 182 p.

DACONTA, M. C.; OBRST, L. J.; SMITH, K. T. **The Semantic Web**: A guide to the future of XML, Web Services and Knowledge Management. Indianapolis: Willey. 2005, 312p.

DAHLBERG, I. Teoria do conceito. **Ciência da Informação**, Rio de Janeiro, v.7, n.2, 1978. p.101- 107.

DESCARTES, R. **Discurso do método**. Tradução de Maria Ermantina Galvão. São Paulo: Martins Fontes, 1996.

DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **arXiv e-prints**, 2018. Disponível em: https://arxiv.org/pdf/1810.04805.pdf. Acesso em: 10 out. 2021.

DONGO, I.; CARDINALE, Y.; CHBEIR, R. RDF-F: RDF Datatype inFerring Framework. **Data Sci. Eng**. n. 3, 2018. Disponível em: https://doi-org.ez67.periodicos.capes.gov.br/10.1007/s41019-018-0064-6. Acesso em: 17 jun. 2020.

DUBUC, R. **Manuel pratique de terminologie**. 4. ed. Canadá: Linguatec, 2002.

DYER, W. T. Thiselton. "Botanical Terminology." **Nature**, London, v.6, n.153, 1872. p.455. Disponível em: https://search.library.wisc.edu/digital/ALBXITYVRTMAPI83. 1869-2021. Acesso em: 15 jan. 2021.

FAO. **The AGROVOC Editorial Guidelines 2020**. 2 ed, Rome, 2022. Disponível em: https://doi.org/10.4060/cb8640en. Acesso em: 16 mai. 2023.

FAULSTICH, E. L. de J. Terminologia: o Projeto Brasilterm e a formação de recursos humanos**. Ciência da Informação**, [S. l.], v. 24, n. 3, 1995. DOI: 10.18225/ci.inf.v24i3.581. Disponível em: https://revista.ibict.br/ciinf/article/view/581. Acesso em: 26 out. 2023.

FELBER. H.. **Terminology Manual**. Lnfoterm: Vienna, 1989.

FERREIRA, L. B.; ROCKEMBACH, M. Abordagens contemporâneas sobre avaliação em Arquivologia e Ciência da Informação: macroavaliação, avaliação do fluxo informacional e modelo indício-evidência-prova. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 22, n. 50, p. 31-43, set. 2017. ISSN 1518-2924. Disponível em: https://doi.org/10.5007/1518-2924.2017v22n50p31. Acesso em: 17 jun. 2020.

FIGUEIREDO, C. de. **Novo dicionário da língua portuguesa**. 2010.

FRAZIER, Priscila Jane. **SKOS**: A Guide for Information Professionals. American Library Association, 2015. Disponível em: http://www.ala.org/alcts/resources/z687/skos. Acesso em: 11 nov. 2022

FRUIT CLASSIFICATION. **Nature,** London, v. 4. n.96, 1871, p. 347-48. Disponível em:. https://search.library.wisc.edu/digital/AHBPSYKXSCSVKR8M/pages/A2FIHXWFCEUSDF9B. Acesso em: 15 jan. 2021.

FUJITA, Mariângela Spotti Lopes; BOCCATO, Vera Regina Casari; RUBI, Milena Polsinelli; GONÇALVES, Maria Carolina. (Org). **A indexação de livros: a percepção de catalogadores e usuários de bibliotecas universitárias**: um estudo de observação do contexto sociocognitivo com protocolos verbais [online]. São Paulo: Editora UNESP; São Paulo: Cultura Acadêmica, 2009. 149 p.

GANDON, Fabien; SCHREIBER, Guus. **RDF 1.1 XML Syntax**. W3C Recommendation, 2014. Disponível em: https://www.w3.org/TR/rdf-syntax-grammar/. Acesso em: 17 jun. 2020.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. - São Paulo: Atlas, 2008.

GILREATH, C.T., The Semantic Valence of Terms: A Systematic Treatment of Multi-Meaning Terms, Standardizing and Harmonizing Terminology: Theory and Practice, **American Society for Testing and Materials, Philadelphia (ASTM)**, 1994.

GILREATH, Ch,T,: Harmonization of terminology. An overview of principles**. Int.Classif**. 19 No.3, 1992, p. 135-139.

GILREATH. C. T. Onometrics: The Formal Evaluation of Terms, Standardizing Terminology for Better Communication: Practice, Applied Theory, and Results. **American Society for Testing and Materials. Philadelphia (ASTM)**, 1993, p. 75-94.

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

GOMES, H. E. Terminologia e estrutura conceitual. **Ponto de Acesso**, [S. l.], v. 15, n. 3, 2021. DOI: 10.9771/rpa.v15i3.47464. Disponível em: https://periodicos.ufba.br/index.php/revistaici/article/view/47464. Acesso em: 17 fev. 2022.

GOMES, H. E.; CAMPOS, M. L. A.; GUIMARÃES, L. D. S. Organização da informação e terminologia: a abordagem onomasiológica. **DataGramaZero**, v. 11, n. 5, 2010 . Disponível em: http://hdl.handle.net/20.500.11959/brapci/7184. Acesso em: 26 out. 2023.

GOMES, H. E.; CAMPOS, M. L. de A. Tesauro e normalização terminológica: o termo como base para intercâmbio de informações. **DataGramaZero** - Revista de Ciência da Informação - v.5 n.6 dez/2004.

GREFENSTETTE, G. Sextant. In: Explorations in Automatic Thesaurus Discovery. **The Springer International Series in Engineering and Computer Science**, vol 278. Boston: Springer, 1994. Disponível em: https://doi.org/10.1007/978-1-4615-2710-7_3. Acesso em: 17 fev. 2022.

GREFENSTETTE, Gregory. Sextant: exploring unexplored contexts for semantic extraction from syntactic analysis. In **Proceedings**… of the 30th annual meeting on Association for Computational Linguistics (ACL '92). Association for Computational Linguistics, USA, 1992, p.324–326. Disponível em: https://doi.org/10.3115/981967.982020. Acesso em: 17 fev. 2022.

GUARINO, N. Ontology and Terminology: how can formal ontology help concept modeling and terminology? In **Proceedings**… EAFT- NordTerm ws on Terminology, Concept Modeling and Ontology, Vaasa, February 10th, 2006.

GUTHRIE , Louise ; GUTHRIE , Joe; LEISTENSNIDER, Jam. **Document classification and routing**. In_____: STRZALKOWSKI, Tomek (Ed). Natural Language Information Retrieval. [s.l.]: Springer Science, 1999.

HAAS, S. W.; TRAVERS, D.; TINTINALLI, J.E.; POLLOCK, D.; WALLER, A.; BARTHELL, E.; BURT, C.; CHAPMAN, W.; COONAN, K.; KAMENS, D.; MCCLAY, J. Toward Vocabulary Control for Chief Complaint. **Academic Emergency Medicine**, v.15, n.5,

p. 476-482, 2008. Disponível em: https://doi.org/10.1111/j.1553-2712.2008.00104.x. Acesso em: 05 mai. 2020.

HADJI, C. **Avaliação desmistificada**. Porto Alegre: Artmed, 2001.

HARRIS, Steve; SEABORNE, Andy. **SPARQL 1.1 Query Language**. W3C Recommendation, 2013. Disponível em: https://www.w3.org/TR/sparql11-query/. Acesso em 05 mai 2020.

HAYES, P. J.; PATEL-SCHNEIDER, P. F. RDF 1.1 Semantics. W3C Recommendation 25 February, 2014. Disponível em: https://www.w3.org/TR/2014/REC-rdf11-mt-20140225/. Acesso em 12 dez. 2021.

HODGE, G. **Systems of Knowledge Organization for Digital Libraries**: Beyond Traditional Authority Files. 2000. Disponível em: www.clir.org/pubs/abstract/pub91abst.html. Acesso em: 17 dez. 2020.

HULLAH, John. On Musical Nomenclature. **Proceedings of the Musical Association**, vol. 1, 1874, p. 74–87. Disponível em: http://www.jstor.org/stable/765408. Accessed 22 Oct. 2022.

INDIA. Ministry of Education. Standing Commission for Scientific and Technical Terminology. **Evolving a National Terminology**: a Monograph. New Delhi: Standing Commission for Scientific and Technical Terminology, Ministry of Education, Govt. of India, 1968.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 25964-1**: **Thesauri and interoperability with other vocabularies**. Part 1: Thesauri for information retrieval. Geneve: International Standard Organization, 2011.

ISAAC, A.; SUMMERS, Ed. (Ed.). **SKOS Simple Knowledge Organization System Primer**: W3C Working Group Note, 18 August 2009. Disponível em: https://www.w3.org/TR/2009/. Acesso em: 17 dez. 2020.

ISO - INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 704**: **Terminology work** — Principles and methods, Switzerland: ISO copyright office, 2022.

ISO - INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 25964-2**: **Information and documentation —** Thesauri and interoperability with other vocabularies — Part 2: Interoperability with other vocabularies. Geneve: International Standard Organization, 2013.

JACQUEMIN, Christian; TZOUKERMANN, Evelyne. Nlp For Term Variant **Extraction: Synergy Between Morphology, Lexicon, And Syntax**. 1999. Disponível

em:10.1007/978-94-017-2388-6_2. In_____: STRZALKOWSKI, Tomek (Ed). Natural Language Information Retrieval. [s.l.]: Springer Science, 1999.

KARLGREN, Jussi. **Stylistic experiments in information retrieval**. In_____: STRZALKOWSKI, Tomek (Ed). Natural Language Information Retrieval. [s.l.]: Springer Science, 1999.

KIM, S.; OH, S. Extracting and applying evaluation criteria for ontology quality assessment. **Library Hi Tech** , v. 37, n. 3, p. 338-354, 2019. Disponível em: https://plosjournal.deepdyve.com/lp/emerald-publishing/extracting-and-applying-evaluation-criteria-for-ontology-quality-QrlNHKyEb6?impressionId=5db6377af3e0b&i_medium=docview&i_campaign=recommendations&i_source=recommendations. Acesso em: 05 abr. 2020.

KITCHENER, Frank E. Botanical Terminology. **Nature,** London, v.6, n.151, 1872, p.413-414. Disponível em: https://search.library.wisc.edu/digital/ALBXITYVRTMAPI83. Acesso em: 05 mar. 2021.

KLESS, D.; MILTON, S. Towards quality measures for evaluating thesauri. In: . Metadata and Semantic Research: 4th International Conference, MTSR 2010, Alcalá de Henares, Spain, October 20-22, 2010. **Proceedings**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 312–319.

KOBASHI, N. Y. **Vocabulário controlado: estrutura e utilização**. [s.l]: ENAP, 2008.

KOBASHI, N. Y.; SMIT, J. W.; TÁLAMO, M. F. G. M. A função da terminologia na construção do objeto da ciência da informação. **DataGramaZero**, v. 2, n. 2, 2001. Disponível em: http://hdl.handle.net/20.500.11959/brapci/4867. Acesso em: 01 dez. 2022.

KRIEGER, M. G.; FINATTO, M. J. B. **Introdução à Terminologia**: teoria & prática. São Paulo: Contexto, v. 1, 2004.

LANCASTER, F.W. **El control del vocabulario en la recuperación de la información**. 2ª ed. Valencia: Universidad de Valencia. 2002.

LARA, M. L. G. de. Propostas de tipologias de KOS: uma análise das referências de formas dominantes de organização do conhecimento. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 20, n. esp. 1, p. 89-107, fev., 2015.

LASSILA, O.; MCGUINESS, D. L. **The Role of Frame-Based Representation on the Semantic Web.** Knowledge Systems Laboratory Report KSL-01-02, Stanford University, 2001.

LEITÃO, D. A. **NLForSpec**: uma ferramenta para geração de especificações formais a partir de casos de teste em linguagem natural. Recife, 2006. Dissertação (mestrado) - Universidade Federal de Pernambuco, 2006.

LEMOS, M. L. V. de. Desenvolvimento de um vocabulário controlado na biblioteca do senado federal. **Ci. Inf.,** Brasília, v.15, n. 2, p. 155-58, jul./dez. 1986.

LEWIS, A. L. On the Evils Arising from the Use of Historical National Names as Scientific Terms. **The Journal of the Anthropological Institute of Great Britain and Ireland**, vol. 8, 1879, pp. 325–35. Disponível em: https://doi.org/10.2307/2841047. Acesso em: 23 mar. 2022.

LOPES, l. L. de A. S. **Análise do uso das linguagens controlada e livre nas estratégias de busca em bases de dados**. 2000. 111 f., il. Dissertação (Mestrado em Ciência da Informação) - Universidade de Brasília, Brasília, 2000.

LORENZON, E. J. **Análise de domínio para avaliação de tesauros**: uma experiência com a cadeia produtiva do calçado no Brasil. 2011. 108 f. Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2011. Disponível em: http://hdl.handle.net/11449/103385. Acesso em: 15 mai. 2020.

MADER, Christian; HASLHOFER, Bernhard; ISAAC, Antoine. Finding quality issues in SKOS vocabularies. In: International Conference on Theory and Practice of Digital Libraries. **Proceedings**...Springer, Berlin, Heidelberg, 2012. p. 222-233.

MANAF, N. A. A.; BECHHOFER, S.; STEVENS, R. **The Current State of SKOS Vocabularies on the Web**. Springer-Verlag: Berlin Heidelberg, 2012. p. 270–284.

MANOLA, F.; MILLER, E; MCBRIDE, B. **RDF Primer**. W3C Nota do Grupo de Trabalho W3C24, 2014. Disponível em: https://www.w3.org/TR/rdf11-primer/. Acesso em: 10 jun. 2021.

MARTÍNEZ GARCÍA, S.: A representação e organização da informação através do dicionário de sinônimos. A interdisciplinaridade como um novo paradigma: desafios para a documentação e a biblioteconomia, **Contribuições para as ciências sociais**, 2009. Disponível em: www.eumed.net/rev/cccss/06/smg.htm. Acesso em: 15 mai. 2020.

MARTÍNEZ-ÁVILA, Daniel; BUDD, John M. Epistemic warrant for categorizational activities and the development of controlled vocabularies. **Journal of Documentation**, v. 73, n. 4, p.700-715, 2017. Disponível em: https://doi.org/10.1108/JD-10-2016-0129. Acesso em: 05 mai. 2020.

MASTORA, A.; PEPONAKIS, M.; KAPIDAKIS, S. SKOS concepts and natural language concepts: an analysis of latent relationships in KOSs. **Journal of Information Science**, v. 43, n.4, p. 492–508. 2017.

MAZZOCCHI, Fulvio. Knowledge organization system (KOS). **Knowledge Organization** 45, no.1: 54-78. Also available in ISKO Encyclopedia of Knowledge Organization, eds. Birger Hjørland and Claudio Gnoli, 2018. Disponível em: http://www.isko.org/cyclo/kos. Acesso em: 20 jun 2021.

MILES, A.; BECHHOFER, S. **SKOS Simples Knowledge Organization System** Reference. W3C Recommendation, 2009. Disponível em: http://www.w3.org/TR/skos-reference/. Acesso em: 20 jun 2021

MOREIRA, Manoel Palhares; MOURA, Maria Aparecida. Construindo tesauros a partir de tesauros existentes: a experiência do TCI - Tesauro em Ciência da Informação. **DataGramaZero** - Revista de Ciência da Informação - v.7 n.4 ago/2006.

MOREIRO GONZÁLEZ, J. A. **Linguagens documentárias e vocabulários semânticos para a web: elementos conceituais**. Salvador: EDUFBA, 2011. 128 p.

MORIN, Edgar. **Os sete saberes necessários à educação do futuro**.Tradução de Catarina Eleonora F. da Silva e Jeanne Sawaya, 2. ed. – São Paulo : Cortez, 2000.

NADKARNI, Prakash M; OHNO-MACHADO, Lucila; CHAPMAN , Wendy W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association**, Volume 18, Edição 5, setembro de 2011, p. 544–551. Disponível em: https://doi.org/10.1136/amiajnl-2011-000464. Acesso 10 jun 2021.

NAYAK, G.; DUTTA, S.; AJWANI, D.; NICHOLSON, P.; SALA, A. Automated assessment of knowledge hierarchy evolution: comparing directed acyclic graphs. **Inf Retrieval J**, v.22, p.256-284, 2019. Disponível em: https://doi-org.ez31.periodicos.capes.gov.br/10.1007/s10791-018-9345-y. Acesso em: 15 mai. 2020.

NIELSEN, J. **Usability Engineering**. Boston: Academic Press, Cambridge, MA, 1993.

NIELSEN, J.; MOLICH, R. **Heuristic evaluation of user interfaces**. Proc. ACM CHI'90 Conf., Seattle, EUA, 1-5 abril, p. 249-256, 1990.

NISO - NATIONAL INFORMATION STANDARDS ORGANIZATION . **Guidelines for the construction, format, and management of monolingual controlled vocabularies**. Baltimore: NISO, 2010. 172 p.

NISO - NATIONAL INFORMATION STANDARDS ORGANIZATION . **Issues in Vocabulary Management**: a technical report of the National Information Standards Organization, 2017. Disponível em: https://groups.niso.org/apps/group_public/download.php/18410/NISO_TR-06-2017_Issues_in_Vocabulary_Management.pdf. Acesso em: 15 mai. 2020.

NUNES, Luis Pablo. Michael Toxites y los vocabularios plurilingües de Onomastica (1574). **New Journal of Hispanic Philology**, v.67, n. 1, 2019. Disponível em: https://doi.org/10.24201/nrfh.v67i1.3463. Acesso em 10 jun 2021.

NYBAKKEN, O. **Greek and Latin in Scientific Terminology**. Iowa State University Press, 1959.

OBRST, L. **Ontologies & the Semantic Web for Semantic Interoperability**. 2004 Disponível em: www.web-services.gov/OntologiesSemanticWebSemInteropSICOP909-Obrst.ppt. Acesso em: 15 mai. 2020.

PAIM, I.; NEHMY, R. M. Q. Questões sobre a avaliação da informação: uma abordagem inspirada em Giddens. **Perspectivas em Ciência da Informação**, [S.l.], v. 3, n. 2, abr. 2008. ISSN 19815344. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/598/367. Acesso em: 17 jun. 2020.

PAIVA, P. T. P. A origem da nomenclatura médica, o texto científico e suas implicações para a tradução médica. **Revista ECOS**, *[S. l.]*, v. 6, n. 1, 2007. Disponível em: https://periodicos2.unemat.br/index.php/ecos/article/view/986. Acesso em: 22 out. 2022.

PARK, J.-R.; RICHARDS, L.L.; BRENZA, A. Benefits and challenges of BIBFRAME: Cataloging special format materials, implementation, and continuing educational resources. **Library Hi Tech**, v. 37, n. 3, pp. 549-565, 2019. Disponível em: https://doi-org.ez31.periodicos.capes.gov.br/10.1108/LHT-08-2017-0176. Acesso em: 05 mai. 2020.

PASTOR-SANCHEZ, Juan-Antonio; MARTÍNEZ-MENDEZ, Francisco Javier; RODRÍGUEZ-MUÑOZ, José Vicente. Advantages of thesauri representation with the Simple Knowledge Organization System (SKOS) compared with other proposed alternatives. **Information Research**, v. 14, 2009. Disponível em: https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/epi.2012.may.04. Acesso em: 23 out. 2023.

PEIRCE, Charles S. **Semiótica**. Tradução de José Teixeira Coelho Neto. São Paulo: Perspectiva, 2015.

PHARMACOPŒIAL NOMENCLATURE. **British Medical Journal**. n. 1.536, 1871. p. 373-74. Disponível em: https://doi-org.ez67.periodicos.capes.gov.br/10.1136/bmj.1.536.373. Acesso em: 10 jan. 2023.

PINTO, M. A user view of the factors affecting quality of thesauri in social science databases. **Library & Information Science Research**, v. 30, p. 216–221, 2008.

PINTO, Mariana de Azevedo. Entropia informacional e desinformação – um estudo acerca da organização da informação aplicada no sistema de informação do Programa Mais Médico. Dissertação (Mestrado) Universidade Federal da Bahia, Instituto de Ciência da Informação, 2017. 111f.

PRUD'HOMMEAUX, Eric; CAROTHERS, Gavin. **RDF 1.1 Turtle**: Terse RDF Triple Language. W3C Recommendation, 2014. Disponível em: https://www.w3.org/TR/turtle/. Acesso em: 05 mai. 2020.

QUARATI, A.; ALBERTONI, R.; MARTINO, M. de. Overall quality assessment of SKOS thesauri: An AHP-based approach. **Journal of Information Science,** v. 43, n.6, p. 816-834, 2016. Disponível em: https://doi-org.ez31.periodicos.capes.gov.br/10.1177/0165551516671079. Acesso em: 05 mai. 2020.

RAMALHO, R. A. S. **Desenvolvimento e utilização de ontologias em bibliotecas digitais**: uma proposta de aplicação. 2010. 145f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010.

RAMALHO, R. A. S. **Ontologias e Simple Knowledge Organization System (SKOS)**: aproximações e diferenças. In: José Augusto Chaves Guimarães; Vera Dodebei. (Org.). Organização do conhecimento e diversidade cultural. 1ed. Marília: ISKO-Brasil; FUNDEPE, v. 1, p. 100-107, 2015.

RAMALHO, Rogério Aparecido Sá; SOUSA, Janailton Lopes. Diretrizes para avaliação de sistemas de organização do conhecimento representados em SKOS. **Inf. Inf**., Londrina, v. 24, n. 2, mai/ago. 2019, p. 126 – 138. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/download/37986/pdf/187516. Acesso em: 30 dez. 2020.

RIGGS, F.W.: Onomantics and Terminology: Formats, borrowed terms and omissions. **Knowl.Org**. v.23, l996, p. 216 -224. Disponível em: https://www.nomos-elibrary.de/10.5771/0943-7444-1996-4-216.pdf?download_full_pdf=1. Acesso em: 15 jun. 2021.

RILOFF, Ellen; LORENZEN, Jeffrey. Extraction-based text categorization: generating domain-specific role relationships automatically. In_____: STRZALKOWSKI, Tomek (Ed). **Natural Language Information Retrieval.** [s.l.]: Springer Science, 1999.

RUGE, Gerda. **Combining corpus linguistics and human memory models for automatic term association**. In_____: STRZALKOWSKI, Tomek (Ed). Natural Language Information Retrieval. [s.l.]: Springer Science, 1999.

RYAN, C. **Thesaurus construction guidelines**: an introduction to thesauri and guidelines on their construction. Publisher: National Library of Ireland, 2014. Disponível em: https://www.researchgate.net/publication/267207310_Thesaurus_construction_guidelines_an_introduction_to_thesauri_and_guidelines_on_their_construction. Acesso em: 15 mai. 2020.

SÁNCHEZ-CUADRADO, S., COLMENERO-RUIZ, M. J.; MOREIRO, J. A. Tesauros: estándares y recomendaciones. **El profesional de la información**, mai / jun, v. 21, n. 3, 2012.

SANTAELLA, L. **O método anticartesiano de C. S Peirce**. São Paulo: Editora UNESP, 2004.

SELLTIZ, C.; WRIGHTSMAN, L. S.; COOK, S. W. **Métodos de pesquisa das relações sociais**. São Paulo: Herder, 1965.

SILVEIRA, D. T.; CÓRDOVA, F. P. **A pesquisa científica**. In_____: GERHARDT, Tatiana Engel; SILVEIRA, Denise Tolfo. Métodos de pesquisa. Porto Alegre: Editora da UFRGS, 2009.

SMIT, J. W.; KOBASHI, N. Y. **Como elaborar vocabulário controlado para aplicação em arquivos.** São paulo: Arquivo do Estado, Imprensa Oficial, 2003. 56 p.

SMITH, B.; WELTY, C. **Ontology: Towards a new synthesis**: in Formal Ontology in Information Systems, Maine: ACM Press. 2001 Disponível em: https://portal.acm.org/citation.cfm?doid=505168.505201. Acesso em: 05 jun. 2020.

SOERGEL, D. **Thesauri and Ontologies in Digital Libraries**: Tutorial. In: Evaluation ofthesauri. Joint Conference on Digital Libraries, Portland, Oregon, July 14, 2002. Disponível em: http://www.dsoergel.com/cv/B63.pdf. Acesso em: 15 mai. 2020.

SOLER-MONREAL, C; GIL-LEIVA, I. Evaluation of controlled vocabularies by inter-indexer consistency. **Information Research**: An International Electronic Journal, v. 16, n. 4, 2011. Disponível em: http://www.informationr.net/ir/16-4/paper502.html. Acesso em: 10 jan. 2020.

SOUSA, Janailton Lopes. **Avaliação do padrão Simple Knowledge Organization System (SKOS) para a representação de vocabulários controlados**. 2019. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de São Carlos, São Carlos, 2019. Disponível em: https://repositorio.ufscar.br/handle/ufscar/11934. Acesso em: 05 out. 2020.

SOUZA, R. R.; TUDHOPE, D.; ALMEIDA, M. Towards a Taxonomy of KOS: Dimensions for Classifying Knowledge Organization Systems. **Knowledge Organization**. 39, 2012, p179-192. Disponível em: https://doi.org/10.5771/0943-7444-2012-3-179. Acesso em: 05 out. 2020.

SOUZA, T. B.; CATARINO, M. E.; SANTOS, P. C. D. Metadados: catalogando dados na internet. **Transinformação**, v. 9, n. 2, 1997. Disponível em: http://hdl.handle.net/20.500.11959/brapci/23620. Acesso em: 10 mar. 2023

STRZALKOWSKI, Tomek (Ed). **Natural Language Information Retrieval.** [s.l.]: Springer Science, 1999. Disponívem em: https://link.springer.com/content/pdf/10.1007/978-94-017-2388-6.pdf. Acesso em: 10 out. 2021.

SUOMINEN, O.; HYVÖNEN, E. Improving the quality of skos vocabularies with skosify. In: International Conference on Knowledge Engineering and Knowledge Management, 18th , 2012, Ireland. **Proceedings**… Ireland, Springer-Verlag, 2012.

SUOMINEN, O; MADER, C. Assessing and Improving the Quality of SKOS Vocabularies. **Journal on Data Semantics**, v. 3 n. 1, p. 47-73, 2014.

SVENONIUS, E. Unanswered questions in the design of controlled vocabularies. J. Am. KOS. **Inf. Sci.,** v.37, p.331-340, 1986. Disponível em: https://doi.org/10.1002/(SICI)1097-4571(198609)37:5<331::AID-ASI8>3.0.CO;2-E. Acesso em: 05 mai. 2020.

TÁLAMO, M. de F. G. M.; LARA, M. L. G. de; KOBASHI, N. Y. Contribuição da terminologia para a elaboração de tesauros. **Ci. Inf, Brasília**, 21(3): 197-200, set./dez. 1992.

TROJAR, M. Wüster's View of Terminology. Slovenski jezik – **Slovene Linguistic Studies,** 11 : p. 55–85. Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, Ljubljana, 2017.

URDICIAIN, B. G. Evaluación del rendimiento de tesauros españoles en sistemas de recuperación de información. **Revista Española de Documentación Científica**, v. 21, n. 3, 1998.

VÁLLEZ, M.; PEDRAZA-JIMÉNEZ, R.; CODINA, L.; BLANCO, S.; ROVIRA, C. Updating controlled vocabularies by analysing query logs, **Online Information Review**, v.

39, n. 7, p. 870-884, 2015. Disponível em: https://doi-org.ez31.periodicos.capes.gov.br/10.1108/OIR-06-2015-0180. Acesso em: 15 mai. 2020.

VERE, Aubrey de. Catholic world. **Making of America Journal Articles** [Volume 18, Issue 104, Nov 1873; pp. 194-195]. Disponível em: http://quod.lib.umich.edu/m/moajrnl/bac8387.0018.104/198 . Acesso em: 15 mai. 2021.

VOGEL, Michely Jabala Mamede; KOBASHI, Nair Yumiko. Tesauro funcional para organização de arquivos administrativos. **Páginas a&b**. S.3, nº 12 (2019) 48-62. Disponível em: https://doi.org/10.21747/21836671/pag12a4. Acesso em: 12 dez. 2021.

WITTGENSTEIN, Ludwig. **Tractatus Logico-Philosophicus**. Tradução e apresentação de José Arthur Giannotti. São Paulo: Companhia Editora Nacional, 1968.

WRIGHT, S. E.**Typology for Knowledge Representation Resources**. in NKOS-CENDI September, 2008.

ZAHAREE, M. Building controlled vocabularies for metadata harmonization. Bul. Am. KOS. **Info. Sci. Tech**., v.39, p.39-42, 2012. Disponível em: https://doi.org/10.1002/bult.2013.1720390211. Acesso em: 05 mai. 2020.

ZAMORANO AGUILAR, Alfonso. La terminología como disciplina: aproximación interpretativa a su evolución epistemológica y metodológica a través de la caología. **Moenia,** n.19, 2013, p. 25-43.

ZENG, M. L. Knowledge organization systems (KOS). **Knowledge Organization:** International journal devoted to concept theory, classification, indexing, and knowledge representation, Frankfurt, v. 35, n. 2-3, p. 160-182, 2008.

ZHANG, Y., OGLETREE, A., GREENBERG, J.; ROWELL, C. Controlled vocabularies for scientific data: Users and desired functionalities. Proc. AsKOS. **Info. Sci. Tech**., 52: 1-8, 2015. Disponível em: https://doi.org/10.1002/pra2.2015.145052010054. Acesso em: 05 mai. 2020.

ZHOU, Joe. **Phrasal terms in real-world ir applications**. In_____: STRZALKOWSKI, Tomek (Ed). Natural Language Information Retrieval**.** [s.l.]: Springer Science, 1999.

APÊNDICE A – Lista de vocabulários controlados recuperados

Vocabulários Controlados em RDF/SKOS

1 Thesaurus Multilingue da União Europeia (EuroVoc) (pt) (download)
2 Tesauro de Objetos Religiosos da Fé Católica (pt) Sparql
3 Global Agricultural Concept Scheme (GACS) (não disponível para download)
4 International Coastal Atlas Network Coastal Erosion Global Thesaurus (2Download ou Sparql)
5 Tesauro Ambiental Multilíngue Geral (GEMET) (pt) (download)
6 Resource Type Vocabulary - (download) ok
7 Electronic Government Controlled Vocabulary (br) (pdf, csv, rdf)
8 LandVoc - the Linked Land Governance Thesaurus (rdf, csv)
9 Thesaurus Unesp
10 CAB Thesaurus (CABT) (parcialmente disponível)
11 European Education Thesaurus (parcialmente disponível em, tematres)
12 Multilingual European Thesaurus on Health Promotion (parcialmente disponível, tematres)
13 AGROVOC (Download ou Sparql)
14 Thesaurus in Labor Law (br) (parcialmente disponível em, tematres)
15 National Agricultural Thesaurus (br) (parcialmente disponível, tematres, mudou de endereço)
16 travel!digital Thesaurus
17 Europeana Fashion Vocabulary ok (versão em RDF compactada)
18 Partage Plus Vocabulary versão em RDF dividida em 3 partes ok

Vocabulários Controlados em PDF

1 Tesauro Brasileiro de Ciência da Informação (IBICT) (**br**) (link mudou (sem acesso))
2 Thesaurus for Gender Studies and Women (**br**) (interface de busca e pdf)
3 LanguaL ™ Thesaurus (interface de busca e pdf)
https://www.langual.org/langual_Thesaurus.asp
4 European Commission for Democracy Through Law Systematic Thesaurus (pdf e csv)
5 Human Rights Information and Documentation Systems Micro-Thesauri
6 Electoral Justice Thesaurus (br)
7 European multilingual thesaurus on AIDS and HIV infection (desatualizado)

Vocabulários Controlados em outros formatos

1 REEEP Climate Smart Thesaurus ok

2 Military Thesaurus of the Union (br) (erros ao exportar em skos)
3 Thesaurus of Clinical Signs (Integrado ao sistema de busca Orphanet) (OWL)
4 Thesaurus Brased (br)"The Brazilian Thesaurus of Education (Brased, integrado ao pergamum)
5 Inter-Active Terminology for Europe (csv)
6 Legal Vocabulary (br, stj) sistema próprio
7 Thesaurus of Folklore and Brasilian Popular Culture (br) Multites
8 Thesaurus of Senado Federal (br)
9 Multilingual Egyptian Thesaurus

Vocabulários Controlados Inativos, Descontinuados Ou Não Localizados

1 INA Web Thesaurus
2 European Migration Network Glossary & Thesaurus
3 Thesaurus of Scientific Instruments in Portuguese
4 Thesaurus of European Education Systems
5 Thesaurus of Disability
6 Multilingual synoptic thesaurus of affections of the soft tissues of the oral cavity (TEMATRES antigo, desatualizado) link quebrado
7 European Heritage Network Multilingual Thesaurus (pdf, migrou de endereço)
8 Thesaurus in Information Science (br) (PUC-MG)
9 Historical Archive of São Paulo Archival Description (br)(TEMATRES antigo, desatualizado)
10 Thesaurus of Biodiversity (br) (TEMATRES 1,5 antigo, desatualizado)
11 Bibliography of Contemporary History of Portugal Thesaurus
12 Children and Youth Literature sem acesso () link quebrado

APÊNDICE B – Dados da pesquisa

1 Link dos Dados

2 Plano de Dados

ANEXOS

THE Subscriptions to the Association for the year 1871 become due in advance on January 1st. All which are not already paid, should be forwarded to the General Secretary, Mr. T. Watkin Williams, 13, New Hall Street, Birmingham; or to the Secretaries of Branches.

BRITISH MEDICAL JOURNAL.

SATURDAY, APRIL 8TH, 1871.

A LAME APOLOGY.

IT was the very unpleasing duty of the JOURNAL a fortnight since, in defending the British Medical Association and its Medical Reform Committee from a long series of insults and attacks on the part of the *Lancet*, to have to declare and to prove that every one of the statements was directly contrary to the truth in essence, in form, and in letter, and to denounce them separately and in plain language as untrue, and as deliberately put forward to pervert and misrepresent the action of the Association and to injure and annoy the well-known public men who have been making great sacrifices of time and labour, at the bidding of the Association, in order to assist in securing certain defined objects of medical reform entrusted to their charge by repeated public and solemn acts of the Association in general meeting. Our contemporary pleads guilty on all the charges of fact, but with the unlooked-for qualification that its statements were of no importance. It was not true that the Committee was an excrescence of the Committee of Council; that it had no representative character; or was a "Rump"; or that it had lost all its importance by the secession of whatever members gave importance to it; or that it had suffered any secession at all; or that Mr. Headlam had intimated a preference for the clauses of the Bill brought forward by Mr. Lush; or that our statements on the subject were incredible, or anything else than strictly accurate; or that the five members of the General Medical Council had done anything more than to decline to act with the General Executive of the Association in urging the principle of direct representation, which is the war-cry of the *Lancet*, borrowed from the Association. But whether any or all of these statements were true, whether the five gentlemen referred to belonged to the Reform Committee or not, or why they took this course, is of no more consequence "than whether they wore light coats or dark ones." Every statement which the *Lancet* has made on this subject, and on which it has founded so much elegant abuse, is false from beginning to end. Compelled this week to start from that confession as a base, it has no other plea to offer than that its statements are "of no importance." It is very well to know the estimate which it puts upon its own fabrications; many persons were disposed beforehand to adopt it: on such excellent authority it will henceforth be difficult to doubt it, on this or any other subject. We presume, however, that our contemporary does not always mean to carry its rage for unimportant fiction and trifling vituperation to the extent of falsifying the acts of a great association, garbling and suppressing the brief official rectifications of fact addressed to it, and coarsely insulting the eminent representatives of a body including a fourth of the profession. At least, the plea of "*vive la bagatelle*" comes in a little lamely as an apology for such a course; it is contrary to the rules of professional conduct and of respectable journalism. By assuming such an attitude, it may hope to avert the destructive anger, but it will hardly save itself from the unmitigated contempt, of the profession.

It opens a new issue by professing for the Association an affection, which we are most willing to accept as real and unforced. In that case we will point out to the conductors of the journal, that some treasonable practices have been at work to falsify that affection. Setting aside the question of medical reform, which it desires to select as a field for agreeable fiction and irresponsible misstatement, we may point out that there is no reason why a journal which is merely just, not to say affectionate, should use the name of the Association so freely for the purposes of abuse, however playful, but carefully erase its name from any paragraph referring to its acts of public usefulness, when the necessity of referring to those acts occurs. For instance, there is no reason which we can discover, founded either in love or justice, why the recent meeting of the Metropolitan Counties Branch, attended by an influential parliamentary and lay auditory, to consider the medical aspects of pauperism, should have been reported as "a meeting at the Charing Cross Hotel"; or why all reference to the acts of the Association has been carefully omitted in the articles on the origin and report of the Royal Sanitary Commission, in the summary of the year, in the articles on the sick-club movement, the Shropshire tariff of fees, and a score of other similar matters. It is impossible to dissimulate affection more closely than by seizing every conceivable opportunity for abusing the beloved object, and suppressing its connexion with any efforts which must be discussed in the opposite spirit. With this hint, we close the discussion.

The sort of apology which the *Lancet* offers is not much more creditable than the offence. There are, however, in its management elements which justify the anticipation that similar proclivities will in future be checked by the good sense and good feeling of gentlemen connected with that journal. The *Lancet* must have printed the official letter of our General Secretary rectifying some of its misstatements, with something of the feeling with which a Parisian tradesman affixes to his doorpost the judgment of the court on his fabricated compounds, of which the public analyst has exposed the character; but at least it has thought well, since our "rectification of facts," to pause in the course of suppressing or mutilating the official corrections which its misstatements compelled.

PHARMACOPŒIAL NOMENCLATURE.

IN an able paper read before the Pharmaceutical Society on Wednesday evening, Dr. Attfield discussed a subject of much medical interest, the alteration in pharmacopœial nomenclature necessitated by the advancement of chemistry. Within the last few years, the views hitherto prevailing of the constitution of matter have undergone radical alteration. There is no small difficulty in adopting for the Pharmacopœia chemical names, explicit, easily understood, and unambiguous, and yet consistent with accepted chemical theories taught in the schools. Dr. Attfield discussed the history of the chemical names employed in the Pharmacopœia, historically, and from the modern stand-point. We need not follow him through this part of his address, in which the facts will have been anticipated by many of our readers, but may refer to the current number of the *Pharmaceutical Journal*, in which it will appear at length, but will only state the conclusions at which he arrives.

The chief alterations in pharmacopœial nomenclature which he proposed amounts to this, that the compounds of the alkali-metals and alkaline-earth-metals, instead of being named as hitherto on two distinct systems, should follow but one:—that instead of salts of potassium and of potash, we should have salts of potassium only; instead of

April 8, 1871.] THE BRITISH MEDICAL JOURNAL. 373

THE Subscriptions to the Association for the year 1871 become due in advance on January 1st. All which are not already paid, should be forwarded to the General Secretary, Mr. T. Watkin Williams, 13, New Hall Street, Birmingham; or to the Secretaries of Branches.

BRITISH MEDICAL JOURNAL.

SATURDAY, APRIL 8TH, 1871.

It opens a new issue by professing for the Association an affection, which we are most willing to accept as real and unforced. In that case we will point out to the conductors of the journal, that some treasonable practices have been at work to falsify that affection. Setting aside the question of medical reform, which it desires to select as a field for agreeable fiction and irresponsible misstatement, we may point out that there is no reason why a journal which is merely just, not to say affectionate, should use the name of the Association so freely

sequence "than whether they wore light coats or dark ones." Every statement which the *Lancet* has made on this subject, and on which it has founded so much elegant abuse, is false from beginning to end. Compelled this week to start from that confession as a base, it has no other plea to offer than that its statements are "of no importance." It is very well to know the estimate which it puts upon its own fabrications; many persons were disposed beforehand to adopt it: on such excellent authority it will henceforth be difficult to doubt it, on this or any other subject. We presume, however, that our contemporary does not always mean to carry its rage for unimportant fiction and trifling vituperation to the extent of falsifying the acts of a great association, garbling and suppressing the brief official rectifications of fact addressed to it, and coarsely insulting the eminent representatives of a body including a fourth of the profession. At least, the plea of "*vive la bagatelle*" comes in a little lamely as an apology for such a course; it is contrary to the rules of professional conduct and of respectable journalism. By assuming such an attitude, it may hope to avert the destructive anger, but it will hardly save itself from the unmitigated contempt, of the profession.

PHARMACOPŒIAL NOMENCLATURE.

In an able paper read before the Pharmaceutical Society on Wednesday evening, Dr. Attfield discussed a subject of much medical interest, the alteration in pharmacopœial nomenclature necessitated by the advancement of chemistry. Within the last few years, the views hitherto prevailing of the constitution of matter have undergone radical alteration. There is no small difficulty in adopting for the Pharmacopœia chemical names, explicit, easily understood, and unambiguous, and yet consistent with accepted chemical theories taught in the schools. Dr. Attfield discussed the history of the chemical names employed in the Pharmacopœia, historically, and from the modern stand-point. We need not follow him through this part of his address, in which the facts will have been anticipated by many of our readers, but may refer to the current number of the *Pharmaceutical Journal*, in which it will appear at length, but will only state the conclusions at which he arrives.

The chief alterations in pharmacopœial nomenclature which he proposed amounts to this, that the compounds of the alkali-metals and alkaline-earth-metals, instead of being named as hitherto on two distinct systems, should follow but one:—that instead of salts of potassium and of potash, we should have salts of potassium only; instead of

85th and 105th, and finally that between the 115th and 140th degrees. But the intensity of colour must vary inversely to the breadth of the stripes, and the three stripes left between the red ones be filled with a pretty vivid blue. This hemisphere placed upon a table with its southern pole pointing towards sunset will afford a tolerable portrait of the aspect of the sky as it appeared immediately after sunset, and continued unchanged for more than a quarter of an hour. The stripes were not visible near the horizon, but were very distinct at an altitude of about fifteen degrees, and almost disappeared about the zenith. No cloud was seen during the occurrence of the phenomenon.

These stripes were certainly parallel in reality, and their apparent divergence may be accounted for by perspective. The reddish stripes may owe their colour to sunlight reflected back from the particles scattered in the atmosphere. But why did the celestial vault show so distinct a blue colour in the intervening bands? Yet, probably, this phenomenon is more easily to be explained than the infinite variation of evening colourings that want a valid explanation to this day.

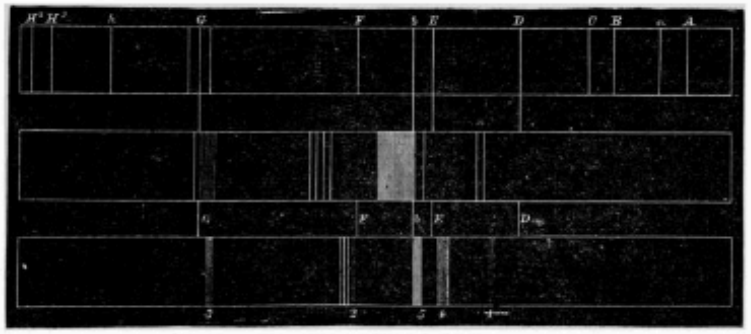
Magdeburg, August 19 A. SPRUNG

THE AURORA

THERE was a very fine display of aurora here on the night of the 21st. It commenced to be visible about 9.30 P.M., reached its maximum about 11, and faded suddenly away about 11.30. In appearance it was of a silvery white, without a trace of that rose colour which characterised the three great displays of last autumn. The main portion of the light was in the north-western quarter of the heavens, and it was sufficiently strong to see large print by. Extending from the north-west and reaching the north-eastern horizon arose three luminous arches concentric with each other, the 1st about 15° altitude, the 2nd about 25° altitude, and the 3rd about 40° altitude. These were connected by radial tongues of light which were ever changing their height. There was another marked and isolated nucleus about and around α Lyræ.

At about 10.45 P.M. there were most curious rays shot up from the arches in the north, and concentric with them. These shooting arches, if I may call them so, had at the horizon an apparent angle of about 150° to 180°, but as they approached the nucleus in Lyra, they contracted and lost themselves in sheets of white light. On applying the spectroscope I found one bright line visible all over the heavens excepting on the south horizon for an altitude of about 25°. The spectrum obtained on the north-west gave five bright lines, of which I send a drawing.

From want of convenient measuring apparatus I had resource to the method of superposed spectra. The light I chose for comparison was that of a tallow candle, from which I got the bright lines of sodium and carburetted



COMPARISON SPECTRA OF SUN, AURORA, AND CANDLE

hydrogen. The instrument I used was one of Browning's direct vision spectroscopes—an instrument that gives the best results with the minimum amount of light. Of the bright lines, two were strong, one was medium, two were very faint. In the accompanying map I have put the solar spectrum at the top and carried the chief lines down for comparison. In putting numbers to the lines I have been directed by their degrees of intensity.

No. 1 is a sharp, well-formed line, visible with a very narrow slit.

No. 2. A line very slightly more refrangible than F. The side towards D is sharp and well-defined, while on the other side it is nebulous.

No. 3. Slightly less refrangible than G, is a broad ill-defined band only seen with a wide slit.

No. 4. A line near E, woolly at the edges, but rather sharp in the centre. This should be at or near the position of the line 1474 of the solar corona.

No. 5. A faint band coincident with *b*, extending equally on both sides of it.

The barometer stood at 29·574 in.; the thermometer at 61°·2. A gentle wind was blowing from the south-west,

*FRUIT CLASSIFICATION**

DR. DICKSON referred to the confessedly unsatisfactory state of fruit-classification, and to the very unnecessary extent of the existing terminology, which is further complicated by a considerable amount of variance among botanists as to the precise application of several of the terms employed. He was of the opinion, which he believed to be a growing one among botanists, that the most convenient method of classification was, in the first place, rigorously to restrict the definition of a "fruit" to the mature or ripe pistil, excluding from that definition the modifications of accessory parts or organs, which, in many cases, are correlative therewith; and, secondly, to base the primary classification upon the general character of the modification undergone by the pars of the pistil in ripening, treating as of minor importance the characters involved in the description of the flower, such as the superior or inferior position of the ovary, &c.

The classification which Dr. Dickson suggests for the consideration of botanists approaches most nearly to that indicated by Schacht in his "Grundriss," of which, indeed, it may be viewed as a modification and expansion. Schacht grouped fruits under three heads—(1) Capsular fruits which dehisce to allow the seeds to escape; (2) splitting fruits or Schizocarps, which

hydrogen. The instrument I used was one of Browning's direct vision spectroscopes—an instrument that gives the best results with the minimum amount of light. Of the bright lines, two were strong, one was medium, two were very faint. In the accompanying map I have put the solar spectrum at the top and carried the chief lines down for comparison. In putting numbers to the lines I have been directed by their degrees of intensity.

No. 1 is a sharp, well-formed line, visible with a very narrow slit.

No. 2. A line very slightly more refrangible than F. The side towards D is sharp and well-defined, while on the other side it is nebulous.

No. 3. Slightly less refrangible than G, is a broad ill-defined band only seen with a wide slit.

No. 4. A line near E, woolly at the edges, but rather sharp in the centre. This should be at or near the position of the line 1474 of the solar corona.

No. 5. A faint band coincident with *b*, extending equally on both sides of it.

The barometer stood at 29·574 in.; the thermometer at 61°·3. A gentle wind was blowing from the south-west, and the sky was free from clouds.

Observatory, Dun Echt, Aberdeen LINDSAY

FRUIT CLASSIFICATION *

DR. DICKSON referred to the confessedly unsatisfactory state of fruit-classification, and to the very unnecessary extent of the existing terminology, which is further complicated by a considerable amount of variance among botanists as to the precise application of several of the terms employed. He was of the opinion, which he believed to be a growing one among botanists, that the most convenient method of classification was, in the first place, rigorously to restrict the definition of a "fruit" to the mature or ripe pistil, excluding from that definition the modifications of accessory parts or organs, which, in many cases, are correlative therewith; and, secondly, to base the primary classification upon the general character of the modification undergone by the parts of the pistil in ripening, treating as of minor importance the characters involved in the description of the flower, such as the superior or inferior position of the ovary, &c.

The classification which Dr. Dickson suggests for the consideration of botanists approaches most nearly to that indicated by Schacht in his "Grundriss," of which, indeed, it may be viewed as a modification and expansion. Schacht grouped fruits under three heads— (1) Capsular fruits which dehisce to allow the seeds to escape; (2) splitting fruits or Schizocarps, which

* "Suggestions on Fruit Classification." By Alex. Dickson, M.D., Regius Professor of Botany in the University of Glasgow. Read before the British Association, 1871.

PHILOSOPHICAL TERMINOLOGY.*

II.

TO THE EDITOR OF THE CATHOLIC WORLD:

IN the letter which I ventured to address to you a short time ago concerning the general conditions required in a good English work of philosophy, I made some observations on the importance and difficulty of wielding the popular language in a strictly philosophical manner. As I apprehend that the title of "Philosophical Terminology," under which that letter was made to appear, is scarcely justified by its very limited contents, I beg leave to add a few other considerations on the same subject, that your intelligent readers may find in these additional remarks a confirmation and a further development of what I said about our need of a more copious philosophical language.

There are two words which can-is popularly used in this philosophical sense.

The word "act" with us primarily signifies that which is produced by action; for all action is the production, or the position, or the making of an act. But all action implies an agent—that is, a being which is already "in act," with its actual power prepared for action. On the other hand, nothing is formally "in act," but through an intrinsic "act," which is the formal principle of its actuality. Accordingly, the word "act," though primarily known to us as expressing the product of action, must, by metaphysical necessity, be applied also to that from which every agent and every being has its actuality.

Hence, philosophers found it necessary to admit two kinds of "acts"—the *essential* and the *accidental*. The essential is that which gives

DR. WILLIAM POLE, F.R.S., IN THE CHAIR.

ON MUSICAL NOMENCLATURE.

BY JOHN HULLAH, ESQ.

I PROPOSE in this address to deal with certain names or terms and epithets in use among English musicians. Many of these, it is certain, have outlived the ideas or things for which they once stood; others now represent to all of us ideas and things different from those they once represented. The time seems to have arrived when we should come to an understanding as to our musical nomenclature. It will not, I think, be found necessary to make any addition to it; at any rate, I have none to propose to you to-day. But I shall simply ask you to consider and, if possible, to decide, which out of many names or terms representing, and epithets qualifying, the same thing, it is desirable to adopt or recommend for adoption. Musical nomenclature has reference, of necessity, to time, to tune, and to expression. I will deal with its application to these separately. Under the head of time, let us first consider the duration names of musical notes. Those which at present concern us are—breve, semibreve, minim, crotchet, quaver, semiquaver, and demisemiquaver. Of these names, the first three have altogether lost their significance; the fourth is no longer appropriate; the fifth, sixth, and seventh are arbitrary. The breve is no longer short, but unusually long; the minim is not now the least or shortest note, but not unfrequently the greatest or longest; the crotchet has now no crotch or hook; and the quaver and its fractions might just as well be called the 'shiver,' the 'half-shiver,' and the 'quarter-shiver,' or by any other names as fantastic or irrelative. The Germans call these notes, beginning from our semibreve, the whole note, the half note, the quarter note, and so on. These appellations, so far as they express the proportion of the first note named to those which follow it, are convenient. They form themselves a time-table, but it is an imperfect one, for they do not show, without further calculation, any intermediate proportions. They show at once that eight quavers equal one semibreve, but not at once that four quavers equal one minim. But I have a much more serious charge to bring against them. They assume what, if not always false, is, as it seems to me, not

For the Library.

From the Institute.—The Canadian Journal. Vol. XV, No. 8.
From the Institution.—Journal of the Royal Institution of Cornwall, No. XX.
From the Society.—Journal and Proceedings of the Royal Society of New South Wales. Vol. XI.
From the Author.—Remarks on the Sedimentary Formation of New South Wales, 4th edition. By the Rev. W. B. Clarke, M.A., F.R.S.
From the Berlin Anthropological Society.—Zeitschrift für Ethnologie, No. 4, 1878.
From Sir John Lubbock, Bart., M.P., D.C.L., F.R.S.—Plan of the Maori Mythology. By John White.
From the Society.—Transactions of the Asiatic Society of Japan. Vol. VI, Part 2.
From J. Park Harrison, Esq., M.A.—Essai de Classification des bruits Articules. By M. Condeseau; Du Transformisme: Note sur le Squelette de Mention. By M. Le Marquis de Nadaillac; Du Mouvement de la Population en France et en Europe. By Miss de Nadaillac; Life in the Southern Isles. By the Rev. W. Wyatt Gill.
From the Right Hon. Lord Arthur Russell, M.P.—Parliamentary Papers relating to the Colony of Fiji, Slave Trade, China, Mauritius, and South Sea Islanders.
From the Editor.—Revue Internationale des Sciences, Nos. 46–47, 1878.
From the Editor.—Revue Scientifique, Nos. 20–21, 1878.

The following paper was read by the author:—

On the Evils *arising from the* Use *of* Historical National Names *as* Scientific Terms. By A. L. Lewis, M.A.I.

Perhaps there are few things that strike the Anthropologist more in reading the periodical literature of the time than the feeling shown by the writers that race problems have much to do with the every-day business of the world, together with the very imperfect understanding of the complex nature of those problems which is usually displayed, and the consequent ease with which they are sometimes settled on paper on the one hand, and the exaggerated importance and influence frequently attached to them on the other.

Thus, the French are usually written of as Gauls, and all their