LIDIANNE VIANNA ALBERNAZ

Uma revisão integrativa sobre taxonomias e *web* semântica: trajetórias, convergências e interdisciplinaridade

Dissertação de mestrado
Dezembro de 2023

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO (UFRJ)
ESCOLA DE COMUNICAÇÃO (ECO)
INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA (IBICT)
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO (PPGCI)

LIDIANNE VIANNA ALBERNAZ

UMA REVISÃO INTEGRATIVA SOBRE TAXONOMIAS E WEB SEMÂNTICA: TRAJETÓRIAS, CONVERGÊNCIAS E INTERDISCIPLINARIDADE

Rio de Janeiro
2023

LIDIANNE VIANNA ALBERNAZ

UMA REVISÃO INTEGRATIVA SOBRE TAXONOMIAS E WEB SEMÂNTICA: TRAJETÓRIAS, CONVERGÊNCIAS E INTERDISCIPLINARIDADE

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação em convênio entre o Instituto Brasileiro de Informação em Ciência e Tecnologia e a Universidade Federal do Rio de Janeiro/Escola de Comunicação, como requisito parcial à obtenção do título de Mestre em Ciência da Informação.

Orientadora: Prof.ª Dr.ª Lillian Maria Araujo de Rezende Alvares

Rio de Janeiro

2023

CIP - Catalogação na Publicação

A331r

Albernaz, Lidianne Vianna
Uma revisão integrativa sobre taxonomias e web semântica: trajetórias, convergências e interdisciplinaridade / Lidianne Vianna Albernaz. - Rio de Janeiro, 2023.
243 f.

Orientador: Lillian Maria Araujo de Rezende Alvares.
Dissertação (mestrado) - Universidade Federal do Rio de Janeiro, Escola da Comunicação, Instituto Brasileiro de Informação em Ciência e Tecnologia, Programa de Pós-Graduação em Ciência da Informação, 2023.

1. Sistemas de Organização do Conhecimento. 2. Taxonomia. 3. Web Semântica. 4. Recuperação da Informação. 5. Interdisciplinaridade. I. Alvares, Lillian Maria Araujo de Rezende, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

LIDIANNE VIANNA ALBERNAZ

UMA REVISÃO INTEGRATIVA SOBRE
TAXONOMIAS E WEB SEMÂNTICA: TRAJETÓRIAS, CONVERGÊNCIAS E INTERDISCIPLINARIDADE

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação em convênio entre o Instituto Brasileiro de Informação em Ciência e Tecnologia e a Universidade Federal do Rio de Janeiro/Escola de Comunicação, como requisito parcial à obtenção do título de Mestre em Ciência da Informação.

Aprovada em 22 de dezembro de 2023.

Prof.ª Dr.ª Lillian Maria Araujo de Rezende Alvares
Orientadora, PPGCI IBICT/UFRJ

Prof.ª Dr.ª Dulce Maria Baptista
Membro Titular Externo, UnB

Prof.ª Dr.ª Rosali Fernandez de Souza
Membro Titular Interno, PPGCI IBICT/UFRJ

Prof.ª Dr.ª Vania Lisboa da Silveira Guedes
Membro Titular Interno, PPGCI IBICT/UFRJ

Às minhas avós, que iluminam cada passo do meu caminhar.

Às minhas três mães, mulheres fortes com um tipo especial de bondade e generosidade, cujas histórias de vida e ensinamentos me inspiram.

AGRADECIMENTOS

Sinceros agradecimentos aos professores, discentes, secretaria acadêmica e colaboradores do Programa de Pós-graduação em Ciência da Informação (PPGCI IBICT-UFRJ), assim como ao Instituto Brasileiro de Informação em Ciência e Tecnologia e à Universidade Federal do Rio de Janeiro e sua Escola de Comunicação.

Aos professores Gustavo Saldanha, Rosali Fernandez de Souza e Luana Sales, que me apresentaram à área de organização do conhecimento e plantaram as primeiras sementes do interesse pelos sistemas de organização do conhecimento. Esta última de forma especial por me acolher nos momentos mais difíceis e a quem atribuo a ideia do tema.

Ao professor Jorge Biolchini, que me apresentou ao mundo das revisões sistemáticas.

À minha orientadora, por quem guardo admiração profunda, pavimentou cada passo do meu caminho e contribuiu significativamente para o desenvolvimento de meus conhecimentos no campo da ciência da informação.

Ao meu amigo e professor Daniel Campos, que é exemplo de pessoa e de profissional, uma inspiração de vida. Esteve comigo quando me formei em gestão de projetos e esteve comigo quando desenvolvi o pré-projeto de pesquisa para ingresso no programa de mestrado.

Também aos professores Dércio Santiago, César Vargas, Liliane Penello, Liliana Lugarinho, Corina Mendes, Suely Deslandes, Romeu Gomes, Martha Moreira, Cynthia Magluta e Maria Gomes, que contribuem de forma especial com minha trajetória de vida.

À Izadora Schetter, que foi meu apoio emocional durante toda a jornada do mestrado, quem me deu confiança para seguir em frente.

Ao Marcio Souza, meu amor, meu companheiro, meu amigo, que passou comigo pelos momentos mais difíceis e foi capaz das formas mais lindas de apoio e carinho.

À todas as minhas famílias.

“Taxonomy: The perfect mix of art and Science”

“Taxonomia: a mistura perfeita entre arte e ciência” (tradução nossa)

Mollee Marcus, Taxonomista

RESUMO

Do “caos documentário” à “*web* dos dados” (como é conhecida a *web* semântica), deparamo-nos com uma explosão informacional que não vai diminuir de intensidade, especialmente no mundo virtual. Trata-se de uma questão que nos demanda a identificação de boas práticas e estratégias para a organização de informações na *web* e disseminação de conhecimento de forma mais ampla e democrática. Na Ciência da Informação (CI) as taxonomias são sistemas lógicos de classificação e categorização de múltiplas aplicações, utilizados principalmente em ambientes digitais para a organização e navegação de conteúdo, compreensíveis por humanos e máquinas. Nesse contexto, delineou-se pesquisa qualitativa, teórica, descritiva e documental empreendida em duas etapas: revisão narrativa nos domínios da organização e recuperação da informação e do conhecimento e da *web* semântica, e Revisão Integrativa (RI), para atualizar e sistematizar o estado da arte das taxonomias na CI e sua relação com a *web* semântica. A RI utilizou métodos recomendados pelos guias de Toronto e Remington e da *Cochrane Library*. Foram realizadas buscas nas bases Oasisbr e Scopus com os termos “taxonom* AND semantic* AND *web**” em combinação para recuperação de estudos primários publicados a partir de 2019, nos idiomas inglês e português, que abordassem a relação entre taxonomias e *web* semântica na CI e áreas correlatas, excluindo-se estudos da biologia, zoologia e botânica. A estratégia resultou em 237 estudos após duplicatas e o processo de seleção contou com duas etapas, uma de escaneamento que resultou em 67 estudos para elegibilidade e outra de leitura completa, com a seleção final de 55 estudos para análise e síntese temática. Os resultados apontaram funções não usuais para as taxonomias na *web* semântica, que com versatilidade servem atualmente ao enriquecimento semântico de grandes *corporas* multimodais, informação de contexto, gestão de riscos, detecção de conflitos e manejo de eletrodomésticos em casas inteligentes. Novos campos se revelam em diálogo com a CI, como ciência de dados, linguística computacional e inteligência artificial. Um glossário e uma taxonomia bilíngue são produtos da pesquisa, que busca contribuir para a produção científica em língua portuguesa, além do desenvolvimento do tema das taxonomias no Brasil.

Palavras-chave: Sistemas de Organização do Conhecimento. Taxonomia. Web Semântica. Web 3.0. Recuperação da Informação. Interdisciplinaridade.

ABSTRACT

From “documentary chaos” to the “*web* of data” (as websemantics is known), we are faced with an informational explosion that will not decrease in intensity, especially in the virtual world. This is a matter that demands the identification of best practices and strategies for organizing information on the *web* and disseminate knowledge in a broader and more democratic way. In Information Science (IS), taxonomies are logical classification and categorization systems with multiple applications, used mainly in digital environments for organizing and navigating content, understandable by humans and machines. In this context, a qualitative, theoretical, descriptive and documentary research was designed, undertaken in two stages: narrative review in the domains of organization and retrieval of information and knowledge and the semantic *web*, and Integrative Review (IR), to update and systematize the state of the art of taxonomies in IS and their relationship with the semantic *web*. The IR used methods recommended by the Toronto and Remington and the Cochrane Library guides. Searches were carried out in the Oasisbr and Scopus databases with the terms “taxonom* AND semantic* AND *web**” in combination to retrieve primary studies published from 2019 onwards, in English and Portuguese, that addressed the relationship between taxonomies and the semantic *web* in IS and related areas, excluding studies in biology, zoology and botany. The strategy resulted in 237 studies after duplicates and the selection process had two stages, a scanning stage that resulted in 67 studies for eligibility and another complete study reading, with the final selection of 55 studies for analysis and thematic synthesis. The results pointed out unusual functions for taxonomies on the semantic *web*, which with versatility currently serve the semantic enrichment of large multimodal corpus, context information, risk management, conflict detection and management of household appliances in smart homes. New fields are revealed in dialogue with IS, such as data science, computational linguistics and artificial intelligence. A glossary and a bilingual taxonomy are products of the research, which seeks to contribute to scientific production in Portuguese, in addition to the development of the topic of taxonomies in Brazil.

Keywords: Knowledge Organization Systems. Taxonomy. Semantic Web. Web 3.0. Information Retrieval. Interdisciplinarity.

LISTA DE FIGURAS

Figura 1 - Organização e Representação da Informação e do Conhecimento.....42
Figura 2 - Uma visão das estruturas e funções dos SOC 51
Figura 3 - Crescimento estrutural da complexidade entre os tipos de vocabulário controlado.....56
Figura 4 - As várias facetas da taxonomia presente na biblioteconomia e ciência da informação.....64
Figura 5 - Desenho original incluído na proposta de Tim Berners-Lee ao CERN em 1989 74
Figura 6 - As seis etapas do processo de revisão integrativa, Toronto e Remington (2020).....100
Figura 7 - Tela de refinamento da estratégia de busca na base Oasisbr 105
Figura 8 - Visão parcial da tela de consulta na Oasisbr: quantitativo de documentos encontrados.....106
Figura 9 - Visão parcial da tela de consulta na Scopus: quantitativo de documentos encontrados.....108
Figura 10 - Visão dos filtros aplicados na tela de consulta na Scopus.....109
Figura 11 - Nuvem de palavras-chave obtidas pela estratégia de busca adotada na base de dados Scopus.....110
Figura 12 - Fluxograma PRISMA da Revisão Integrativa sobre taxonomias na *web* semântica.....114
Figura 13 - Taxonomia da revisão integrativa sobre taxonomias na *web* semântica (2019-2023).....122
Figura 14 - Camadas de aprendizagem ontológica, traduzida de Yang et al., 2019132
Figura 15 - Funções das taxonomias na arquitetura do Microsoft Concept Graph, adaptado de Ji et al., 2019 145
Figura 16 - Uso de taxonomias na arquitetura do modelo TEAL proposto por Wang et al., 2019 151
Figura 17 - Representação do fluxo de processamento dos principais módulos do sistema NEO, de Giabelli et al., 2020.....154

LISTA DE QUADROS

Quadro 1 - Compreensões do conceito de taxonomia na ciência da informação.....57
Quadro 2 - Comparativo de tipos de estudos de revisão 93
Quadro 3 - Critérios de elegibilidade adotados na revisão integrativa 102
Quadro 4 - Processo de análise temática por etapa da revisão integrativa 120
Quadro 5 - Categorias selecionadas para apresentação das sínteses da RI.....121
Quadro 6 - Localização dos autores dos estudos incluídos na RI no mundo.....126

LISTA DE TABELAS

Tabela 1 - Diferenças entre os tipos de revisão mais comuns 88
Tabela 2 - Estratégia de busca por base de dados 103
Tabela 3 - Total de estudos avaliados por fonte de informação 124
Tabela 4 - Total de estudos incluídos por fonte de informação e idioma 124
Tabela 5 - Total de estudos incluídos na RI por ano e país(es) 125
Tabela 6 - Distribuição conceitual dos estudos incluídos na RI por ano e área identificada 126

LISTA DE ABREVIATURAS E SIGLAS

AIAAIC	Artificial Intelligence, Algorithmic and Automation Incidents And Controversies
AIID	Artificial Intelligence Incident Database
AIRO	Artificial Intelligence risk ontology
AliCG	Alibaba Conceptual Graph
API	Application Programming Interface
ARDC	Australian Research Data Commons
ARK	Acess Risk Knowledge
AS-TEAL	Supervised Taxonomy Enhanced Adversarial Supervised Model
BRAPCI	Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação
CAPEC	Common Attack Pattern Enumeration and Classification
CASP	Critical Appraisal Skills Program
CBF	Collaboratively Boosting Framework
CC	Ciência da Computação
CI	Ciência da Informação
CNN	Convolutional Neural Networks
CSE	Cyber Security Exercises
CWE	Common Weakness Enumeration
DEA	Drug Enforcement Administration
DL	Descrição Lógica
DLM	Deep Language Modeling
DSSN	Deep Semantic Segmentation Network
DUI	Drug-Use Insights
EEO	Ontologia de Equipamentos de Engenharia
ENISA	European Union Agency for Cybersecurity
ESCO	The European Skills, Competences, Qualifications and Occupations
FAIR	Findable, Accessible, Interoperable, Reusable
FTP	File Transfer Protocol

HTML	Hypertext Markup Language
HSS	Hierarchical Semantic Similarity
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
IoT	Internet of Things
ITMWP	Iterative Transfer Multi-Wahba Projection
ISCO	International Standard Classification of Occupations
IP	Internet Protocol
KB	Knowledge Base
KES	Knowledge Explorer System
KG	Knowledge Graph
KGIS	Knowledge Graph Induction Service
KO	Knowledge Organization
LBSN	Location-Based Social Networks
LD	Linked Data
LHA	Lesbian Herstory Archives
LOD	Linked Open Data
MAG	Microsoft Academic Graph
MARC	Machine-Readable Cataloging
MEET	Method for Embeddings Evaluation for Taxonomic Data
MOQA	Metric based Ontology Quality Analyzer
MWP	Multi-Wahba Projection
Oasisbr	Portal Brasileiro de Publicações e Dados Científicos em Acesso Aberto
OBO	Open Biological and Biomedical
OC	Organização do Conhecimento
OccO	Occupational Classification Compliant Ontology
OCDE	Organização para a Cooperação e o Desenvolvimento Econômico
OI	Organização da Informação
OWL	Web Ontology Language
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses

PLN	Processamento de Linguagem Natural
POC	Processos de Organização do Conhecimento
POI	Point Of Interest
QE	Query Expansion
RC	Representação do Conhecimento
RDA	Research Data Alliance
RDA	Research Data Australia
RDF	Resource Description Framework
RI	Representação da Informação
RS	Remote sensing
SEI	Software Engineering Institution
SKOS	Simple Knowledge Organization System
SOC	Sistema de Organização do Conhecimento
SPARQL	SPARQL Protocol and RDF Query Language
SSS	Smart Spreadsheet
SVM	Support Vector Machine
TAM	Technology Acceptance Model
TEAL	Taxonomy Enhanced Adversarial Learning
TCP	Transmission Control Protocol
TIC	Tecnologias de Informação e Comunicação
TMN	Triplet Matching Network
TMWP	Transfer Multi-Wahaba Projection
UFRJ	Universidade Federal do Rio de Janeiro
UK SOC	The UK National Statistics Standard Occupational Classification
UN-GGIM	The United Nations initiative on Global Geospatial Information Management
URI	Uniform Resource Identification
USP	Upwards Score Propagation
U-TEAL	Unsupervised Taxonomy Enhanced Adversarial Learning

WoS	Web Of Science
WWW	World Wide Web
XAI	EXplainable Artificial Intelligence
XBRL	eXtensible Business Reporting Language
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO 21

1.1 PROBLEMA E JUSTIFICATIVA 25
1.2 OBJETIVOS 28
1.2.1 Objetivo Geral 28
1.2.2 Objetivos Específicos 28
1.3 CONTRIBUIÇÕES DA PESQUISA 29
1.4 RESTRIÇÕES 29
1.5 PREMISSA E HIPÓTESES 30

2 REVISÃO DE LITERATURA 31

2.1 ORGANIZAÇÃO E REPRESENTAÇÃO DA INFORMAÇÃO E DO CONHECIMENTO 31
2.1.1 Princípios Classificatórios 43
2.1.1.1 Categorização 44
2.1.1.2 Diferenciação entre termos e conceitos 45
2.1.1.3 Relacionamentos hierárquicos 46
2.1.2 Sistemas de Organização do Conhecimento 48
2.2 TAXONOMIAS 53
2.2.1 Taxonomias e suas relações 68
2.2.1.1 Com tesauros 68
2.2.1.2 Com metadados 69
2.2.1.3 Com ontologias 70
2.2.1.4 Com folksonomias 71
2.3 WEB SEMÂNTICA 71
2.3.1 A web 1.0 73
2.3.2 A web 2.0 75
2.3.3 A web 3.0 79
2.3.4 RDF e SKOS: relações entre web semântica e taxonomias 82

3 METODOLOGIA 86

3.1 DESENHO DA PESQUISA E PROCEDIMENTOS 86
3.2 REVISÃO NARRATIVA 91
3.3 REVISÃO INTEGRATIVA 97

3.3.1 Critérios de elegibilidade 101
3.3.2 Fontes de informação 102
3.3.3 Estratégias de busca 103
3.3.3.1 Base de dados Oasisbr 104
3.3.3.2 Base de dados Scopus 106
3.3.4 Aspectos metodológicos e resultados preliminares dos processos de avaliação de estudos 110
3.3.4.1 Processo de seleção 111
3.3.4.2 Processos de coleta e análise de dados 114
3.3.4.3 Avaliação de riscos de vieses e confiabilidade dos estudos 120
3.3.4.4 Métodos de síntese 121
4 RESULTADOS: REVISÃO INTEGRATIVA 123
4.1 CARACTERÍSTICAS DOS ESTUDOS 123
4.2 SÍNTESES 128
4.2.1 Classificação para contextualizar, compreender e recuperar informações 128
4.2.2 Bases de conhecimento (knowledge bases) e vetores de palavras (embeddings) 141
4.2.3 Grafos do conhecimento (knowledge graphs) 143
4.2.4 Linked (Open) Data 146
4.2.5 Inteligência Artificial 149
4.2.6 Serviços Web e IoT 162
5 DISCUSSÃO E CONCLUSÃO 164
REFERÊNCIAS 171
GLOSSÁRIO 185
APÊNDICE A - RELATÓRIO *CHECKLIST* PRISMA 187
APÊNDICE B - ESTUDOS AVALIADOS PARA ELEGIBILIDADE E QUALIDADE METODOLÓGICA 190
APÊNDICE C - CONTEXTUALIZAÇÃO CONCEITUAL DOS ESTUDOS INCLUÍDOS NA REVISÃO INTEGRATIVA 198

APÊNDICE D - RESULTADOS E APLICABILIDADE DOS ESTUDOS INCLUÍDOS NA REVISÃO INTEGRATIVA 209

APÊNDICE E - ABORDAGENS UTILIZADAS PARA CRIAÇÃO, GESTÃO E/OU USO DE TAXONOMIAS NA *WEB* SEMÂNTICA NOS ÚLTIMOS CINCO ANOS .. 223

ANEXO 1 - CASP QUALITATIVE CHECKLIST (2018) 239

1 INTRODUÇÃO

Do “caos documentário”, marcado por um crescente volume de documentos e desafios para a classificação, à revolução tecnológica, marcada pela interação entre os mundos real e digital, a evolução de dois elementos intelectuais intrinsecamente relacionados desafiam persistentemente nossos modos de ser, de trabalhar e de viver: informação e conhecimento (Souza, 2007). Partindo deste ponto de vista, somos cada vez mais capazes de gerar e disseminar informação (e desinformação) e cada vez menos capazes de organizar e representar tanto informação quanto conhecimento para melhor compreensão e uso efetivo. Há quase uma década, Marjorie Hlava sinalizava para uma explosão informacional que atinge cada dia mais a nossa realidade e não apresenta indícios de que possa desaparecer, diminuir de intensidade ou desacelerar, pelo contrário, se mostra contínua e impõe desafios ao nosso futuro. Isso se torna especialmente evidente quando precisamos entender a informação, refinar a análise e extrair valor real desse volume de informações. A autora nos remete à importância de organizar a informação e o conhecimento, conduzindo-nos numa perspectiva de que a geração de valor agregado a partir da informação é geração de conhecimento (Hlava, 2014).

Na Ciência da Informação (CI) as taxonomias são compreendidas como sistemas lógicos de classificação e categorização, que possuem múltiplas aplicações e são utilizados principalmente em ambientes digitais para a organização e navegação de conteúdo, sendo compreensíveis tanto por humanos quanto por máquinas. Nesse sentido, acredita-se que as taxonomias possam contribuir significativamente para melhor compreensão das informações no ambiente *web*, entre outras coisas, por meio da ampliação da capacidade semântica de seus objetos digitais, tornando a própria rede mais efetiva. Assim, a presente pesquisa, com aportes teóricos de diferentes campos do saber, em especial da ciência da computação e da administração, busca compreender o papel desempenhado pelas taxonomias no contexto da *web* semântica, a *web* que vivemos atualmente, tendo como ponto de partida o estudo da evolução desses conceitos.

Segundo Lambe (2007), há confusão conceitual entre os termos “conhecimento”, “informação” e “dados”, o que gera confiança excessiva nas tecnologias de processamento de informações e um sentimento de menos valia em relação ao

impacto dos fatores “pessoas”, “processos”, “cultura” e “crenças” na maneira como o conhecimento é criado, compartilhado e aplicado. Conforme Lima e Alvares (2012, p. 25), “a informação obtida por um indivíduo, para se transformar em conhecimento, dialoga com a sua cultura, seus valores e princípios, seu modo de ser e sua maneira de ver e compreender o mundo”. Le Coadic (2004, p. 4) corrobora apontando que o conhecimento está associado ao saber, enquanto a informação é um elemento de sentido, “[...] um conhecimento inscrito (registrado) [...] graças a um sistema de signos (a linguagem)”, ideia complementada por Fogl (1979) ao associar o conhecimento à cognição (Café; Barros; Fonseca, 2014). Há que se considerar também o aspecto do conhecimento acumulado ao longo do tempo, conforme sinalizam Lima e Alvares (2012) ao afirmarem que o termo conhecimento comporta não apenas os processos de significação mental e individual, mas especialmente um conjunto de saberes em constante desenvolvimento e evolução, sendo compartilhados e difundidos em todos os domínios da atividade humana. Na visão dos autores, os princípios que norteiam a organização do conhecimento só tem sentido na CI a partir da visão social de conhecimento. Em relação ao conceito de dados Borgman (2007) apresenta de maneira ampla como a menor unidade de informação. A autora defende que a complexidade do dado reside no fato de que seu contexto de uso ditará sua conceituação, enquanto as diferentes definições existentes para informação traduzem sua complexidade (Sayão; Sales, 2020). Assim, para fins de organização da informação e do conhecimento no presente trabalho, no campo da CI, será adotado o aspecto social dos conceitos de conhecimento, informação e dado, sejam eles aplicados no cotidiano ou nas organizações para fins estratégicos.

Ideias correlatas às de Souza (2007) no âmbito da CI são utilizadas pelo Consórcio *World Wide Web* (W3C) ao se referirem à *web* semântica. Desta vez localizadas no mundo virtual da internet, fala-se sobre a evolução da “*web* de documentos” à “*web* dos dados”, esta também conhecida como “*web* 3.0”[1]. O conceito de *web* semântica conferido pelo W3C envolve desde a interação entre computadores

[1] O termo Web 3.0 refere-se à *web* de dados ou *web* semântica, enquanto web3 é um conceito mais novo cunhado no contexto da tecnologia de *blockchain*. De modo a evitar confusões conceituais entre os termos, o presente trabalho adotará os sinônimos escolhidos pelo W3C, *web* semântica e *web* de dados. O termo *web* 3.0 somente será utilizado na abordagem sobre a evolução do ambiente *web* até sua versão 3.0 (Hedden, 2022; W3C, 2015).

e pessoas até o desenvolvimento de tecnologias, linguagens, padrões e recomendações que tornem as informações legíveis por máquinas (Berners-Lee; Hendler; Lassila, 2001; Catarino; Souza, 2012; Ferreira; Santos, 2013; W3C, 2013, 2015). A concepção de *web* semântica assemelha-se também à visão que teve Paul Otlet, considerado o pai da biblioteconomia e da CI, sobre o *Mundaneum* no início do século XX. O *Mundaneum* era uma instituição que tinha como objetivo reunir, organizar e compartilhar todo o conhecimento humano, porém não pode ser concretizado em sua magnitude original. Muitas das ideias contidas em seu modelo anteciparam fundamentos da organização do conhecimento e da própria CI, além de recursos e funcionalidades da *web* que conhecemos atualmente, a *web* semântica, cuja visão otimiza o acesso à informação no ambiente *web* e a interação homem-máquina.

Os paradigmas históricos evidenciam a evolução da busca e do acesso ao documento e seu conteúdo intelectual para um ambiente mais amplo e irrestrito de busca e acesso a um volume maior de informações em tempo real, idealmente de maneira conectada e relacionada. Entretanto, esse complexo ambiente de compartilhamento de conhecimento “exige um repensar de instrumentos e métodos de intermediação necessários para a produção, organização e disseminação de informação”(Souza, 2007, p. 120).

Ainda, Souza (2007) sinaliza que a CI se caracteriza como campo de investigação científica acerca dos complexos problemas de organização da informação e do conhecimento, o que inclui recuperação. O acesso à informação está diretamente relacionado à facilidade com que ela pode ser encontrada e aos meios pelos quais pode ser visualizada e processada. Zins (2007, p. 479) complementa quando afirma que “o campo da Ciência da Informação está em constante mudança. Portanto, os cientistas da informação devem revisar regularmente – e se necessário – redefinir seus blocos de construção fundamentais”.

A área de organização do conhecimento contém um sem número de recursos e métodos capazes de lidar em algum nível com as questões concernentes à produção, organização e disseminação de informação e conhecimento. No ambiente *web* destacam-se as taxonomias e as ontologias, ambas classificadas como Sistemas de Organização do Conhecimento (SOC). Segundo Zeng (2008), os SOC são

incorporados no ambiente *web* como serviço, agindo como mapas semânticos, que facilitam a descoberta e a recuperação de recursos, orientando usuários humanos e máquinas numa linguagem comum. Lambe (2007), Vignoli, Souto e Cervantes (2013), Hlava (2014) e Hedden (2022), ao comparar taxonomias e ontologias, demonstram suas diferenças e complementaridades em distintos âmbitos de atuação. Ontologias possuem tanto uma acepção filosófica, quando se referem ao estudo do ser e da existência, quanto informacional, quando compreendidas como sistema de classificação, e computacional, quando se traduzem em modelos de representação e compromissos ontológicos processados por agentes inteligentes. Enquanto as ontologias direcionam-se à leitura e ao processamento por máquinas, tratando, portanto, do relacionamento profundo entre dados, as taxonomias são meios flexíveis e amigáveis de representação e acesso à informação e ao conhecimento dentro e fora do ambiente *web*, podendo ser mais ou menos complexas e enriquecidas de elementos tecnológicos, compreensíveis por humanos e máquinas.

Em meio às tecnologias emergentes[2] voltadas à análise e processamento de dados que já se tornaram realidade, muitos autores se perguntam se o desenvolvimento das ontologias no campo da Ciência da Computação (CC) tornou as taxonomias obsoletas. Porém, a confiança excessiva a que se refere Lambe (2007), nas tecnologias de processamento de informações, é um aspecto a ser considerado ao questionarmos se na *web* de dados as taxonomias dariam lugar às ontologias. As tecnologias emergentes vem contribuindo de maneira expressiva para as atividades diárias dos indivíduos e organizações, na automatização de tarefas, na manipulação de grande volume de dados e no aumento da segurança cibernética, por exemplo. No entanto, conforme o uso se torna mais frequente, uma série de limitações e desafios também se colocam, valorizando a intervenção humana na sua superação.

O estudo de Saumure e Shiri (2008), sobre as tendências nas eras pré e pós-*web* dos estudos da organização do conhecimento, demonstra a predominância de temas voltados para a indexação e resumo na era pré-*web*. Já após o advento da *web* dominam temas como informação corporativa; organização do conhecimento assistida

[2] No âmbito das Tecnologias de Informação e Comunicação (TIC), algumas das consideradas tecnologias emergentes são: Inteligência Artificial, *Cloud Computing* e *Machine Learning*. São denominadas emergentes por se tratar de avanços contemporâneos e inovações em variadas áreas.

por máquinas; atuação de bibliotecários como organizadores do conhecimento na *web*; interoperabilidade; classificação das informações no ambiente *web*; aplicação e uso de metadados; entre outros. É possível observar que os principais temas trabalhados possuem forte relação com as taxonomias, o que corrobora a realização de novos estudos sobre a temática nesse contexto. Os autores apontam também que, com o advento da *web* e a exposição das pessoas a um volume informacional maior, se tornou necessário explorar a capacidade do computador de classificar materiais com base no significado e no contexto. No entanto, logo surgiram críticas de que os computadores nem sempre indexavam materiais de maneiras semanticamente significativas, capacidade relacionada aos bibliotecários, e no presente trabalho estendida aos cientistas da informação. Apesar do receio inicial, na era pós-*web* as competências na organização do conhecimento ganham força. À medida que as bibliotecas digitais e o ambiente digital se expandem em complexidade, a categorização efetuada por intervenção humana ganha crescente importância, provavelmente pela limitação dos computadores em apreender verdadeiramente as nuances semânticas sutis presentes na linguagem humana.

Hjørland (2016) afirma que o desenvolvimento da área de organização do conhecimento vem sendo impulsionado pela tecnologia. A *web* semântica surgiu há cerca de vinte e cinco anos atrás e até hoje seu potencial vem sendo explorado e desenvolvido no ensino, na pesquisa, no dia a dia de trabalho das pessoas e nas organizações. A principal questão que se coloca refere-se à melhor compreensão das bases pelas quais organizamos o conhecimento no ambiente *web*, que se tornou parte considerável da vida em sociedade. As taxonomias e as ontologias são conceitos centrais nesse sentido (Hedden, 2022; Hlava, 2014; Saumure; Shiri, 2008; Vignoli; Souto; Cervantes, 2013), sendo estas objeto de estudos em diversas áreas, enquanto poucos esforços são direcionados ao esquema de classificação que pode ser mais facilmente compreendido e utilizado pelas pessoas e pelas organizações, ponto que será melhor analisado adiante (Lambe, 2007; Lopes; Aganette; Maculan, 2020; Mendes; Bentes, 2019).

1.1 PROBLEMA E JUSTIFICATIVA

Deparamo-nos com uma explosão informacional que não vai diminuir de

intensidade, desacelerar ou até mesmo desaparecer (Hlava, 2014), especialmente no mundo virtual, onde predominam atividades sociais e organizacionais, dos processos produtivos e do mundo do trabalho. Trata-se de uma questão que nos demanda a identificação de boas práticas e estratégias para a organização de informações na *web*, não apenas para melhoria da experiência do usuário na busca e recuperação da informação disponível, mas também para a disseminação do conhecimento de forma mais ampla e democrática.

As taxonomias são estruturas comuns no ambiente *web*, onde podem ser lidas e compreendidas tanto por humanos quanto por máquinas. “Fornecem um excelente meio para organizar informações específicas do assunto em um formato facilmente navegável” (Saumure; Shiri, 2008, p. 658). Aquino, Carlan e Brascher (2009, p. 204) apontam que o aparecimento e uso de taxonomias nos ambientes digitais relacionam-se às formas automatizadas de criação da informação, tornando-as alvo de estudos da CI, campo onde as áreas de organização e recuperação da informação e do conhecimento são vistas como centrais para o ensino e para a pesquisa. Segundo as autoras o termo taxonomia é o mais adotado, especialmente neste campo. Porém outros termos como taxionomia e sistemática também podem ser encontrados, este de forma mais frequente no campo originário do conceito, a biologia.

Lambe (2007, p. 1, tradução nossa) lembra que a atividade de “organizar vai além de fazer sentido [...] o que nos habilita a agir sistemática e intencionalmente em relação ao ambiente e isso é atividade primária de gestão”. A atividade de gerir algo significa não apenas controlar seus aspectos, mas também acessar, utilizar meios e recursos e resolver problemas com eficácia. Agir sistematica e intencionalmente em relação à informação vincula-se diretamente com o ato de classificar, que para Piedade (1977, p.9) é “dividir em grupos ou classes, segundo as diferenças ou semelhanças. É dispor os conceitos, segundo suas semelhanças e diferenças, em certo número de grupos metodicamente distribuídos”. Enquanto estruturas classificatórias, as taxonomias “representam os propósitos de organização intelectual de um dado contexto” (Campos; Gomes, 2007, p. 4). Aquino, Carlan e Brascher (2009) afirmam também que os sistemas de classificação são a base para organização e a representação do conhecimento e para isso utilizam elementos oriundos de diferentes campos do saber, que passaram por um aprofundamento teórico-conceitual ao longo

dos anos. Dessa forma, as autoras reforçam a robustez dos fundamentos por meio dos quais as taxonomias se desenvolvem e sinalizam sobre a importância da adoção de princípios classificatórios explícitos e consistentes na sua elaboração.

Birger Hjørland compreende a organização do conhecimento a partir de dois aspectos principais, Processos de Organização do Conhecimento (POC) e Sistemas de Organização do Conhecimento (SOC). Tal distinção enfatiza a importância de compreender não apenas os sistemas, mas também os processos envolvidos na organização do conhecimento, o que pode ajudar a desenvolver abordagens mais efetivas para lidar com a complexidade do conhecimento em diferentes áreas. Enquanto SOC, as taxonomias são produtos desses processos (como descrição, indexação e classificação), cujo resultado é a recuperação da informação (Hjørland, 2016).

Assim, estima-se que as taxonomias possam contribuir para além da recuperação, desempenhando outros papéis como construção de conhecimento, identificação de novas relações de conhecimento, criação de sentido, gerenciamento de complexidade, diagnóstico e tomada de decisão (Lambe, 2007). Aquino, Carlan e Brascher (2009), concordam com Terra et al. (2005), que corroboram essa visão ao apontar alguns dos objetivos das taxonomias: agilizar a comunicação entre especialistas e outros públicos; encontrar o consenso; controlar a diversidade de significação; construir relacionamento semântico; e mapear uma área, servindo como guia em processos de conhecimento. Para os autores, uma taxonomia é acima de tudo um instrumento para alocação, recuperação e comunicação das informações dentro de um sistema sob uma premissa lógica. No contexto da CI, esses autores conceituam taxonomia como um sistema para classificar e facilitar o acesso à informação. Tais afirmações habilitam as taxonomias ao enfrentamento dos desafios do ambiente *web* atual em relação à organização e à recuperação da informação e do conhecimento.

No entanto, observa-se pouco amadurecimento do tema na literatura brasileira, com escassez de revisões sistemáticas e outros estudos sobre taxonomias na ciência da informação, especialmente quando relacionadas ao ambiente *web*, onde prevalecem as ontologias, por seu perfil computacional. Estudo de Lopes, Aganette e Maculan (2020), por exemplo, apontou para o quantitativo pouco representativo de

pesquisas sobre taxonomias na CI, neste caso referindo-se a Teses e Dissertações da Capes. Foram recuperadas apenas sete dissertações no campo no período de 15 anos e nenhuma tese. Além disso, o estudo evidenciou a existência de uma lacuna no conhecimento da ciência da informação sobre o estado da arte dos conceitos, usos e aplicações das taxonomias corporativas e facetadas, compreendidas pelos autores como os tipos de taxonomias mais enfocadas neste campo. Uma das poucas revisões sistemáticas, um estudo de Mendes e Bentes (2019) sobre a presença da taxonomia na biblioteconomia e na CI, apontou maior emprego das taxonomias no âmbito internacional, destacando como provável motivo o desconhecimento das potencialidades que a taxonomia possa trazer para as organizações de um modo geral. Esses trabalhos trazem importantes achados, que evidenciam a necessidade de mais estudos sobre taxonomias na ciência da informação e se somam à ausência no idioma português de classificações dos tipos de relações semânticas na organização e representação do conhecimento, conforme apontaram Maia, de Lima e Maculan (2017).

Diante deste cenário, torna-se fundamental maior produção científica em língua portuguesa sobre a temática, atualização do estado da arte das taxonomias na CI, ampliação do conhecimento sobre taxonomias enquanto estruturas classificatórias no ambiente *web* e o levantamento de lacunas do conhecimento, possibilitando o desenvolvimento de taxonomias mais eficientes e adaptáveis, mitigando os principais desafios vividos na *web* de dados.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Atualizar o estado da arte das taxonomias no campo da ciência da informação no contexto da *web* semântica, a partir da produção científica disponível identificada por meio de revisão integrativa de literatura.

1.2.2 Objetivos Específicos

Mapear a trajetória e a evolução do conceito de taxonomia na Ciência da Informação.

Analisar a evolução da *web* (1.0, 2.0 e 3.0) em direção à *web* semântica.

Identificar e analisar as relações entre taxonomias e *web* semântica.

1.3 CONTRIBUIÇÕES DA PESQUISA

A presente pesquisa busca contribuir com:

a) a produção científica em língua portuguesa sobre a temática das taxonomias e da *web* semântica;

b) a atualização do estado da arte das taxonomias na Ciência da Informação;

c) a ampliação do conhecimento sobre taxonomias enquanto estruturas classificatórias no ambiente *web*;

d) o levantamento de lacunas do conhecimento e sugestões de novos estudos em organização e recuperação da informação e do conhecimento, mitigando os principais desafios vividos na *web* de dados.

Uma pesquisa não se pauta apenas nos interesses acadêmicos do autor, mas pelas contribuições que o trabalho pode oferecer à sociedade e à ciência. Dessa forma são produtos:

a) uma revisão de literatura sobre organização e representação da informação e do conhecimento;

b) um esquema que relaciona e delimita as áreas de OC, OI, RC e RI;

c) uma revisão integrativa sobre o papel das taxonomias na *web* semântica;

d) um glossário com os principais conceitos surgidos ou atualizados no âmbito da revisão integrativa sobre o papel das taxonomias na *web* semântica;

e) uma taxonomia bilíngue das relações entre taxonomias e *web* semântica.

1.4 RESTRIÇÕES

Com a consciência do desenvolvimento de uma pesquisa fundamentalmente teórica, não se pretende esgotar o assunto no presente trabalho. Cabe colocar também que são compreendidas e consideradas as restrições acerca dos métodos e ferramentas disponíveis para a realização da pesquisa, bem como do tempo em que poderá ser conduzida, dentro do que propõe o programa de mestrado, mesmo diante

das possibilidades de extensão.

No que concerne à revisão integrativa, Toronto e Remington (2020) sinalizam que ainda não há padrões e diretrizes robustos o suficiente para guiar a condução de estudos de revisão integrativa, o que pode afetar a qualidade dos resultados da revisão. Para mitigar essa questão, as escolhas metodológicas basearam-se no Guia *Cochrane handbook for systematic reviews of interventions*, o mais bem estabelecido para condução de revisões sistemáticas, da Cochrane Library. A Cochrane Library é uma respeitada biblioteca da área da saúde (medicina e outras especialidades) fornecida pela Colaboração Cochrane, uma rede global composta por outras organizações, pesquisadores, profissionais e sociedade civil interessados no tema. É uma organização sem fins lucrativos e tem como enfoque a tomada de decisão baseada em evidências científicas, tendo se especializado ao longo de cerca de trinta anos na produção de resumos de evidências confiáveis. O arcabouço Cochrane Library, composto por guias, protocolos, ferramentas e *software*, é o padrão internacional para condução de revisões sistemáticas.

1.5 PREMISSA E HIPÓTESES

Partindo da premissa de que as taxonomias são sistemas de organização do conhecimento flexíveis e adaptáveis, formulam-se as seguintes hipóteses:

- a construção de uma taxonomia amplia a capacidade semântica dos objetos digitais disponíveis na *web*, bem como sua interoperabilidade, tornando a rede mais eficiente;
- a evolução das tecnologias da *web* semântica promove novos usos para as taxonomias em meio ao aumento do volume informacional;
- o estabelecimento da *web* semântica gera desafios que exigem a intervenção humana no manejo de taxonomias para organização do conhecimento em diferentes áreas do saber.

2 REVISÃO DE LITERATURA

Esta revisão narrativa de literatura se estrutura em duas partes: a primeira se volta aos principais conceitos da área de Organização do Conhecimento (OC), onde se localizam as taxonomias na Ciência da Informação (CI), distinguindo também as áreas de Organização da Informação (OI), Representação da Informação (RI) e Representação do Conhecimento (RC); a segunda trata da *web* semântica e suas relações com as taxonomias.

2.1 ORGANIZAÇÃO E REPRESENTAÇÃO DA INFORMAÇÃO E DO CONHECIMENTO

Em Aristóteles (384 a.C.-322 a.C.) e Platão (428 a.C.-347 a.C.), marcos referenciais da filosofia ocidental, encontramos as origens das noções de termo, conceito, informação, categoria e conhecimento, desenvolvidas por meio de questões fundamentais acerca da natureza da realidade, da existência humana e do saber. Similaridades e diferenças são encontradas entre os dois pensadores, porém para ambos essas (e outras) ideias são essenciais para a compreensão do mundo e para a comunicação entre as pessoas. O trabalho de Aristóteles, discípulo de Platão, contribuiu para a sistematização do saber humano e para o desenvolvimento da área de classificação, influenciando os estudos de Porfírio (232 d.C. - 304 d.C.) e Ranganathan (1892 - 1972). Em *Isagogé* (Introdução às Categorias) surge a conhecida “Árvore de Porfírio" – técnica de organização da informação e do conhecimento utilizada para representar classes e subclasses de um conceito – cuja hierarquia se tornou a base para a Classificação Decimal Universal (CDU) de Paul Otlet (1868 - 1944). À evolução, os princípios fundamentais de Ranganathan possibiitaram a classificação facetada mais precisa e flexível, revolucionando a organização e representação da informação e do conhecimento em bibliotecas e outros espaços, dentre eles o ambiente *web* (Lima, 2016; Moraes; Arcello, 2000).

De fato, nos fundamentos teóricos e sistemas de classificação se encontram as raízes da Organização do Conhecimento (OC) (Aquino; Carlan; Brascher, 2009; Lima,

2016), que se estabelece a partir de Bliss (1929 e 1936)[3], Soergel (1971[4]) e Dahlberg (1973[5]), os primeiros pesquisadores a adotarem esse termo (Lima, 2016; Lima; Alvares, 2012; Ohly, 2020). Além deles, Mazzocchi (2018) aponta também para as importantes contribuições de Cutter, Richardson e Sayers para a área no início do século XX. A OC foi reconhecida como área pela *International Society of Knowledge Organization* (ISKO)[6] em 1973, teve Dahlberg como sua primeira presidente, e se firmou com a substituição do periódico *International Classification*, publicado de 1974 a 1992, que passou a se chamar *Knowledge Organization*, o veículo internacional com o maior prestígio na área. Assim, dada a proximidade entre as áreas de Classificação e OC, seus embasamentos teóricos muitas vezes se confundem (Dahlberg, 2006; Lima, 2016; Lima; Alvares, 2012). Não se pretende, no presente trabalho, delimitar os fundamentos de cada área, porém é essencial a compreensão da profunda convergência teórica entre ambas.

Delimitação conceitual

A literatura aponta também para a necessidade de delimitação dos conceitos de OI e OC, termos muitas vezes utilizados em conjunto ou como sinônimos. Para Bräscher e Café (2008) é possível a utilização dos dois termos de maneiras distintas no domínio da ciência da informação, uma vez que suas delimitações estão ligadas à distinção entre os conceitos de informação e conhecimento, amplamente explorados pela comunidade científica há décadas (Alvares, 2022; Buckland, 1991; Lima, 2016; Moreira, 2018; Zins, 2007). As autoras adotam a compreensão de Fogl (1979) sobre estes conceitos, onde a informação é uma unidade constituída simultaneamente pelo conteúdo, que é o conhecimento; pela linguagem, a partir da qual é expressa; e pelo suporte físico, a partir do qual ela existe e exerce função social; enquanto o conhecimento é o conteúdo da consciência humana, resultado da cognição, do pensamento. Desse modo, definem OI como um processo que possibilita o acesso ao

---

[3] Cabe apontar os trabalhos de Evelyn Bliss, *The Organization of Knowledge and the System of the Sciences* (OKSS), de 1929, e *The Organization of Knowledge in Libraries*, de 1936.
[4] A tese de doutorado de Dagobert Soergel, denominada *Organization of knowledge and documentation*, publicada na Alemanha em 1971.
[5] A tese de doutorado de Ingetraut Dahlberg, denominada *Foundation of Universal Organization of Knowledge*, de 1973.
[6] Sociedade científica internacional, responsável pelos principais eventos científicos da área de organização do conhecimento.

conhecimento contido na informação, que para ser organizada precisa ter seus aspectos material e de conteúdo descritos, sendo a linguagem aquela que engloba tanto um quanto outro aspecto. Além disso, as autoras compreendem a OI como um processo específico ligado a um documento ou objeto informacional, enquanto a OC estaria ligada não a unidades de informação, mas a unidades de pensamento, preocupando-se com a modelagem de um domínio através de conceitos.

> Em nossa visão, temos dois tipos distintos de processos de organização, um que se aplica às ocorrências individuais de objetos informacionais - o processo de organização da informação, e outro que se aplica a unidades do pensamento (conceitos) - o processo de organização do conhecimento (Bräscher; Café, 2008, p. 6).

Lima (2016, p. 71) concorda quando afirma a complementaridade entre as áreas da OC e da OI:

> [...] e complementam, pois, enquanto na área de OC organiza-se o conhecimento a partir da modelagem de domínios, representando sua estrutura com base na descrição de conceitos e dos relacionamentos semânticos entre eles, na área de OI ocorre o processo de representação do conteúdo de um documento específico, o qual tem o conhecimento registrado, para que ele seja posteriormente recuperado pelo usuário em um Sistema de Recuperação da Informação (SRI).

Para tanto, as autoras se fundamentam nos estudos de Ingetraut Dahlberg e Birger Hjørland, que apresentam visões distintas, porém complementares. Ambos os autores assumem a OC como foco principal e orientador. Hjørland (2016), cuja leitura é mais aplicada, se refere à OC como campo de pesquisa, ensino e prática. Dahlberg aprofunda teoricamente a discussão e, em seu trabalho "*Knowledge Organization: A New Science?*" (2006), aborda a OC na busca por uma atuação conjunta da comunidade científica no que a autora considera que poderia ser uma nova ciência.

Para Hjørland (2016, p. 475, tradução nossa) OC "é sobre descrever, representar, arquivar e organizar documentos, representações de documentos, assuntos e conceitos tanto por humanos quanto por programas de computador". Esta compreensão evidencia a OI (mais específica) como parte da OC (mais ampla) e reforça a atribuição que o autor faz da OC como descrição, indexação e classificação, uma visão do que o autor denomina Processos de Organização do Conhecimento (POC). Complementa ao abordar também os Sistemas de Organização do

Conhecimento (SOC), onde se localizam os conceitos e os relacionamentos semânticos indicados entre eles, dando origem a diferentes tipos de SOC, dentre eles as taxonomias. POC e SOC são, segundo o autor, os dois principais aspectos da OC.

Em trabalhos mais recentes Hjørland atualiza e discute as tradições de pesquisa, as abordagens e as questões teóricas básicas em OC. No artigo *Knowledge Organization* (KO) (2016) o autor afirma que as abordagens de OC são geralmente divididas em duas vertentes: abordagem baseada em humanos, que se refere aos métodos que demandam intervenção humana, e abordagem baseada em máquinas, que se refere aos métodos que se utilizam de algoritmos e outras tecnologias computacionais. O autor afirma que essas abordagens não são suficientemente claras na prática, além de possuírem subtipos muito diferentes entre si, o que dificulta sua compreensão. Como tanto humanos quanto máquinas podem ou não estar envolvidos em atividades de OC, o autor sugere oito categorias de abordagens consideradas as mais influentes e importantes atualmente dentro e fora da OC (Hjørland, 2016, p. 476–480, tradução nossa).

Abordagens internas à organização do conhecimento: abordagens advindas de estudos realizados pela comunidade científica de OC:

- Abordagens práticas e intuitivas: tem como enfoques a estabilidade e a centralização, priorizando o aspecto prático[7] nas escolhas e valorizando a intuição do classificacionista. O modelo apontado nessas abordagens é o mais adotado mundialmente segundo o autor, o Sistema de Classificação Decimal de Dewey (CDD)[8];
- Abordagens baseadas em consenso: adotam a visão de Henry E. Bliss, que defendia a classificação das bibliotecas pautada pelo consenso científico, de modo que a literatura acadêmica fosse a base para a identificação de padrões de concordância;
- Abordagens analitico-facetadas: referem-se ao paradigma analítico-

[7] Hjørland sinaliza que o termo escolhido, prático, não deve ser compreendido como pragmático, uma vez que, na visão do autor, este seria um conceito à parte, diferente da noção de praticidade.
[8] Nome original da Classificação de Melvil Dewey, *Dewey Decimal Classification system (DDC)*, cuja primeira edição foi publicada em 1876.

facetado[9], desenvolvido por Shiyali Ramamrita Ranganathan e pelo Classification Research Group, baseado em princípios milenares de divisão lógica e considerado pelo autor a abordagem teórica que provavelmente seja a mais explícita e "pura" já desenvolvida para a organização do conhecimento na ciência da informação. Tem como pontos fortes seus princípios lógicos e a forma como provê estrutura aos sistemas de organização do conhecimento, apesar de não apresentar bases empíricas fortes e metodologia rigorosa de ordenamento do conhecimento;

- Abordagens cognitivas e baseadas no usuário: trata das tradicionais abordagens baseadas em estudos de usuário, porém o autor diferencia essa família de abordagens daquelas denominadas *user-friendly* (fácil de usar) e relativiza sua importância, já que sistemas de organização bem sucedidos muitas vezes baseiam-se em outros tipos de critérios como estatísticos e lógicos, e não em estudos empíricos de usuários. Para o autor, a característica mais importante é ser amigável e não baseada na opinião do usuário;
- Abordagens epistemológicas ou de análise de domínio: seguindo sentido contrário às abordagens baseadas em consenso científico, nas análises de domínio o autor considera que o princípio-chave é a coexistência de diferentes perspectivas sobre um mesmo objeto, uma vez que cada área envolvida contribui com uma visão diferente. Aponta, porém, que ainda assim o consenso é um elemento valorizado cientificamente e importante no trabalho de análise de um domínio. O autor coloca que, mesmo que diferentes domínios possam apresentar diferentes graus de consenso, o mais importante é considerar sempre tanto os pontos fortes quanto as fragilidades existentes em diferentes perspectivas, teorias e paradigmas ao longo do desenvolvimento do domínio. Esse ponto de vista se associa, e é citado pelo autor, à visão de Dahlberg sobre a organização do

[9] Em estudos de outros autores e também nos estudos anteriores de Hjørland é possível encontrar o termo sistema analítico-facetado, porém nesse estudo o autor adotou o termo paradigma analítico-facetado.

conhecimento como uma ciência da ciência, isto é, parte do método científico, “ferramenta” fundamental na compreensão de um campo;

Abordagens externas à organização do conhecimento: abordagens concorrentes desenvolvidas fora da área de OC:

- Abordagens bibliométricas: abordagens como altimetria, infometria, webometria e cientometria, consideradas pelo autor áreas bastante desenvolvidas, com pontos fortes e fracos, que pode ser vista como concorrente ou complementar à organização do conhecimento, uma vez que as análises realizadas e as informações geradas possuem importante valor agregado, especialmente no que concerne à recuperação da informação;
- Abordagens de recuperação da informação: trata-se, segundo o autor, de uma abordagem que migrou da ciência da informação para a ciência da computação e estuda a relação estatística entre termos, documentos e coleções de documentos, representando abordagens concorrentes à OC e muitas vezes consideradas mais bem-sucedidas.
- Outras abordagens: neste ponto o autor escolhe duas abordagens dentre várias como enfoque e exemplo, a abordagem ontológica, que envolve a criação de um sistema estruturado com especificações formais de definições e conceitos para representar o conhecimento em um campo específico, e a abordagem de gênero, que enfatiza a importância de compreender o contexto de atividade daqueles que produzem e usam conhecimento e informação.

As oito categorias de abordagens sugeridas por Hjørland revelam diferentes aspectos da organização do conhecimento, facilitando seu entendimento, além de demonstrarem uma vez mais a forte interdisciplinaridade da área.

No trabalho de Dahlberg destacam-se contribuições seminais para a área da OC, o sistema de classificação denominado *Information Coding Classification* (ICC) e a Teoria do Conceito, os quais, segundo Gomes e Campos (2019, p.8), “fornecem elementos para sistematizar o Universo não somente por categorias, mas também, pela natureza da atividade de pesquisa”. As autoras apontam que Dahlberg buscava

desenvolver seus estudos numa perspectiva científica, o que corresponde à uma compreensão disciplinar de OC, onde seu desenvolvimento se pauta em métodos científicos verificáveis e consensos. “Desta forma, conceito para Dahlberg é uma unidade de conhecimento, pois é necessário que exista um acordo entre as pessoas para que algo possa ser comunicável” (Gomes; Campos, 2019, p. 39). Nesse sentido, as autoras afirmam que o entendimento de Dahlberg da OC possui três dimensões: (1) a organização do conhecimento em si; (2) a organização de cada universo específico de conhecimento; e a (3) organização das unidades de conhecimento no interior de um universo específico. O artigo da enciclopédia ISKO sobre a autora esclarece os conceitos que envolvem essa compreensão:

> De acordo com Dahlberg, organização do conhecimento é a ciência da estruturação e arranjo sistemático de unidades de conhecimento (conceitos) de acordo com seus elementos de conhecimento inerentes (características) e a aplicação dos conceitos e classes de conceitos assim ordenados para a descrição de conteúdos que valem a pena conhecer de objetos de qualquer tipo (Ohly, 2020, n.p., tradução nossa).

Para Dahlberg, a OC está relacionada à construção de sistemas conceituais, enquanto a OI se refere à capacidade de relacionamento entre as unidades desses sistemas e os objetos de informação de qualquer tipo, tais como objetos de um museu, um documento (físico ou virtual), um animal, entre outros (Lima; Alvares, 2012). A referência de Wüster sobre os “sistemas parciais de conceitos" é evoluída por Dahlberg, que aprofunda os tipos de relacionamento. “A Teoria do Conceito que vai fornecer o método de identificação do conceito e as bases para relacionamentos consistentes” (Gomes; Campos, 2019, p. 52). A OC em Dahlberg envolve, assim: (a) unidades de conhecimento (termos) e (b) objetos relacionados de todos os tipos, que por sua vez estão relacionados a termos correspondentes ou classes conceituais, sendo os termos a aparência externa de algo que pode e deve ficar explícito (Gomes; Campos, 2019; Ohly, 2020).

O “Dicionário de Organização do Conhecimento” de Barité et al. (2015, p. 120) reforça e se baseia nos achados anteriormente citados a partir de diversos autores da área. É importante notar que o dicionário apresenta apenas a definição de OC, não uma específica para a OI. A definição de OC apresentada reúne elementos tanto das compreensões de OI, quanto de OC, conforme a seguir:

> Área do conhecimento de formação recente, que estuda as leis, princípios e procedimentos pelos quais o conhecimento especializado em qualquer disciplina é estruturado, com a finalidade de representar tematicamente e recuperar as informações contidas em documentos de qualquer tipo, por meios eficientes que dêem uma resposta rápida às necessidades dos usuários.
>
> A Organização do Conhecimento é alimentada pelas contribuições recebidas da Ciência da Computação, Linguística, Terminologia e Ciências da Informação. García Marco (1995) aponta que é "uma disciplina científica (isto é, sujeita ao método científico) e social, de natureza aplicada", pois seu objetivo é "otimizar a circulação do conhecimento na sociedade.
>
> O objeto de estudo da Organização do Conhecimento é o conhecimento socializado ou registrado, e no que se refere à Biblioteconomia, Documentação e Ciência da Informação, responde por: a) desenvolvimento teórico prático para a construção, gestão, uso e avaliação de sistemas de organização do conhecimento (classificações, taxonomias, nomenclaturas, tesauros, listas, ontologias de assuntos e outros vocabulários); b) a teoria e prática dos processos de classificação e indexação; c) a análise temática da informação em geral, considerando aspectos semânticos, cognitivos formais e computacionais (Barité et al., 2015, p. 120).

As contribuições de Dahlberg e Hjørland estruturaram e ampliaram a compreensão da OC como prática construída social e culturalmente, que considera diferentes perspectivas e objetivos. Lima e Alvares (2012, p. 27) explicam essa construção ao indicar interseções entre os estudos da organização social do conhecimento e da organização do conhecimento, que compartilham “teorias, tradições, ideologias e paradigmas”. Trata-se de um fenômeno que impacta desde a divisão social do trabalho até as inovações e está presente em tantos aspectos da nossa vida cotidiana que se tornou natural agrupar, separar, hierarquizar e centralizar informações e ideias de forma intuitiva e contínua (Hjørland, 2017, 2019; Mazzocchi, 2018; Moreira, 2018; Silva; Baptista, 2015).

Uma vez definido o conceito mais abrangente (OC), cabe retomar os elementos constituintes da OI e sua definição para melhor compreensão. Conforme exposto, tanto a OC quando a OI tem como alvo finalístico o atendimento das necessidades do usuário (físico ou virtual) por meio da recuperação da informação ou dos objetos informacionais, que possuem o conhecimento registrado. Lima e Alvares (2012, p. 35) apontam também que “outra perspectiva do objetivo da organização da informação é

fornecer acesso ao conhecimento estruturado". Sinalizam que, para ser acessada, a informação necessita estar estruturada, isto é, decomposta, descrita, analisada. Essa noção se relaciona ao que Bräscher e Café (2008) apontam quando dizem que, para que a informação seja organizada é necessário que seus aspectos material e de conteúdo sejam descritos. A OI está, portanto, ligada a processos individuais de organização dos objetos de acervos tradicionais ou eletrônicos, por meio da descrição física e de conteúdo desses objetos. O resultado desse processo, conforme concordam os autores, é a informação representada, que poderá então ser recuperada (Bräscher; Café, 2008; Café; Barros; Fonseca, 2014; Lima, 2016; Lima; Alvares, 2012).

E quanto à Representação?

Para Lima e Alvares (2012, p. 22), "representar é o ato de utilizar elementos simbólicos – palavras, figuras, imagens, desenhos, mímicas, esquemas, entre outros – para substituir um objeto, uma ideia ou um fato", que no âmbito da CI diz respeito às formas de simbolizar a informação (RI) e o conhecimento (RC). Bräscher e Café (2008, p. 5) consideram que: "no contexto da OI e da RI, temos como objeto os registros de informação. Estamos, portanto, no mundo dos objetos físicos, distinto do mundo da cognição, ou das ideias, cuja unidade elementar é o conceito". Já no contexto da OC e da RC, por representação do conhecimento:

> "entendemos não apenas a estrutura lógica da representação conceitual, mas também todas as questões de nomeação dos conceitos pelos termos mais adequados, onde as questões da terminologia também devem ser consideradas" (Dahlberg, 2006, p. 12, tradução nossa).

Lima e Alvares (2012) apontam problemas teóricos em relação ao uso do termo "representação do conhecimento", para cujo conceito não há padronização de terminologia e de compreensão. Apesar disso, os autores definem a RC como "um esforço, nas diversas ciências, de materializar o que ocorre na mente humana e na dinâmica do conhecimento" (2012, p. 33), relacionando então a RC aos modelos de representação:

> [...] que podem ser descritivos, matemáticos, visuais, informatizados, dentre outros, conforme as necessidades e especificidades de cada área para se aproximar dos processos cognitivos que envolvem a aprendizagem, o raciocínio, as relações que ocorrem entre os conhecimentos disponíveis na memória... (Lima; Alvares, 2012, p. 34).

Tem-se então que, enquanto os modelos de abstração ou modelos de representação são resultado da RC, os elementos que descrevem as características de determinado objeto informacional são resultado da RI. Segundo Toutain (2007, p. 90), o termo representação apresenta funções documental e simbólica e “consiste em perceber, descrever, gravar e interpretar uma informação”. Em última análise, representar é resultado do processo descritivo necessário à organização (Bräscher; Café, 2008). Citando Bräscher e Café, o “Tesauro Brasileiro de Ciência da Informação” traz como significado do termo representação da informação os: “elementos descritivos que representam os atributos de um objeto informacional específico ou de seu conjunto, tendo como produtos catálogos, índices, bases de dados etc.” (Pinheiro; Ferrez, 2014, p. 196). Baptista (2007, p. 187) complementa:

> Em qualquer ambiente (convencional ou digital), no entanto, o que se procura é a precisão e a eficácia na representação do objeto, tendo em vista a sua transformação em recurso informacional, o que deve resultar, por sua vez, em agilização e autonomia operacional no acesso à informação. Para que esse resultado seja alcançado, contudo, torna-se imprescindível a utilização eficiente das linguagens adequadas (documentárias/metalinguagens) que [...] produzirão um registro fiel, - um verdadeiro sucedâneo, mediante o qual o recurso poderá ser rapidamente identificado, analisado e utilizado e de acordo com o interesse específico do usuário.

Nesse sentido, Catarino e Souza (2012, p. 84) apontam que a RI “tem como principal objetivo proporcionar a comunicação entre os documentos e os usuários da informação”, podendo ser subdividida de forma complementar em representação descritiva, onde se enquadram as características individuais do documento como a descrição bibliográfica, e representação temática, responsável pela atribuição de assuntos, indexação e elaboração de resumos. Sinalizam ainda importante dificuldade conceitual existente na literatura sobre estes termos, tratados de forma separada e provavelmente adotados como uma ampliação do conceito de catalogação, compreendida como o processo que gera o registro bibliográfico (Catarino; Cervantes; Almeida, 2015; Catarino; Souza, 2012; Lima; Alvares, 2012). Lima e Alvares (2012) concordam com a subdivisão ao indicar que a RI torna explicita a identificação dos autores, do contexto de produção e do conteúdo de determinado objeto informacional, e envolve dois passos principais de análise que tem como resultado: (a) uma expressão linguística que caracteriza o documento, e (b) a atribuição de conceitos ao

documento.

Baptista (2007, p. 187) aponta também para o aspecto interdisciplinar que a evolução das tecnologias de informação e comunicação incutiram à área de RI:

> Tal representação extrapolou a alçada exclusiva de profissionais da informação, especialmente os bibliotecários, passando a exigir uma integração de diferentes tipos de profissionais, entendidos como atores envolvidos no processo mais amplo da organização da informação. Trata-se, efetivamente, de um conjunto de atividades interdisciplinares, colaborativas e descentralizadas, as quais demandam integração de esforços, conhecimentos, e utilização intensiva de tecnologia.

Por este motivo, diversos autores apontam que há na CI uso inadvertido e ambiguidades no emprego dos termos organização do conhecimento; organização da informação; representação do conhecimento; representação da informação (Alvarenga; Silva, 2010; Bräscher; Café, 2008; Lima; Alvares, 2012). Há forte interação entre esses conceitos e intersecções dentro do campo da CI com as ciências cognitivas, com a linguística, com a semiótica, com a ciência da computação (Lima; Alvares, 2012).

Além disso, Moreira (2018, p. 26) sinaliza a indissociabilidade entre a organização e a representação, indicando que:

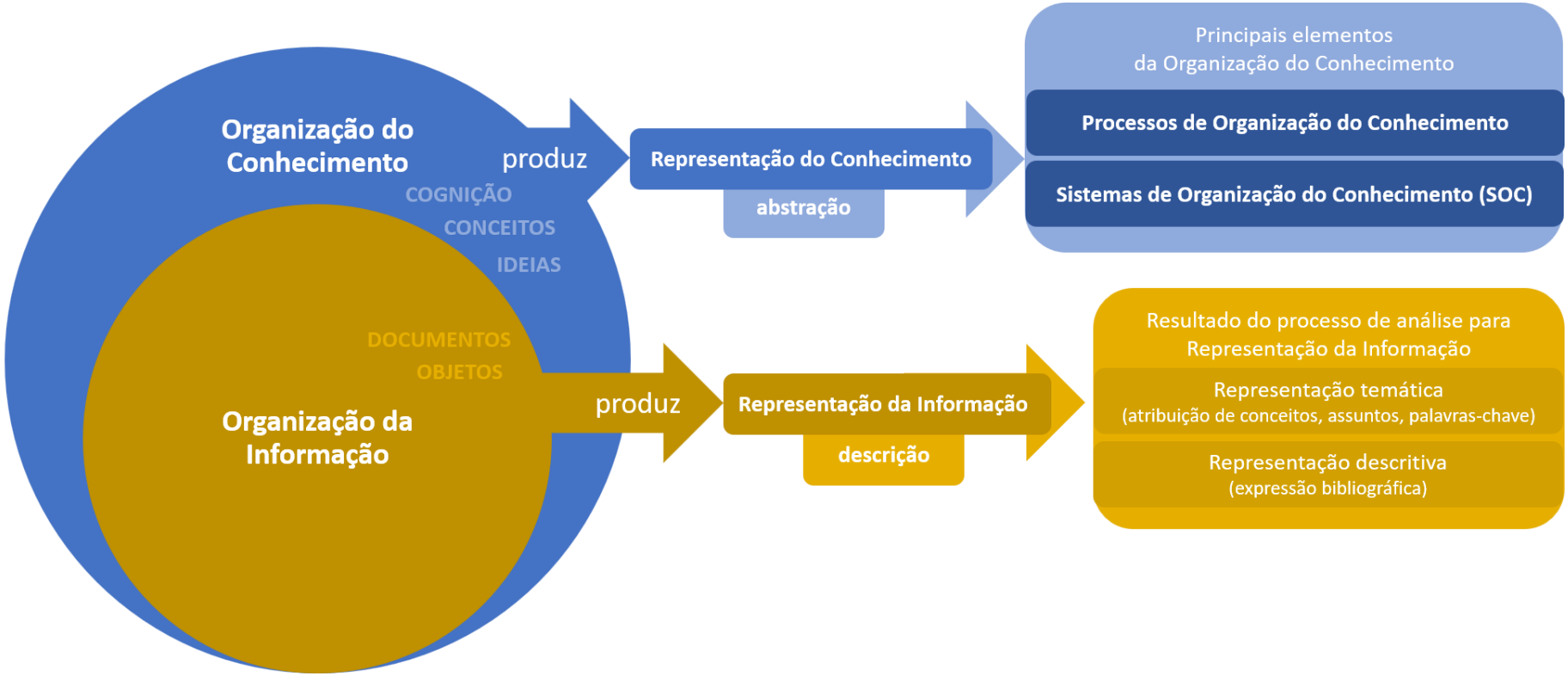
> “o conceito de representação está presente tanto nos sistemas quanto nos processos de organização do conhecimento [...], uma vez que o ‘conhecimento’ que é objeto da organização, em sentido estrito, é materializado (representado)[...]”.

Conforme as definições de OI e RI já citadas é possível que a mesma indissociabilidade ocorra, uma vez que se torna difícil separar a análise relativa ao processo de organização da informação do resultado material que é a representação.

Recuperação da informação: objetivo comum

Em resumo, em suas diferentes abordagens e no âmbito da CI, a OC tem como assuntos centrais os sistemas de classificação e os sistemas conceituais e, enquanto área, inclui atividades de OI e RI, relativas aos objetos informacionais, e RC, relacionadas aos conceitos, ambas com vistas à recuperação da informação (Barité et al., 2015; Bräscher; Café, 2008; Lima, 2016), visão demonstrada na Figura 1.

Figura 1. Organização e Representação da Informação e do Conhecimento



Fonte: elaborado pelo autor (2023)

2.1.1 Princípios Classificatórios

Para Hjørland (2008) a biblioteconomia e a ciência da informação somente serão capazes de avançar na compreensão da OC a partir do aprofundamento teórico sobre os princípios classificatórios que a regem. Aquino, Carlan e Brascher (2009) concordam ao afirmarem que os sistemas de classificação são a base para organização e a representação do conhecimento e para isso utilizam princípios classificatórios oriundos de diferentes campos do saber, são eles: a categorização; a diferenciação entre termos e conceitos, o estabelecimento de hierarquias e a multidimensionalidade dos termos.

Desde a antiguidade o homem possui o hábito de classificar, o que ao longo dos tempos tornou a classificação parte do cotidiano social (Aquino; Carlan; Brascher, 2009). Schiessl e Shintaku (2012) propõem uma divisão entre a compreensão do termo e do conceito de classificação. Sobre o uso do termo classificação, argumentam que ele pode assumir quatro acepções de distinção tênue entre si: (*i*) colocar em determinada classe, o que corresponde ao processo de rotulagem; (*ii*) determinar a classe à qual pertence uma entidade; (*iii*) definir classes, processo que difere da rotulagem e se relaciona à estruturação dos conceitos; e (*iv*) alcançar determinado nível, que se refere às decisões sobre pertinência e ordenação. Já no âmbito da conceituação, Piedade (1977, p.9) entende que classificar é “dividir em grupos ou classes, segundo as diferenças ou semelhanças. É dispor os conceitos, segundo suas semelhanças e diferenças, em certo número de grupos metodicamente distribuídos”.

Os sistemas de classificação, segundo Zeng (2008, p. 161, tradução nossa) são “arranjos hierárquicos e facetados de notações numéricas ou alfabéticas para representar tópicos amplos”. São também chamados esquemas de classificação, diferenciando-se, segundo Schiessl e Shintaku (2012, p. 63), no sentido de constituírem “a representação gráfica das categorias, classes, suas divisões e arranjos”, apesar de serem utilizados muitas vezes como sinônimos pela literatura. A classificação bibliográfica é um tipo específico de classificação que surge com o objetivo de organizar informações em bibliotecas, sua estrutura dá origem aos sistemas de classificação bibliográfica, onde os mais comuns são: Classificação Decimal de Dewey (CDD); Cassificação expansiva de Cutter; Classificação Decimal Universal (CDU); Classificação do congresso americano ou *Library of Congress*

Classification (LCC); Classificação de assuntos de Brown; Classificação bibliográfica de Bliss; e Classificação de dois pontos de Ranganathan, a primeira classificação facetada (Aquino; Carlan; Brascher, 2009; Schiessl; Shintaku, 2012). Esta última representou, segundo Aquino, Carlan e Brascher (2009), uma verdadeira ruptura na metodologia vigente, uma vez que as classificações existentes até então uniam os assuntos basicamente por meio de relacionamentos hierárquicos. A adoção de uma classificação analítico-sintética baseada em facetas possibilitou o estabelecimento de relações semânticas e a organização do conhecimento por meio de grandes classes e conceitos básicos, ou elementos, de acordo com certas características. Para Dahlberg (2006), a essência da formação de classes é determinar a proximidade, semelhança ou parentesco especial de diferentes itens para a possibilidade de agrupá-los e organizá-los.

2.1.1.1 Categorização

Conforme Campos e Gomes (2007, p. 200):

> categorização é o processo mental de dividir os conceitos em grupos gerais ou categorias amplas compreendendo certos componentes que partilham similaridades em termos de atributos num determinado contexto. As categorias são elementos agregadores que reúnem os conceitos mais altos de um sistema de classificação, permitindo a visualização de uma área como um todo.

As autoras afirmam que a noção de categorias foi introduzida nos sistemas de classificação por Ranganathan, por meio de sua proposta baseada em cinco Categorias Fundamentais, que seriam as categorias mais genéricas e capazes de classificar todos os objetos da natureza conhecidos pelo homem. São conhecidas pelo acrônimo PMEST: Personalidade, que distingue um assunto, normalmente entidades, objetos ou sistemas; Matéria, material físico do qual um assunto pode ser composto, substância, componentes; Energia, ações, atividades, processos que ocorrem em torno de um assunto; Espaço, que corresponde ao componente geográfico da localização de um assunto; e Tempo, ou período associado a um assunto. “As manifestações dessas categorias nas diversas classes que compõem o sistema de classificação, recebem o nome de facetas” (Moreira, 2018, p. 91).

Conforme apontado por Moreira (2018), é por meio dos SOC que se estabelecem sistemas categoriais que permitem processos de organização do conhecimento, como

a classificação e a indexação, além da estrutura de relacionamentos hierárquicos e associativos. A categorização atua como um mecanismo cognitivo dedutivo e fundamental, que simplifica a experiência de compreensão de um domínio ou a busca e recuperação da informação num sistema. O reconhecimento das semelhanças dos itens de uma classe e a distribuição das categorias permitem que agentes humanos e máquinas encontrem ou estabeleçam uma ordem em um ambiente complexo. A categorização, por meio da utilização de um SOC, implica manifestar formalmente um ponto de vista em relação às particularidades e às configurações epistemológicas de um determinado campo ou domínio.

2.1.1.2 Diferenciação entre termos e conceitos

Segundo Dahlberg (2006) e Hjørland (2008), a organização do conhecimento teve sua fundamentação teórica como disciplina científica embasada nas relações entre conceitos, fundamentados na Teoria do Conceito da primeira autora. Está intimamente ligada também à Teoria da Classificação Facetada (TCF) de Ranganathan (1978) e à Teoria Geral da Terminologia (TGT) de Wüster (1981), visto que ambas buscaram estabelecer uma relação clara entre termos e conceitos, no sentido de uma comunicação precisa e consistente, impactando diretamente a compreensão das ideias dentro de um SOC e o processo de sua construção.

Segundo Aquino, Carlan e Brascher (2009, p. 201) a TGT entende o termo como “a unidade de comunicação que representa o conceito e pode ser constituído de uma ou mais palavras, uma letra, um símbolo gráfico, uma abreviação, uma notação”, funcionando também como símbolo linguístico do conceito dentro de um sistema de conceitos; enquanto conceitos são como uma unidade de pensamento. Dahlberg (1978, p.12) define conceito como “a unidade de conhecimento que surge pela síntese dos predicados necessários relacionados com determinado objeto e que, por meio de sinais linguísticos, pode ser comunicado”.

Conceitos, termos e relacionamentos são elementos essenciais da representação temática da informação em qualquer SOC, são eles que permitem que as informações sejam organizadas, recuperadas e compreendidas. Desse modo, enquanto os conceitos são ideias abstratas representadas por termos e atuam como unidades básicas de conhecimento, seus relacionamentos, isto é, as conexões entre

essas ideias, constituem unidades dos SOC, conferem semântica ao sistema. Essa compreensão revela o papel dos termos e dos conceitos na recuperação precisa e eficiente de informações, constituindo a base para a criação de vocabulários controlados. É o relacionamento entre termos e conceitos existente nos vocabulários controlados que possibilita o estabelecimento de uma linguagem comum entre diferentes comunidades discursivas e a descoberta de assuntos relacionados. Além disso, contribuem de maneira indireta para a descoberta de novos conhecimentos e para interoperabilidade entre diferentes sistemas, especialmente no ambiente *web*. Os relacionamentos serão abordados adiante, com enfoque no tipo hierárquico, característico dos SOC denominados taxonomias (Aquino; Carlan; Brascher, 2009; Melo e Bräscher, 2014).

2.1.1.3 Relacionamentos hierárquicos

Aquino, Carlan e Brascher (2009, p. 202) apontam que as relações de hierarquia "surgem da necessidade de especificar, na maioria dos casos, termos genéricos e específicos e desta forma, estabelecer os mais diversos tipos de relações". Referindo-se às relações categoriais e aos termos as autoras indicam duas propriedades lógicas: a extensão e a compreensão. De modo que a extensão conceitual existente a partir de um termo, traduza o grau de compreensão que se faz dele e vice versa. Essa distinção expõe as diferentes complexidades que um conceito pode ter, a importância do estabelecimento de termos assertivos e, principalmente a necessidade de realizar divisões que facilitem a apreensão do conhecimento por meio de processos de categorização ou classificação. É o relacionamento entre os termos (enquanto expressões dos conceitos) que constituem os relacionamentos hierárquicos e estabelecem relações de subordinação, superordenação e coordenação em uma estrutura de organização do conhecimento. Conforme descrevem as autoras:

> Num sistema de classificação hierárquico há classes que são subdivididas em subclasses, que também são divididas em unidades subseqüentes, sempre agregando uma característica ou propriedade particular que a distingue da classe posterior (Aquino; Carlan; Brascher, 2009, p. 202).

Desse modo, conforme Zeng (2008), no nível de superordenação uma classe mais ampla inclui uma ou mais específicas; na subordinação, uma classe é um exemplo ou subtipo de outra mais ampla; enquanto na coordenação as classes

guardam uma relação de mesmo nível hierárquico entre si. Nesta lógica, os relacionamentos hierárquicos dividem-se geralmente em:

- relacionamento gênero-espécie: relacionamento entre conceitos de mesma natureza, um conceito mais geral (gênero) inclui um ou mais conceitos mais específicos (espécie);
- relacionamento partitivo: ou relação todo-parte, onde a compreensão da parte depende da compreensão do todo em algum grau; e
- relacionamento polihierárquico: também denominada multidimensionalidade, ocorre quando os itens pertencem a mais de uma classe e por isso seus relacionamentos são mapeados para mais de um conceito de superordenação.

Nesse contexto, Lambe (2007) sinaliza que a noção de hierarquia possui múltiplos significados em diferentes disciplinas. Cita a sociologia e antropologia, onde hierarquia significa a disposição relativa de poder, e as ciências puras, onde as hierarquias tem um sentido muito preciso e diferente, que inclui critérios claros. O mesmo ocorre na biblioteconomia e na CI, onde hierarquias e polihierarquias são também formas taxonômicas. As taxonomias, que serão abordadas em maior profundidade nos tópicos seguintes, são estruturas classificatórias, isto é, um tipo de SOC. Na atividade taxonômica a hierarquia é um tipo específico de estrutura em árvore, onde cada nó representa um conceito ou termo e seus ramos representam relações hierárquicas entre os conceitos. Tal distinção nos habilita à compreensão dos requisitos científicos, citados pelo autor, para que uma estrutura seja considerada hierárquica:

- inclusão: a categoria de topo inclui todas as categorias subordinadas na árvore;
- consistência relacional: o tipo de relacionamento entre cada nível da hierarquia é exatamente o mesmo;
- herança: as categorias subordinadas numa hierarquia herdam todos os atributos das categorias de maior nível;
- mutuamente exclusivas: uma entidade pertence a uma e somente uma classe - esse aspecto das hierarquias elimina totalmente as ambiguidades.

Na visão de Lambe (2007), os requisitos científicos revelam as funções dos SOC,

reforçando a consistência e identificando as regras de relacionamento que produzem previsibilidade. Ao mesmo tempo, no mundo real as categorias se rebelam contra esse nível de rigor, apresentando-se inconsistentes e ambíguas sob diferentes pontos de vista, o que culmina nas polihierarquias e outras estruturas mais flexíveis (porém não menos consistentes). As polihierarquias são hierarquias que quebram as próprias regras de inclusão, consistência relacional e exclusividade ao permitirem que um mesmo termo específico esteja subordinado a mais de um termo genérico. Assim os termos ou assuntos podem ser organizados em diferentes categorias ou facetas, com as quais guardará ligações semânticas estabelecidas na estrutura conceitual na qual está inserido (Aquino; Carlan; Brascher, 2009).

2.1.2 Sistemas de Organização do Conhecimento

Anteriormente foram delimitados aspectos relativos à organização e representação da informação e do conhecimento, perspectiva pela qual Vignoli, Souto e Cervantes (2013, p. 62) abordam os Sistemas de Organização do Conhecimento (SOC) ao afirmar que eles:

> [...] atendem tanto a OI e a RI, quanto a OC e a RC, e isso sempre ocorrerá quando o conhecimento for representado em informação, num devir incessante. [...] são úteis para organizar, padronizar e possibilitar a recuperação da informação e do conhecimento, ou vice-versa, por meio de sistemas como a ontologia e a taxonomia (Vignoli; Souto; Cervantes, 2013, p. 62).

Ingetraut Dahlberg, Elaine Svenonius e Birger Hjørland contribuiram substancialmente para a área abordando em seus trabalhos aspectos espistemológicos e as bases teorico-metodológicas sobre as quais os SOC foram desenvolvidos ao longo do tempo. Dahlberg “estabeleceu as bases para que a organização do conhecimento fosse reconhecida como um campo ou ciência distinta, que se preocupa com a ordem do conhecimento (ou seja, uma metaciência)” (Mazzocchi, 2018, n.p.); já Svenonius, “contribuiu para estabelecer as bases intelectuais da organização do conhecimento (ou, como em suas palavras, informação)” (2018, n.p.); enquanto Hjørland enfatiza a importância da teoria dos conceitos de Dahlberg para a área, sendo eles as unidades básicas de pensamento e conhecimento e as relações entre eles as unidades dos SOC (Mazzocchi, 2018, n.p.).

Segundo Hjørland (2008), os SOC compartilham das diferentes compreensões

(ampla e restrita) e dos aspectos interdisciplinares da área de organização do conhecimento. Assim, no sentido amplo há a organização social do conhecimento (por exemplo, em disciplinas acadêmicas e na divisão social do trabalho) e a organização intelectual/cognitiva do conhecimento em princípios e teorias (por exemplo, a tabela periódica da química e a taxonomia biológica); este sentido é abordado também por disciplinas como sociologia do conhecimento, linguística e metafísica.

Já no sentido restrito está a leitura feita nos campos da Biblioteconomia e da Ciência da Informação (Mazzocchi, 2018), em cuja literatura se encontram definições que caracterizam os SOC com base em diferentes termos, os quais correspondem às estruturas e funções exercidas conforme contexto de aplicação: representações, modelos de abstração, sistemas conceituais, conjunto de termos/conceitos e ferramentas semânticas (Carlan e Bräscher, 2011; Hjørland, 2016; Mazzocchi, 2018; Vignoli; Souto; Cervantes, 2013; Zeng, 2008). Em estudo mais recente, Hjørland aprofundou questões epistemológicas da área, no sentido de aproximação entre as visões ampla e restrita dos SOC e melhor compreensão entre suas relações e as preocupações pautadas na realidade de convivência social com novas tecnologias (Hjørland, 2021).

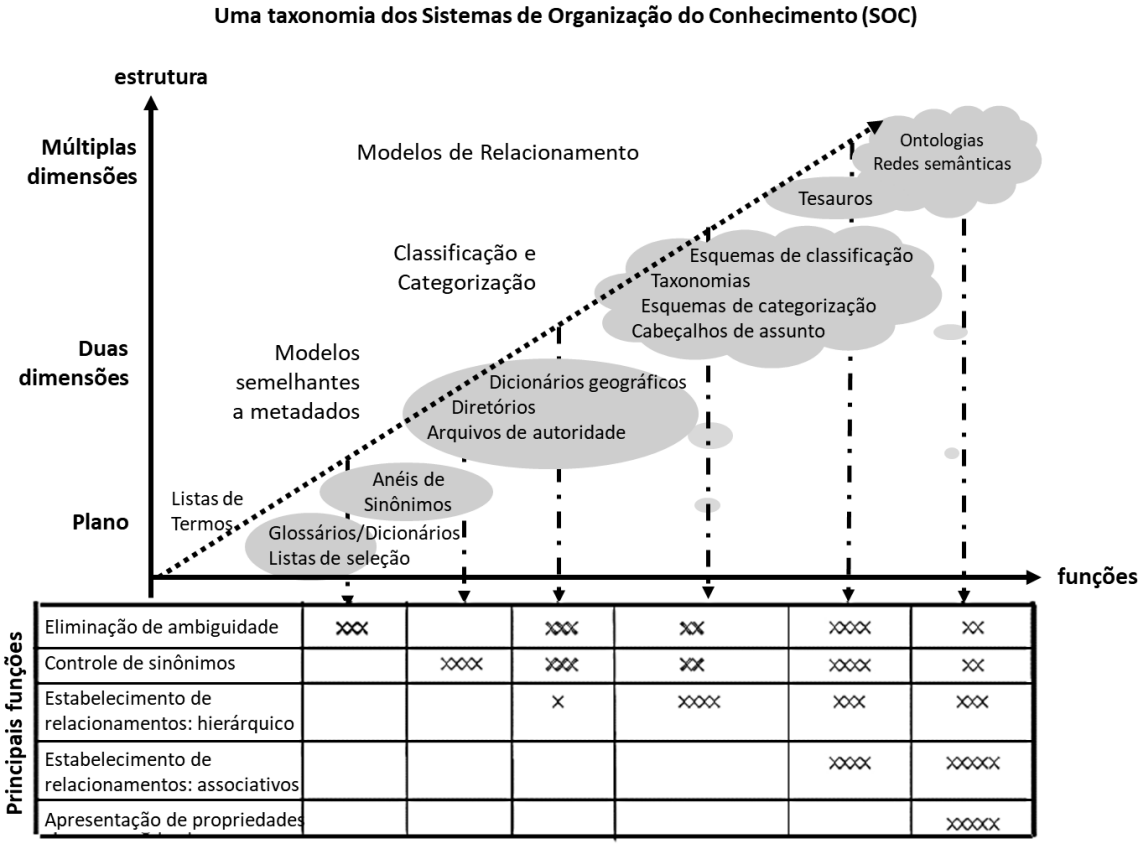
Schiessl e Shintaku (2012) descrevem a classificação dos SOC de acordo com a visão de Vickery (2008), que via na evolução dos sistemas de informação as mudanças de rumos nas funções exercidas pelos SOC em diferentes momentos. Segundo o autor, na era da pré-coordenação os SOC se voltavam ao atendimento das necessidades das bibliotecas físicas e seus sistemas manuais de classificação; na era da pós-coordenação eles evoluíram para atender às necessidade de maior dinamismo, permitindo a manipulação de seus elementos de maneira independente ao representar os assuntos de um documento; já na era da internet destacam-se os SOC do tipo taxonomia, que fornecem classificações hierárquicas que orientam a navegação do usuário pelos assuntos; e na era da *web* semântica os agentes inteligentes ganham força, com destaque para as ontologias.

Mazzocchi (2018) ratifica ao apontar que os variados tipos de SOC foram concebidos em diferentes momentos históricos e para distintas finalidades. Exemplos de SOC como ontologias, tesauros, taxonomias e glossários apresentam diferentes estruturas, funções, relações com a tecnologia e contextos de uso. Suas bases

teórico-metodológicas e sua utilidade vem sendo constantemente desafiadas pelo aumento progressivo da produção científica e cultural, culminando na reavaliação de sua capacidade de atendimento às novas necessidades de informação.

Marcia Zeng (2008) descreve os SOC de acordo com a complexidade de suas estruturas e por suas principais funções, subdividindo-os em grupos considerados por ela mutuamente excludentes como: modelos de relacionamento; classificação e categorização; modelos semelhantes a metadados; e listas de termos. A Figura 2 apresenta esta subdivisão na visão da autora, que se baseou na taxonomia adotada pelo *Networked Knowledge Organization Systems/Services/Structures* (NKOS), grupo dedicado à discussão dos modelos funcional e de dados que possibilitam sistemas e serviços de organização do conhecimento; na norma NISO Z39.19-2005 para vocabulários controlados monolíngues e num estudo de revisão ligado ao *Joint Information Systems Committee* (JISC), que analisa vocabulários de diferentes tipos, diretrizes de melhores práticas, pesquisas sobre serviços de terminologia e projetos relacionados ao tema. São evidenciadas na figura as estruturas dos SOC, que vão de planas a multidimensionais, e as principais funções por tipo de SOC, que são: a eliminação de ambiguidades; o controle de sinônimos ou equivalentes; o estabelecimento explícito de relações semânticas, sejam elas hierárquicas ou associativas; e a apresentação tanto de relações quanto de propriedades dos conceitos (Zeng, 2008).

Figura 2. Uma visão das estruturas e funções dos SOC



Fonte: tradução nossa de Zeng (2008, p. 161)

Zeng (2008) discorre sobre cada uma das funções dos SOC:

- **Eliminação de ambiguidade**: com exceção dos anéis de sinônimos, todos os demais SOC contém a função eliminar ambiguidade, o que significa que buscam definir termos, conceitos e rótulos claros e inequívocos, de modo a evitar situações onde seja possível mais de um significado ou interpretação;
- **Controle de sinônimos**: a função de controle de sinônimos ou equivalentes é um problema comum dos SOC, que resulta em dificuldades de recuperação de informações e diz respeito à garantia de que diferentes termos ou frases que significam a mesma coisa sejam usados e compreendidos de forma consistente dentro do sistema. Esse controle inclui nomenclaturas comuns e técnicas, mudanças no uso de termos ao longo do tempo, termos de diferentes idiomas, acrônimos e

grafias variantes. No âmbito de um vocabulário controlado os sinônimos podem ser considerados verdadeiros, isto é, palavras ou frases de mesmo significado, e equivalentes ou próximos, quando os significados são ligeiramente diferentes;

- **Estabelecimento de relacionamentos hierárquicos**: é a principal característica que distingue um SOC do tipo taxonomia ou tesauro de outras formas mais simples, como listas e anéis de sinônimos, e surge da necessidade de especificar, na maioria dos casos, termos genéricos e específicos;
- **Estabelecimento de relações associativas**: compreensão mais completa e diferenciada entre os conceitos num SOC, numa relação desse tipo há a conexão de conceitos que estão relacionados entre si de alguma forma, mas não em um relacionamento hierárquico ou equivalente.
- **Apresentação de propriedades**: se refere às características ou atributos de um conceito, como sua definição, notas de escopo e sinônimos, que conferem enriquecimento semântico ao sistema.

Há que se destacar também o papel da evolução da linguagem natural — que usamos no cotidiano para nos comunicar — à linguagem controlada, onde estão presentes funções como eliminação de ambiguidade e controle de sinônimos. A CI se utiliza tanto das linguagens convencionais quanto das controladas para representação e recuperação de informações (Schiessl; Shintaku, 2012). Segundo Currás (2008), um tipo de linguagem controlada é a linguagem documentária ou de indexação, que se volta aos sistemas de informação e bibliotecas com o propósito de sintetizar o conteúdo de documentos por meio de um vocabulário controlado. Os vocabulários são conjuntos de termos que descrevem em detalhes e com precisão o conteúdo de um objeto informacional, de modo a facilitar sua recuperação. Em resumo, vocabulários em qualquer nível são elemento fundamental de todo e qualquer tipo de SOC, conforme demonstrado na Figura 2, tornando-se condição essencial ao estabelecimento de relacionamentos semânticos em SOC de maior complexidade. Os vocabulários controlados tangibilizam a relação entre termos e conceitos.

Considerando os diferentes tipos de SOC, Carlan e Bräscher (2011) afirmam que na CI eles servem como ferramenta para transformação do conteúdo dos documentos originais e completos em um esquema estruturado sistematicamente que represente esse conteúdo. Segundo Zeng (2008) atuam como roteiros semânticos em serviços *web* ao modelar a estrutura subjacente de um domínio e fornecer significado, navegação e tradução, por meio de rótulos, definições, digitação, relacionamentos e propriedades para conceitos, padronizando a orientação para indexadores e futuros usuários, sejam eles humanos ou máquinas.

2.2 TAXONOMIAS

Do ponto de vista conceitual, as taxonomias apresentam uma diversidade de definições, características, origens e aplicações práticas que variam de acordo com o ambiente informacional em que são utilizadas (Lopes; Aganette; Maculan, 2020). De maneira geral, são compreendidas no campo da CI como sistemas lógicos de classificação e categorização, que possuem múltiplas aplicações e são utilizados principalmente em ambientes digitais para a organização e navegação de conteúdo, sendo compreensíveis tanto por humanos quanto por máquinas. Sua estrutura e processos apresentam importante potencial de contribuição para as estratégias de organização do conhecimento em diferentes contextos, especialmente nas instituições e no ambiente *web* (Campos e Gomes, 2007; Hedden, 2022; Hlava, 2014; Lambe, 2007; Lopes; Aganette; Maculan, 2020; Saumure e Shiri, 2008; Vignoli; Souto; Cervantes, 2013).

Quanto ao uso do termo taxonomia, a literatura apresenta pequenas variações. Segundo Aquino, Carlan e Brascher (2009) os termos taxonomia e taxionomia tem sua origem no grego *táxis* (ordem) e *onoma* (nome). Segundo Schiessl e Shintaku (2012), ambas as variações são aceitas na língua portuguesa e etimologicamente são compostas, além de *taxis* (ordem ou disposição), por *noma* (norma ou regra). Já para Hjørland (2017) o termo taxonomia origina-se de *taxis* (ordem ou arranjo) e *nomos* (lei ou ciência). De acordo com Lambe (2007, p. 4) a base para a compreensão do termo se encontra na obra lexicográfica greco-inglesa de Liddell e Scott, que define *nomos* (lei ou ordem) e *taxis* com os mesmos significados já apresentados. Para Currás (2008), Aganette, Alvarenga e Souza (2010) e Hlava (2014) deriva de *táxis* (ordem) e

nomos (lei ou norma). No entanto, a compreensão comum parece indicar as ideias de regras ou normas para ordem ou organização.

Schiessl e Shintaku (2012) apresentam notas históricas sobre o surgimento das taxonomias, iniciando seu percurso na Grécia de Aristóteles (384 a.C.-322 a.C.), onde se originaram como estrutura hierárquica de classificação dos seres vivos e das classes de palavras. Sobre a origem grega, Mendes e Bentes (2019) destacam que a importância da classificação revelada pelos filósofos gregos delineou a gênese para as taxonomias. “Entretanto, sabe-se que o fazer cientifico é um processo coletivo e a construção do conhecimento requer exames realísticos de sua existência” (2019, p. 37). Assim, Schiessl e Shintaku (2012) em seguida localizam no ensino de biologia, especificamente na botânica, a existência de um grupo de estudiosos dedicados ao tema, como Teofrasto (371 a.C.-287 a.C.) na Grécia; na Itália, Andrea Cesalpino (1519-1603); na Inglaterra, John Ray (1627-1705); e na França, Joseph Tournefort (1656-1708). Segundo os autores, foi a partir dos estudos do sueco Carl von Linné[10] (1707-1778) que surgiu a moderna classificação científica, também denominada sistemática, taxonomia biológica ou classificação biológica. Aganette, Alvarenga e Souza (2010) apontam que esse foi o primeiro uso do termo, datado de 1735, com a publicação da versão inicial da obra *Systema Naturae*. A classificação desenvolvida por Linné durante o século XVIII classificou e hierarquizou os seres vivos conforme suas características distintivas, ficando conhecida também como “Taxonomia de Lineu”. Outro uso é apontado por Hjørland (2017) e Currás (2008) como o primeiro, porém data de 1813, com o botânico suíço de família francesa Augustin-Pyramus de Candolle. Schiessl e Shintaku (2012) apontam ainda a taxonomia de Bloom, que surge em 1956, quase dois séculos depois propondo a classificação dos objetivos educacionais e, segundo os autores, popularizando o termo taxonomia. Segundo Aquino, Carlan e Brascher (2009) a origem dos sinônimos taxionomia e taxonomia estende-se da biologia e da botânica para a zoologia e para a ciência da informação, onde o termo taxonomia é o mais adotado. Em artigo sobre os marcos históricos da ciência da informação Lillian Alvares e Rogério Henrique de Araújo Júnior (2010, p. 202) sintetizaram cronologicamente os pioneiros, as obras clássicas e os eventos

[10] Em Português, Carlos Lineu, como é conhecido.

fundamentais da CI, citando a disciplina da taxonomia, localizada na biologia, como uma das que deram origem ao campo da CI. Além disso, por vezes os termos classificação e categorização também são utilizados como sinônimos de taxonomia nas diferentes áreas.

Na discussão da CI sobre taxonomias encontram-se comparações conceituais com a classificação e definições que as indicam como um tipo de vocabulário controlado (Aganette; Alvarenga; Souza, 2010; Hjørland, 2017). Isto provavelmente ocorre pela origem do termo na classificação dos seres vivos e pela explosão de intranets corporativas ocorrida em 1994, quando, segundo Hjørland (2017, n.p.), “muitos profissionais da informação começaram a usar seus próprios tesauros, esquemas de títulos de assunto e esquemas de classificação, renomeando-os como ‘taxonomia’”. Aganette, Alvarenga e Souza (2010, p. 80) apontam algumas diferenças em relação à classificação e apresentam o ponto de vista das taxonomias como vocabulários controlados.

Sobre a distinção entre taxonomias e classificação os autores apontam que “enquanto a classificação agrupa os termos utilizados, a taxonomia agrupa termos com interatividade” (Aganette; Alvarenga; Souza, 2010, p. 80). Por meio das definições de outros autores Hjørland (2017) complementa que o termo taxonomia é geralmente utilizado como sinônimo de sistemática na biologia e que envolve três aspectos, sendo a classificação um deles. Currás (2008), dedicada aos estudos do campo da documentação, ao conceituar taxonomia realiza o exercício prévio de distinção entre os termos ordenação e sistema de classificação, sobretudo classificação. A autora aponta a existência de inúmeros estudos sobre classificação e poucos sobre taxonomia, especialmente que não estejam ligados à biologia. À ordenação é chamado o processo físico que consiste em alocar cada objeto, mesmo abstrato, em um local onde possa ser encontrado. À classificação é conferido o sentido de método de ordenação, onde está implícito um processo de comparação para demarcar similaridades e diferenças que determinarão a posição dos elementos entre si. Já o sistema de classificação, na visão da autora se dá *a priori*, como um sistema pré-coordenado utilizado pelas teorias para classificar entes reais e abstratos. E a taxonomia é ao mesmo tempo sistema e método, parte importante do campo da documentação.

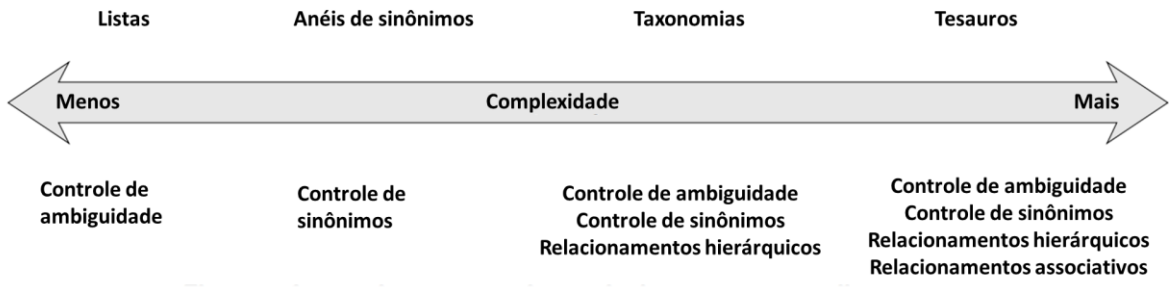
Já para analisar a visão de taxonomia na CI como vocabulário controlado Aganette, Alvarenga e Souza (2010) partem das definições apresentadas na norma ANSI/NISO Z39.19 (2005), revista em 2010. Afirmam que, segundo a norma, vocabulário controlado é uma lista de termos com definições claras e explícitas, onde há regras que garantem a não ambiguidade e não redundância dos termos. A definição de taxonomia aparece como:

> Um tipo de vocabulário controlado consistindo de termos preferenciais, ou ainda uma coleção de termos de vocabulário controlado organizados em uma estrutura hierárquica. Cada termo em uma taxonomia está em uma ou mais relações do tipo pai/filho (geral/específico) em relação a outro termo desta mesma taxonomia (National Information Standards Organization, 2010, p. 9)

A norma traz ainda um esquema que demonstra a complexidade das estruturas compreendidas como vocabulários controlados, Figura 3, onde figuram as taxonomias. As autoras analisam que o objetivo de um vocabulário controlado, conforme definição apresentada, é garantir consistência na descrição de objetos e do conteúdo, de modo a facilitar a recuperação da informação desejada. Hjørland (2017, n.p., tradução nossa) também aborda essa questão com base na mesma definição de taxonomia, porém com um posicionamento crítico:

> Essa maneira de usar a terminologia é confusa; em vez de tentar esclarecer os termos e buscar o consenso no uso de termos para determinados conceitos, faz o oposto (Hjørland, 2017, n.p., tradução nossa).

Figura 3. Crescimento estrutural da complexidade entre os tipos de vocabulário controlado



Fonte: tradução nossa de National Information Standards Organization, 2005

No campo da CI as taxonomias são classificadas como sistemas de organização do conhecimento, compartilhando dos objetivos de um vocabulário controlado, como controle de ambiguidades e sinônimos, relacionamentos hierárquicos e associativos e

apresentação de propriedades. No entanto, os objetivos de uma taxonomia parecem extrapolar os objetivos de um vocabulário. Enquanto as taxonomias se concentram em organizar conceitos em uma estrutura hierárquica e classificatória, que pode apresentar diferentes formas de representação, o vocabulário controlado se concentra apenas no estabelecimento de relações precisas entre termos e conceitos específicos. Esse controle terminológico é apresentado por Aquino, Carlan e Brascher (2009) como um dos princípios classificatórios a serem seguidos para a construção de taxonomias eficientes, podendo também ser entendido como um critério de qualidade.

Sobre as formas que uma taxonomia pode assumir Lambe (2007) aponta a hierárquica como a mais tradicional, além da representação no modelo de lista, árvores, sistemas de mapas, entre outras. Além disso, enquanto taxonomias de diferentes tipos parecem capazes de lidar com grandes quantidades de dados, o vocabulário controlado parece constituir-se como um dos elementos inerentes ao desenvolvimento de uma taxonomia, garantindo a precisão e a especificidade necessárias ao objetivo maior da organização do conhecimento que é a recuperação da informação. O Quadro 1 apresenta as compreensões de diferentes autores sobre o conceito de taxonomia.

Quadro 1. Compreensões do conceito de taxonomia na ciência da informação

Autor	Compreensão
Campos e Gomes (2007, p. 1)	“Taxonomia é, por definição, classificação, sistemática e está sendo conceituada no âmbito da Ciência da Informação como ferramenta de organização intelectual. É empregada em portais institucionais, bibliotecas digitais como um novo mecanismo de consulta, ao lado de ferramentas de busca. Além destas aplicações, a taxonomia é um dos componentes em ontologias A organização das informações através do conceito de Taxonomia permite alocar, recuperar e comunicar informações dentro de um sistema de maneira lógica através de navegação”.

Autor	Compreensão
Lambe(2007, p. 5-8; 123, tradução nossa)	“Uma taxonomia é uma forma de esquema de classificação” (p. 5) “Uma taxonomia é uma espécie de mapa do conhecimento [...] Estas características de uma taxonomia (classificação, semântica, um mapa de um domínio de conhecimento que pode ser apreendido como um todo) conferem-lhe mais do que apenas poder de navegação. Uma taxonomia também atua como um dispositivo de memória artificial” (p. 8) “As taxonomias são ferramentas para a eficácia [*organizacional*]” (p. 123) “Em outras palavras, um tesauro é simplesmente a sua taxonomia em formato de dicionário. Por outro lado, uma taxonomia é um dicionário de sinônimos com todos os rótulos organizados por assunto. Uma taxonomia representa ou mapeia visualmente o assunto como um todo, enquanto um tesauro explica os tópicos um por um” (p. 6–7)

Autor	Compreensão
Zeng (2008, p. 161; 169, tradução nossa)	"divisões de itens em grupos ordenados ou categorias com base em características particulares" (p. 161). "Uma taxonomia é um tipo de SOC que consiste em termos preferenciais, todos conectados em uma hierarquia ou polihierarquia [...] Hoje, o termo taxonomia é aplicado em um sentido mais amplo e geral e agora pode se referir à classificação das coisas, bem como aos princípios subjacentes a tal classificação [...] Na construção de taxonomias, tanto os aspectos científicos da categorização quanto os aspectos cognitivos da categorização precisam ser levados em consideração [...] Atualmente, a abordagem da taxonomia está sendo aplicada a muitos domínios e disciplinas. Com ou sem notações, essas estruturas têm princípios classificatórios totalmente empregados e relações hierárquicas para representar o conhecimento de um domínio. Alguns SOC estão tentando fornecer uma organização taxonômica de alto nível da qual muitos esforços podem se beneficiar" (p. 169).
Aganette, Alvarenga e Souza (2010, p. 74)	“é usada para estruturação de informações e considerada como uma importante ferramenta para entendimento de como uma área de conhecimento é organizada e, principalmente, como essa área se relaciona e interage com outras”.
National Information Standards Organization, (2010, p. 30)	“Uma taxonomia é um vocabulário controlado que consiste em termos preferenciais, todos conectados em uma hierarquia ou polihierarquia”.

Autor	Compreensão
Schiessl e Shintaku (2012, p. 82–83)	“taxonomia apresenta-se como ciência, processo e produto: ciência por incluir conceitos que sistematizam o processo, embasando-o teoricamente; processo por indicar as etapas a serem cumpridas e, finalmente, produto por denominar uma estrutura representativa do conhecimento. A amplitude das taxonomias revela a interdisciplinaridade desse tópico, que possui aplicações em vários domínios”
Vignoli, Souto e Cervantes (2013, p. 68)	“ontologias e taxonomias são ferramentas para organizar e representar o conhecimento em SOC que visam ao usuário final, em sua busca e recuperação da informação. [...] As taxonomias podem ser representadas na Arquivologia e na Biblioteconomia por meio da estruturação de termos com intuito de classificar”.
International Organization for Standardization (2013, n.p.)	“esquema de categorias [...] e subcategorias que podem ser usadas para classificar e organizar itens de conhecimento ou informação Nota 1 de entrada: As taxonomias variam do muito simples ao muito complexo. Nas taxonomias mais simples as categorias não estão necessariamente divididas em subcategorias, enquanto nas complexas podem ser encontrados múltiplos níveis hierárquicos. Outros recursos também podem estar presentes, como todos os recursos do dicionário de sinônimos na ISO 25964-1 e/ou os recursos comumente encontrados em esquemas de classificação. Fora desta Norma Internacional, o termo é frequentemente usado de forma vaga para se referir a qualquer tipo de vocabulário estruturado”.

Autor	Compreensão
Hlava (2014, p. 11; 128)	“As taxonomias são a base para rotular informações para armazenamento, para pesquisa e recuperação precisa, para os rótulos usados em um *site* e para navegação no site. Os termos de taxonomia usados como metadados do assunto suportam 100% de recuperação de informações relevantes” (p.11). “Conforma definida em [*glossário*] ANSI/NISO Z39.19-2010R, as taxonomias são a base para rotular informações para armazenamento, para pesquisa e recuperação precisa, para os rótulos usados em um *site* e para navegação no site. Os termos de taxonomia usados como metadados do assunto suportam 100% de recuperação de informações relevantes” (p. 128).
Mazzocchi (2018, n.p., tradução nossa)	“qualquer agrupamento ordenado de itens com base em características particulares [...] Não está claro o que constitui uma → taxonomia, ou qual é o significado do termo taxonomia. Gilchrist (2003) apontou que tal termo é usado com pelo menos cinco significados diversos (mas sobrepostos) para se referir a (i) diretórios da *web*, (ii) taxonomias para dar suporte à indexação automática, (iii) taxonomias criadas por categorização automática, (iv) filtros front end, e (v) taxonomias corporativas. É sintomático que, como argumentam Souza et al. (2012), quando comparadas com tesauros, as taxonomias podem ser vistas alternativamente como menos (Daconta et al. 2005; Guarino 2006; Obrst 2004) ou mais (Bergman 2007; Smith; Welty 2001) estruturadas”.

Autor	Compreensão
Mendes e Bentes (2019, p. 38; 44)	“entende-se que a taxonomia pode ser incorporada em todo o domínio da atividade humana [...] Com o avanço do sistema *web* a taxonomia ganha maior importância para a representação, organização e recuperação da informação, particularmente no ambiente *web* semântica” (2019, p. 38) “seus princípios permitem estabelecer padrões de classificação e ordenação da informação, versáteis, adaptando-se conforme a evolução do conteúdo e identificando-se as facetas de: suporte, ferramenta, método, serviços, instrução e mapeamento, a depender do objeto de estudo na área da Biblioteconomia e da Ciência da Informação” (2019, p. 44)
Lopes, Aganette e Maculan (2020, p. 162)	“Pode-se afirmar que uma taxonomia é um modo de organizar, classificar e agrupar coisas, por meio de atributos específicos, em seguida, atribuir nomes a esses grupos de coisas”.
Hedden (2022, p. 24; 584, tradução nossa)	“1) um sistema hierarquicamente estruturado de organização de conceitos por seus rótulos preferenciais. 2) qualquer vocabulário controlado, seja ele hierárquico ou não, que envolva o controle de nomes de conceitos” (p. 584) “taxonomias são definidas como qualquer sistema de organização do conhecimento (vocabulário controlado, anéis de sinônimos, tesauros, árvore hierárquica de termos, ou ontologias) usada para suporte à informação/localização de conteúdo, descoberta e acesso” (p. 24)

Fonte: elaborado pelo autor (2023)

A ideias apresentadas nas diferentes definições do Quadro 1 destacam as taxonomias como técnica utilizada para organizar e classificar informações de uma

determinada área do conhecimento, estabelecendo padrões de ordenação e agrupamento. As taxonomias aparecem como ferramentas fundamentais na busca e recuperação de informação, sendo aplicadas em diversas áreas como biblioteconomia, arquivologia, ciência da informação e também na tecnologia da informação, especificamente no desenvolvimento de software. Destaca-se a mais recente compreensão de Hedden (2022), provavelmente pautada pela já referida multiplicidade de características e aplicações das taxonomias.

Nos ambientes digitais o desenvolvimento das taxonomias se direcionou para as formas automatizadas de criação da informação, o que motivou o enfoque de estudos da CI nesta temática (Aganette; Alvarenga; Souza, 2010; Aquino; Carlan; Brascher, 2009; Lopes; Aganette; Maculan, 2020; Schiessl e Shintaku, 2012). Apesar disso, a produção científica encontrada no campo está atualmente escassa, conforme aponta o estudo recente de Lopes, Aganette e Maculan (2020) sobre usos e aplicações de dois tipos de taxonomias na ciência da informação no Brasil. Foi realizado levantamento bibliográfico que recuperou apenas sete dissertações e nenhuma tese com esta temática no campo num período de 15 anos. Um dos resultados foi a existência de uma lacuna no conhecimento sobre o estado da arte dos conceitos, usos e aplicações das taxonomias corporativas e facetadas, compreendidas pelos autores como os tipos de taxonomias mais enfocadas na CI. Os achados evidenciam a necessidade de desenvolvimento do tema taxonomias no contexto da CI.

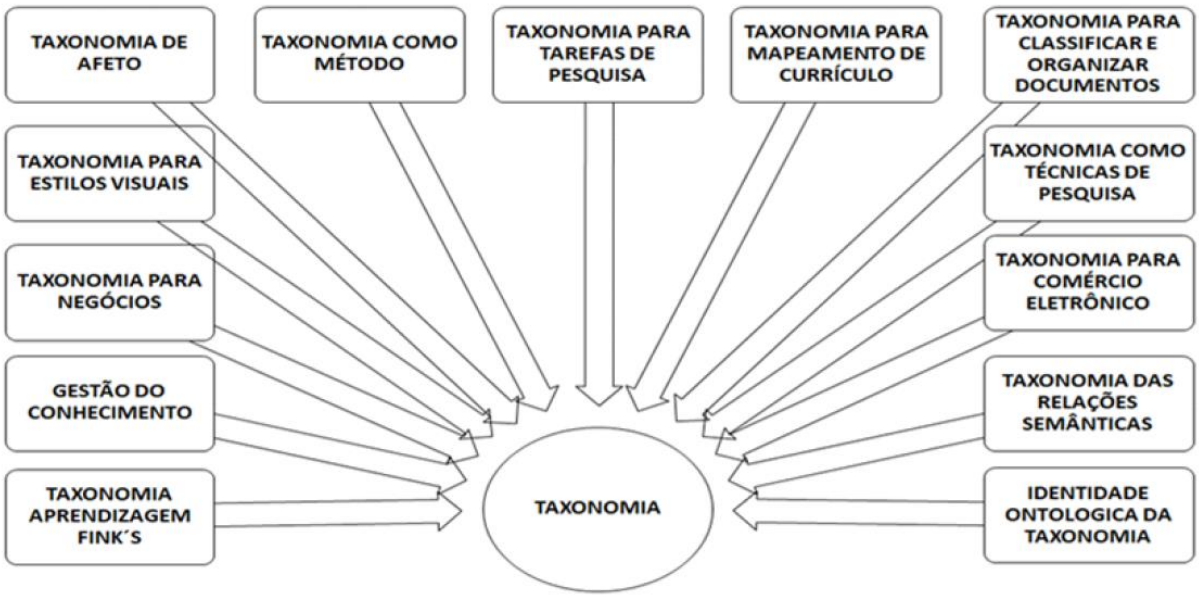
Segundo Campos e Gomes (2007), no contexto da CI as taxonomias se caracterizam por:

- conter uma lista estruturada de conceitos/termos de um domínio;
- incluir termos sem definição, somente com relações hierárquicas (gênero-espécie e parte-todo);
- possibilitar a organização e a recuperação de informação através de navegação;
- permitir a agregação de dados;
- evidenciar um modelo conceitual do domínio;
- ser um instrumento de organização intelectual;

- ser um mecanismo de consulta por navegação;
- ser produto de um projeto conduzido por especialistas, com o objetivo de representar um domínio em determinado contexto;
- conter princípios classificatórios explícitos e consistentes, fruto de gestão centralizada do sistema de classificação;
- ser um instrumento que permite evitar ambiguidades.

Estudo de revisão sistemática de Mendes e Bentes (2019), que analisou a presença da taxonomia na literatura cientifica das áreas da biblioteconomia e da ciência da informação, apontou as várias facetas da taxonomia presentes nos dois campos científicos, conforme demonstrado na Figura 4.

Figura 4. As várias facetas da taxonomia presente na biblioteconomia e ciência da informação



Fonte: Mendes e Bentes (2019, p. 42)

Um resumo dos achados de Mendes e Bentes (2019) aponta usos e aplicações das taxonomias em diferentes contextos: o estudo de Banks e Henderson (2019) apresentou uma taxonomia de aprendizagem significativa denominada Fink como uma estratégia para que os alunos soubessem como acessar o *site* da biblioteca e identificar tipos de fontes; Abukhader (2019) debruçou-se sobre as taxonomias na gestão do conhecimento; Tucker et al. (2018) relataram um projeto de construção taxonômica da estrutura acadêmica do departamento de ciência da informação; Ullah

e Ameen (2018) abordaram taxonomia no sentido de vocabulário controlado; Steimer e Luz (2016) elaboraram uma taxonomia de e-commerce para classificar informações em ambiente digital; Maia, Lima e Maculan (2017) compilaram os tipos de relações semânticas em uma única taxonomia para evitar ambiguidades na língua portuguesa; Simões et al. (2016) destacaram a utilidade da taxonomia como ferramenta ontológica na organização do conhecimento no meio digital e para controle da informação, especialmente ao nível terminológico, contribuindo para a recuperação da informação.

Campos e Gomes (2007) já corroboravam os achados 12 anos antes da publicação do estudo, quando indicavam que as taxonomias, conforme conceito da CI, podem ser aplicadas para diferentes finalidades e são comumente encontradas em portais institucionais, nas bibliotecas digitais, na *web* semântica, nas ontologias e na gestão da informação e do conhecimento. As autoras sinalizam, porém, que as taxonomias não são neutras, tendo seu recorte determinado pelas características de quem a detém e seus objetivos. Afirmam que: “toda taxonomia é fruto de um processo representacional e classificatório e [...] produto de uma construção que representa o estado e visão do conhecimento de seus elaboradores” (2007, p. 10). Estudiosos da temática das taxonomias como Patrick Lambe, Heather Hedden e Marjorie Hlava corroboram em seus trabalhos as ideias colocadas pelas autoras.

Campos e Gomes (2007, p. 10) apresentam também três tipos de taxonomias, sendo as últimas duas adotadas no campo da CI e de natureza policotômica: (a) taxonomia canônica de unidades sistemáticas ou taxonomia científica, originária do campo da biologia, comumente utilizada na botânica e na zoologia, onde não é permitida a agregação de novos assuntos depois da taxonomia construída; (b) taxonomia de domínio, que pela complexidade de sua estrutura requer um primeiro recorte por categorias e no interior destas várias divisões e subdivisões são possíveis (policotômica); e (c) taxonomia corporativa, que contribuem ao “reconhecer e relacionar atividades agregadoras de valor, diminuindo esforços na produção e utilização do conhecimento [nas organizações]” (2007, p. 3).

Ainda no âmbito corporativo, a obra "*Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness*" de Patrick Lambe (2007) destaca a importância do desenvolvimento de taxonomias para a organização e gestão do conhecimento, especialmente no que diz respeito à efetividade organizacional.

Sinaliza que uma taxonomia somente se mantém relevante, útil e atual através de processos bem estruturados de manutenção, uma vez que são estruturas vivas, que evoluem junto com o conhecimento e a organização.

Aquino, Carlan e Brascher (2009) destacam a importância da adoção de princípios classificatórios explícitos e consistentes para a elaboração de taxonomias. Nesse sentido, algumas etapas para a construção de taxonomias no campo da CI são apresentadas com base em Campos e Gomes (2007) e Lambe (2007):

- considerações pré-taxonomia: que problema está tentando responder, tipo e alcance da informação, volume de conteúdo, disponibilidade dos especialistas, arquitetura da informação e informática.
- captura do conhecimento, onde são relevantes o levantamento de assuntos e a definição de uma metodologia de entrevista com especialistas da área onde a taxonomia será aplicada;
- análise dos documentos/informações que serão agregados à taxonomia;
- definição da tipologia ou estrutura classificatória, que deve estar em conformidade com os princípios classificatórios de categorização, diferenciação entre termos e conceitos, hierarquias e multidimensionalidade; e
- mecanismos de validação da taxonomia desenvolvida.

Quanto às vantages de uma taxonomia Campos e Gomes (2007) adotam uma visão ampla, que pode estar associada aos diversos ambientes onde se aplicam taxonomias. São elas:

- permitir a visualização do conteúdo de um sistema de forma integrada;
- minimizar o tempo gasto na busca;
- padronizar a ordem de arquivamento das informações;
- padronizar a terminologia utilizada;
- visualizar a relação entre os assuntos/documentos;
- apresentar a informação acoplada à estrutura classificatória;
- permitir que a estrutura hierárquica da taxonomia revele informações úteis para

a gestão;

- ter uma estrutura classificatória voltada exclusivamente para as necessidades de seus usuários, bem como para o objetivo da Organização;
- possibilitar uma interface amigável ao usuário;
- permitir o gerenciamento e a difusão dos fluxos de trabalho (no caso da taxonomia de tarefa);
- promover a interoperabilidade entre sistemas quando fazem parte de uma ontologia;
- propiciar que o próprio usuário exerça o papel de classificador, enquanto o bibliotecário assume o papel de classificacionista;
- oferecer ao usuário um guia ao assunto a ser investigado, encorajando o pensamento associativo e como consequência a descoberta de informações.

Lambe (2007) complementa a visão das autoras exemplificando a importância do uso de taxonomias na pronta resposta mais eficiente a desastres naturais e acidentes, que a depender da abrangência podem demandar a articulação de diferentes atores governamentais, como postos policiais, bombeiros, hospitais e outros departamentos. O principal destaque se encontra no fato de que se houvesse uma taxonomia de apoio a estas instituições, haveria maior potencial articulador, aumentando a eficiência da pronta resposta, que poderia ser mais rápida e assertiva. O autor aponta que o atendimento da população nos desastres geralmente mobiliza questões sociais que pressupõem registros confiáveis nos diferentes sistemas utilizados pela polícia, pelos hospitais e outros agentes governamentais. O fato de não existir uma taxonomia de base significa que não há padronização de linguagem, não há interoperabilidade entre sistemas, não há entendimento claro entre as partes, não há comunicação eficiente, o que impacta diretamente na eficiência do atendimento. O exemplo citado pelo autor é alusivo a uma série de outras situações em diferentes contextos de aplicação, no sistema público e no sistema privado, dentro e fora das organizações.

Como principais desafios enfrentados na construção de taxonomias, especialmente às voltadas ao ambiente organizacional, Lambe (2007) destaca a

dificuldade em equilibrar o rigor estrutural com a flexibilidade exigida pelo cotidiano, especialmente em ambientes de intenso volume informacional, onde surgem a todo momento novas informações e conceitos.

Por fim, cabe destacar os resultados da revisão sistemática de Mendes e Bentes (2019) que encontrou as características de multifuncionalidade e abrangência nas taxonomias presentes na biblioteconomia e na ciência da informação, com amplo potencial de aplicabilidade em diferentes domínios do conhecimento. Apontam que seus princípios possibilitam o estabelecimento de padrões versáteis de classificação e ordenação da informação, adaptando-se conforme a evolução do conteúdo.

2.2.1 Taxonomias e suas relações

Com vistas a esclarecer brevemente interseções guardadas pelas taxonomias com outros conceitos fundamentais para a organização e representação do conhecimento, comumente observadas na literatura, abaixo são apresentadas as relações entre taxonomias e: tesauros, metadados, ontologias e folksonomias.

2.2.1.1 Com tesauros

Para muitos autores os tesauros podem ser definidos como taxonomias com funções adicionais. Trata-se de um vocabulário controlado altamente estruturado que envolve termos preferenciais, que representam conceitos mais amplos ou mais restritos; termos não preferenciais, que diz respeito aos termos que não estão atribuídos a documentos, mas são pontos de entrada em um tesauro; relações hierárquicas; relações associativas; e notas de escopo sempre que necessárias. Os tesauros contam com o maior grau de consenso em termos de sua definição e para recuperação da informação possuem especificações padrão, como a ISO 25964 e versões, que abordam o desenvolvimento e a manutenção de tesauros, além de sua interoperabilidade com outros vocabulários (Dextre Clarke, 2019; Hedden, 2022; Hlava, 2014).

O significado tradicional de tesauro é um tipo de dicionário, em que cada entrada possui sinônimos, expressões alternativas e antônimos. O tesauro para recuperação da informação compartilha dessas características, conforme apontado, porém com a abrangência delimitada pelo seu domínio de cobertura. Assim, é exigido que um tesauro para recuperação da informação especifique quais termos poderão ser

utilizados como sinônimos, quais são considerados termos próximos ou restritos, quais são conceitos amplos e quais se constituem como termos relacionados (Hedden, 2022).

Para Lambe (2007, p. 6–7, tradução nossa), “um tesauro é simplesmente a sua taxonomia em formato de dicionário. Por outro lado, uma taxonomia é um dicionário de sinônimos com todos os rótulos organizados por assunto”. O autor considera que a principal diferença entre tesauros e taxonomias está nas funções complementares que exercem, de modo que enquanto uma taxonomia mapeia o assunto como um todo, um tesauro define os tópicos um por um. De modo geral, os tesauros são projetados tanto para indexação quanto para pesquisa, com foco no conteúdo semântico, enquanto as taxonomias podem servir a propósitos de categorização mais amplos. Além disso, o desenvolvimento de taxonomias nem sempre segue os padrões e diretrizes estabelecidos para os tesauros (Dextre Clarke, 2019; Lambe, 2007).

2.2.1.2 Com metadados

Os metadados ou dados estruturados, são amplamente definidos como dados sobre dados e desempenham um papel fundamental na descrição de objetos no ambiente *web*. Eles fornecem informações de contexto abrangentes e valiosas para a compreensão das informações por máquina, o que impacta fortemente a interoperabilidade entre os recursos na Internet. Quanto mais detalhados ou enriquecidos são os metadados, maior será a capacidade de compreensão dos recursos informacionais, sua origem, proveniência, propósito, autoria, história, licenças e detalhes técnicos (Baptista, 2007; Hedden, 2022; Rocha; Sales; Sayão, 2017)

Há diferentes tipos de metadados: (1) descritivos, quando se referem à descrição de um recurso para fins de sua descoberta e identificação, incluindo elementos como título, resumo, autor e palavras-chave; (2) estruturais, que indicam como os objetos compostos são montados e ordenados; (3) administrativos, que dizem respeito à gestão do ciclo de vida do recurso; e (4) de preservação, que contêm informações necessárias para arquivar e preservar um recurso. Este último tipo compreendido por alguns autores como um subconjunto do tipo administrativo (Baptista, 2007; Hedden, 2022; Hlava, 2014; Rocha; Sales; Sayão, 2017).

Plataformas de conteúdo e repositórios geralmente adotam esquemas de metadados personalizados ou padrão, como Dublin Core ou MARC, com o objetivo de facilitar o compartilhamento e a interoperabilidade entre diferentes conjuntos de dados e instituições. Esses esquemas compreendem o conjunto de elementos ou campos de metadados definidos para seu domínio de aplicação e as regras necessárias para sua completa descrição (Hedden, 2022).

Segundo Hlava (2014) e Hedden (2022), os termos de uma taxonomia se constituem como metadados de assunto e por isso comportam a totalidade das informações relevantes recuperadas em um sistema. Quando uma taxonomia é implementada, ela se torna um meio de utilizar metadados para descrever um objeto, uma vez que sua hierarquia de conceitos e outros elementos de metadados, apesar de não constituírem a informação em si, proporcionam descrições valiosas.

2.2.1.3 Com ontologias

Ontologias são estruturas formais de representação do conhecimento, que definem conceitos e suas interrelações em profundidade, permitindo a compreensão compartilhada e a interoperabilidade de dados. Constituem-se como uma das principais tecnologias da *web* semântica, onde também especificam regras lógicas para raciocinar sobre os dados. A taxonomia é um dos componentes das ontologias, a estrutura hierárquica base para as primeiras camadas de relacionamento entre conceitos. Pode-se dizer de maneira ampla que as ontologias se concentram nas relações e regras semânticas dentro de um domínio, enquanto as taxonomias priorizam uma categorização estruturada das informações, isto é, fornecem um sistema de classificação hierárquico, organizando termos do geral ao específico para facilitar a recuperação de informações (Mazzocchi, 2018; Vignoli; Souto; Cervantes, 2013; Zeng, 2008).

As ontologias abrangem uma estrutura conceitual mais ampla. Apesar disso, “nos negócios e na indústria são cada vez mais criadas como uma extensão das taxonomias existentes, permitindo assim que as taxonomias sirvam a mais propósitos” (Hedden, 2022, p. 20). Mesmo distintas, ambas as abordagens fornecem estrutura, são instrumento de classificação, modelam um domínio e representam conceitos e relações, atuam como vocabulário controlado, estabelecem conjuntos de termos e

contribuem para a semântica em algum nível (Vignoli; Souto; Cervantes, 2013).

2.2.1.4 Com folksonomias

O termo folksonomia vem do inglês *folksonomy*, onde “folk” significa povo e “sonomy” advém do termo taxonomia, formando seu conceito, surgido no contexto da *web* 2.0, abordada adiante. Trata-se do produto da livre atribuição de palavras-chave por usuários da *web* para organização e recuperação de informações em websites, mídias sociais e outros tipos de plataformas baseadas em conteúdo. A este processo denomina-se etiquetagem, do inglês *tagging*, ou marcação (Hedden, 2022; Lopes eAlbuquerque, 2022).

Segundo Lopes e Albuquerque (2022), as folksonomias permitem a organização do conhecimento de determinado domínio sob o ponto de vista dos usuários, o que apresenta riscos e potencialidades. Dentre os principais riscos estão a falta de padronização e inconsistências promovidas pela subjetividade inerente ao seu processo de construção, o que também corresponde a inevitáveis vieses; difícil escalabilidade; e dependência de uma massa de colaboradores para que seja efetiva (Hedden, 2022). Por outro lado, como uma ferramenta criada de maneira colaborativa, enfatiza a liberdade e os aspectos socioculturais, ajudando a suprir lacunas encontradas em vocabulários controlados. Além disso, podem representar formas mais simples e menos custosas de categorização; contribuem para a identificação de tendências e interesses; quando mais amplas e com volume de participação podem funcionar como uma forma de inteligência coletiva; podem auxiliar na formação de comunidades; e podem contribuir com SOC no ambiente *web* enriquecendo-os. Os autores afirmam ainda, que os problemas inerentes às folksonomias podem ser mitigados pela associação de um ou mais vocabulários controlados como taxonomias (Hedden, 2022; Lopes eAlbuquerque, 2022).

2.3 WEB SEMÂNTICA

Ao longo de 39 anos foram desenvolvidas as tecnologias que possibilitariam o desenho da *World Wide Web* (WWW) ou simplesmente *web*. Alguns dos principais eventos do que podemos chamar de pré-*web* merecem destaque. A história inicia, segundo o W3C, em 1945, com o fim da Segunda Guerra Mundial, quando Vannevar

Bush (1890-1974) propõe no artigo denominado "*As We May Think*" o desenvolvimento do Memex, um computador analógico que permitia ao usuário guardar e recuperar documentos armazenados em microfilmes (ASIS&T, 2023; W3C, 2020). Vinte anos mais tarde surgiria o termo hipertexto, cunhado por Ted Nelson na 20ª Conferência Nacional da área de computação, ocorrida em Nova Iorque. O hipertexto talvez seja o conceito-chave mais importante para o modelo de *web* e sua evolução até a *web* 3.0 ou *web* semântica, como conhecemos atualmente. Mais quinze anos à frente, já com noções de sistemas online disseminadas, pesquisas em redes da ARPANET[11] expandidas e os protocolos IP[12] e TCP[13] criados, Tim Berners-Lee escreve em 1980 um programa que já utilizava as noções de hipertextos. Em 1984, Paul Mockapetris introduz o *Domain Name System* (DNS), o sistema de nomes utilizado na Internet. E em 1989 Tim Berners-Lee descreve a primeira concepção de *web* por meio de uma solução de gestão da informação endereçada ao CERN[14] (W3C, 2020). Cerca de dois anos mais tarde, em agosto de 1991, o software *web* é disponibilizado na Internet por meio do protocolo FTP[15]. Um evento ainda mais importante mundialmente ocorreu em abril de 1993, quando o CERN disponibilizou gratuita e livremente a tecnologia WWW à sociedade (W3C, 2020).

Conforme sinalizam Souza e Alvarenga (2004, p. 132):

> Surgida no início dos anos 90, a World Wide Web*, ou simplesmente Web, é hoje tão popular e ubíqua, que, não raro, no imaginário dos usuários, confunde-se com a própria e balzaquiana Internet – a infraestrutura de redes, servidores e canais de comunicação que lhe dá sustentação.

Com vista à melhor compreensão do conceito de *web*, cabe sua breve distinção em relação à Internet, que se constitui como a infraestrutura por meio da qual a conexão entre computadores é realizada globalmente. Enquanto a *web* é formada por

---

[11] A *Advanced Research Projects Agency Network* (ARPANET) foi a primeira rede de computadores, construída para fins militares.
[12] O Internet Protocol (IP) é um endereço de rede exclusivo, que identifica um dispositivo dentro de um rede de computadores pública ou privada.
[13] O *Transmission Control Protocol* (TCP) é um protocolo de comunicação essencial para o funcionamento da Internet, uma vez que controla o envio e o recebimento de dados na *web*.
[14] CERN é o acrônimo estabelecido a partir do francês *Conseil Européen pour la Recherche Nucléaire* para a Organização Europeia de Pesquisa Nuclear.
[15] *File Transfer Protocol* (FTP) é um protocolo de controle de transferência de arquivos entre diferentes sistemas na *web*. Ele permite, por exemplo, que os usuários façam o upload (envio) e o download (recebimento) de arquivos através da Internet.

páginas – que podem conter textos, imagens, vídeos e outros tipos de conteúdo – interligadas por links (Berners-Lee, 1990; Leiner et al., 2009). Segundo Leiner et al. (2009), a Internet foi criada como suporte a uma gama de diferentes funções, que vão desde o compartilhamento de arquivos ao acesso remoto para compartilhamento de recursos.

> Um conceito-chave da Internet é que ela não foi projetada para apenas um aplicativo, mas como uma infraestrutura geral na qual novos aplicativos podem ser concebidos, conforme ilustrado posteriormente pelo surgimento da *World Wide Web* (Leiner et al., 2009, p. 25, tradução nossa).

Dessa forma a *web* se constitui como um serviço dentro da Internet, que permite a visualização e o acesso de conteúdo por meio de navegadores. O surgimento e a evolução da *web* demonstram a flexibilidade e robustez do projeto da Internet ao mesmo tempo em que transformam a comunicação, a criação e a colaboração global ao longo do tempo. Sobre o desenvolvimento da *web* de sua primeira versão até a *web* semântica, Souza e Alvarenga (2004, p. 132) afirmam:

> buscava oferecer interfaces mais amigáveis e intuitivas para a organização e o acesso ao crescente repositório de documentos que se tornava a Internet. Entretanto, o enorme crescimento – além das expectativas – do alcance e tamanho desta rede, além da ampliação das possibilidades de utilização, fazem com que seja necessária uma nova filosofia, com suas tecnologias subjacentes, além da ampliação da infraestrutura tecnológica de comunicação (Souza eAlvarenga, 2004, p. 132).

2.3.1 A web 1.0

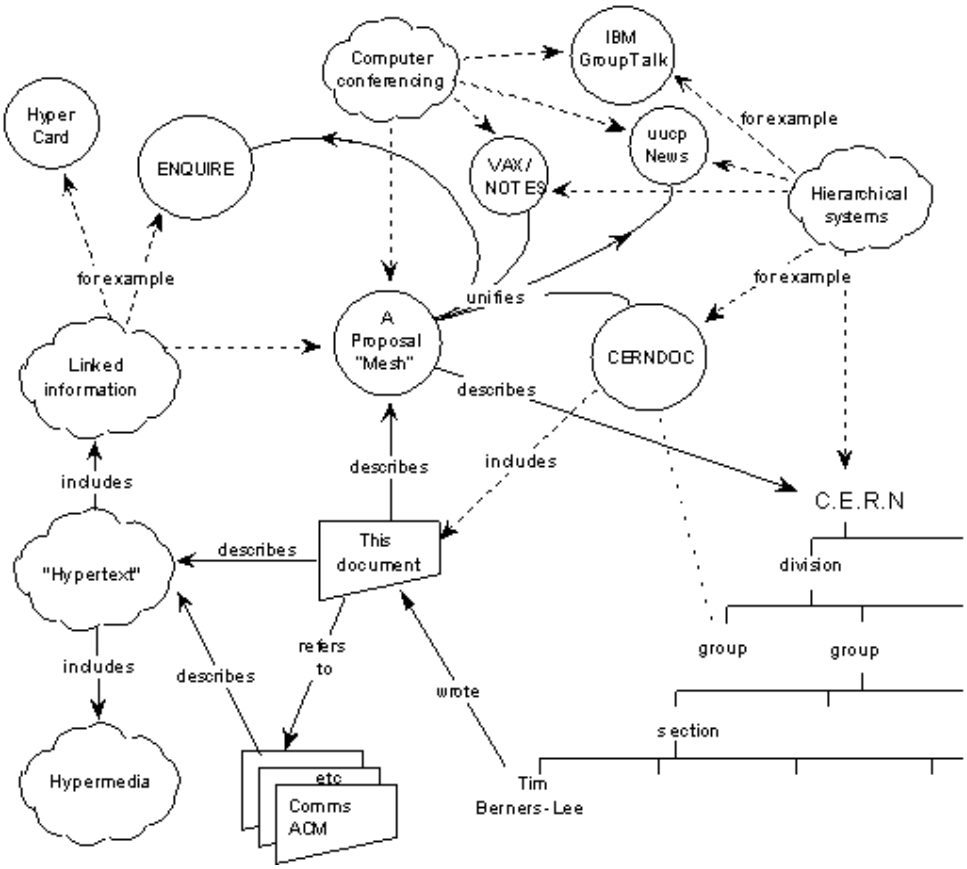
A criação da *web* tem origem em um documento e um script escritos por Tim Berners-Lee em março de 1989 e maio de 1990, respectivamente. O documento, entitulado "*Information Management: A Proposal"* (Gestão da Informação: uma proposta) foi endereçado à administração do CERN numa tentativa de persuadi-los a implantar um sistema global de hipertexto como solução para o gerenciamento de projetos complexos da instituição. A proposta, inicialmente denominada como "Mesh", apresentava uma visão baseada na conexão de informações de diferentes áreas por meio de um sistema de hipertexto distribuído. O objetivo principal era reduzir perda de informação dentro da cadeia organizacional, uma vez que o CERN era visto pelo autor como um modelo em miniatura daquilo que se tornaria o mundo poucos anos à frente.

Assim, o documento discutia tanto questões gerenciais relacionadas a pessoal e infraestrutura, quanto acesso à informação, integração entre sistemas de informação e colaboração em rede, o que incluía a definição de papéis individuais e coletivos na criação e gestão do que mais tarde se tornaria a *web* (Berners-Lee, 1990).

Conforme Figura 5, o detalhamento da proposta original já incluía os principais elementos constituintes da posterior *web* semântica, como a descrição do que seria chamado *Resource Description Framework* (RDF) e a noção de *Linked Data* (LD) (Berners-Lee, 1990).

> O sistema de que precisamos é como um diagrama de círculos e setas, onde círculos e setas podem representar qualquer coisa. Podemos chamar os círculos de nós e as setas de links. Suponha que cada nó seja como uma pequena nota, artigo de resumo ou comentário. Não estou muito preocupado aqui se tem texto ou gráficos ou ambos. Idealmente, representa ou descreve uma determinada pessoa ou objeto (Berners-Lee, 1990, n.p., tradução nossa).

Figura 5. Desenho original incluído na proposta de Tim Berners-Lee ao CERN em 1989



Fonte: Berners-Lee (1990, n.p.)

Também abordava a importância dos vocabulários controlados por meio do que o autor denominou “o problema com as palavras-chave”, que demandavam que o usuário conhecesse previamente o sistema para que soubesse quais expressões utilizar na busca. Segundo o autor, na época os sistemas já exigiam o registro das palavras-chave, o que foi considerado um passo na direção certa, porém ainda sem vinculação entre elas e os demais elementos do sistema. A vinculação permitiria que uma palavra-chave fosse encontrada por meio dos nós aos quais ela estaria relacionada e esses vínculos poderiam ser com documentos, conceitos ou outros tipos de elementos (Berners-Lee, 1990, n.p., tradução nossa).

Segundo o autor, à época da proposta ainda não havia o conhecimento de que suas ideias de hipertexto já eram nomeadas como tal. Referindo-se especificamente a este conceito, o autor descreve no documento uma experiência pessoal datada de 1980, quando escreveu um programa de computador chamado Enquire. Neste sistema, “para encontrar informações, avançava-se através dos links de uma folha para outra” (Berners-Lee, 1990,n.p., tradução nossa).

Aponta-se no Enquire e nas ideias do modelo proposto o uso de uma taxonomia. Apesar de o autor criticar fortemente a estruturação hierárquica dos sistemas vigentes à época, é possível reconhecer o uso de polihierarquias na abordagem de palavras-chave e “linkagem” de dados, por exemplo. Há evidências do uso de conceitos-chave da ciência da informação - como termos, conceitos e multidimensionalidade - na construção da primeira versão da *web*, aspectos da interdisciplinaridade com a ciência da computação. Ao abordar o uso de hipertextos para controle de documentos, o autor afirma que “há muito apoio da indústria editorial e de bibliotecários, cujo trabalho é organizar a informação” (Berners-Lee, 1990, n.p., tradução nossa). As conclusões apresentadas no documento complementam o ideal por trás da *web*:

> Devemos trabalhar em direção a um sistema de informação universal vinculado, no qual a generalidade e a portabilidade são mais importantes do que técnicas gráficas sofisticadas e instalações extras complexas (Berners-Lee, 1990, n.p., tradução nossa)

2.3.2 A web 2.0

A literatura não aponta consenso em relação ao início da segunda geração da *web* ou ao surgimento do termo “*web* 2.0” (Toledano, 2013), porém converge para um

cenário de desenvolvimento que tem origem nos primeiros navegadores (Figo, 2020; O'Reilly, 2005; W3C, 2020). Apenas um ano após a criação da primeira proposta do que se tornaria a *web*, uma segunda foi publicada e foi desenvolvido o primeiro navegador *web*, que culminou na primeira comunicação cliente-servidor pela Internet, em dezembro de 1990. Desde então, foram necessários apenas mais dois anos para que o número de navegadores crescesse já em janeiro de 1993, ainda com interfaces pouco amigáveis, com texto puro e hipertextos. Em março do mesmo ano foi disponibilizada uma versão do Mosaic para o sistema operacional mais conhecido, o Windows. A novidade que o Mosaic apresentava era a evolução de uma interface com texto puro para uma com elementos gráficos, como ícones e imagens, além de funcionalidades de organização de conteúdo como as listas de favoritos (NCSA MOSAIC, 2021; W3C, 2020). As novas possibilidades apresentadas pelo Mosaic aqueceram o setor de tecnologia e o fim de 1993 foi marcado por discussões sobre o futuro da *web* (Figo, 2020; W3C, 2020).

O ano de 1994 é igualmente importante, uma vez que foi neste mesmo ano que: (a) os criadores do Mosaic lançaram o primeiro navegador comercial de apelo popular chamado Netscape, que é considerado o primeiro passo no desenvolvimento de negócios baseados na *web* por conta de sua popularidade; (b) os sistemas *dial-up* tradicionais, como AOL, começaram a fornecer acesso à Internet discada; e (c) em outubro, foi criado o *World Wide Web Consortium* (W3C), comunidade internacional liderada pelo inventor e diretor da Web, Tim Berners-Lee, para o desenvolvimento de padrões, protocolos e diretrizes que garantam o crescimento de longo prazo da *web* (W3C, 2020). Com o início das atividades do W3C ocorre também o desenvolvimento de tecnologias fundamentais da *web*, como *Hypertext Markup Language* (HTML), em 1995, e *Extensible Markup Language* (XML), em 1996 (W3C, 2020), que serão analisadas adiante. Tais tecnologias passaram a estar presentes nos sites de maior sucesso, o que tornou o ambiente ainda mais atrativo ao desenvolvimento de novos negócios, havia mais facilidade de uso das interfaces por parte daqueles que não detinham conhecimentos técnicos, foi o início de um ambiente convidativo à interação dos usuários por meio do uso e compartilhamento de conteúdo (Figo, 2020; Macaskill eOwen, 2006; W3C, 2020).

Os eventos citados contribuíram para o início de uma avalanche de

investimentos em empresas de tecnologia, que neste período adotavam geralmente o sufixo “.com” em seus endereços *web*, na tentativa de parecerem mais digitais, atraindo a atenção dos consumidores. A avalanche de investimentos valorizou as ações das chamadas “PontoCom”, que no dia 10 de março de 2000 alcançaram alta máxima histórica (à época), o que mudou drasticamente em seguida causando a quebra de grande parte delas, que perderam até 75% de seu valor de mercado. O episódio ficou conhecido como o “estouro da Bolha PontoCom” (Figo, 2020).

Apesar de a literatura não apontar consenso em relação ao surgimento do termo “*web* 2.0” ou ao seu início como uma segunda geração, todas as citações fazem referência a Tim O'Reilly como um dos primeiros a utilizá-lo (Cope; Kalantzis; Liam, 2011; Macaskill e Owen, 2006; Toledano, 2013). Segundo o próprio O’Reilly:

> O conceito de "Web 2.0" começou com uma sessão de brainstorming entre a O'Reilly e a MediaLive International. Dale Dougherty, pioneiro da Web e vice-presidente da O'Reilly, observou que, longe de ter "travado", a Web era mais importante do que nunca, com novos aplicativos e sites empolgantes surgindo com regularidade surpreendente. Além do mais, as empresas que sobreviveram ao colapso pareciam ter algumas coisas em comum. Será que o colapso das pontocom marcou algum tipo de ponto de virada para a *web*, de modo que um apelo à ação como "Web 2.0" possa fazer sentido? Concordamos que sim, e assim nasceu a Web 2.0 Conference.
>
> No ano e meio desde então, o termo "Web 2.0" claramente se consolidou, com mais de 9,5 milhões de citações no Google. Mas ainda há muita discordância sobre o que a Web 2.0 significa, com algumas pessoas condenando-a como uma palavra da moda de marketing sem sentido e outras aceitando-a como a nova sabedoria convencional (O’Reilly, 2005, n.p., tradução nossa).

Estudo de Toledano (2013, p. 1) propôs uma revisão de literatura acerca do termo “*web* 2.0”, definindo-o como:

> “a segunda geração da World Wide Web (www), caracterizada pela comunicação bidirecional e pela participação do usuário. Comumente chamada de *web* participativa, [...] é a evolução dos sites estáticos da Internet primitiva”.

A autora também associa conceitos-chave a esta nova geração da *web*, são eles: participação, conversação, conteúdo gerado pelo usuário e comunicação bidirecional.

Segundo Cope, Kalantzis e Liam (2011, p. 136, tradução nossa) a *web* 2.0 é:

> a aplicação mais interativa e extensivamente sociável da internet, [que] aponta para redes mais amplas de participação, maior capacidade de resposta a comentários, corpos de conhecimento mais profundamente integrados e culturas de conhecimento mais dinâmicas, receptivas e de movimento mais rápido.

Já para Macaskill e Owen (2006, p. 2, tradução nossa):

> Esta nova ou segunda geração de desenvolvimento *web* é muitas vezes contrastada com uma "Web 1.0" anterior, caracterizada por uma tecnologia em que computadores conectados permitiam que os internautas pesquisassem passivamente, recuperassem informações proprietárias (em grande parte baseadas em texto) de páginas estáticas da *web*.
>
> De homepages pessoais a blogs, de computadores pessoais a comunidades on-line, de diretórios a "tags", uma Web mais interativa hoje oferece aos usuários serviços que convidam à participação ativa por meio do uso de ferramentas que eliminam a necessidade de aprender código de programação. E os sites e projetos da Web 2.0 estão aumentando em número e popularidade.

Com base nas definições e nos eventos históricos envolvidos, é possível observar que apesar de não haver um limite claro em termos de fase evolutiva entre a *web* 1.0 e a 2.0, independente do termo, há que se verificar uma mudança no sentido da popularização da *web* no mundo social, com maior uso pelas pessoas das possibilidades de comunicação e colaboração que o ambiente oferecia, e no mundo organizacional, com a visão da *web* como um novo ambiente de negócios em produtos e serviços.

Conforme Hendler[16] (2009, p. 111, tradução nossa) sinaliza:

> Um ambiente aberto com um grande número de adotantes iniciais ansiosos, a Web é uma plataforma incrível para prototipagem rápida, integração de aplicativos e inovação. Assim como a mídia e as empresas estão enfrentando os fenômenos da Web 2.0, como as redes sociais, a Wikipedia e seus muitos descendentes, o crescimento da blogosfera e até mesmo o microblog, uma nova geração de tecnologias está surgindo. Tem havido alguma confusão sobre como chamar essa onda crescente de inovação, que é em parte uma realização da visão da Web Semântica expressa em 2001.

Toledano (2013) atribui a Darcy DiNucci o primeiro uso do temo "*web* 2.0" em

---

[16] Cabe diferenciar o autor Jim Hendler (2009) de James Hendler, que acompanha Tim Berners-Lee e Ora Lassila na criação da *web* semântica (Berners-Lee; Hendler; Lassila, 2001).

1999, no artigo *Fragmented Future*, onde é cunhado no contexto do *web design* e antecipa características do que seria a *web* do futuro, como o uso de dispositivos móveis, a necessidade de formatos padrão para leitura e processamento de informações por meio de qualquer dispositivo e a importância de considerar o tamanho e a resolução da tela ao projetar interfaces.

2.3.3 A web 3.0

A história da *web* continua então a partir da criação do W3C, em outubro de 1994, que trouxe consigo o desenvolvimento e a evolução de uma série de tecnologias para a realização da *web* semântica (Hendler, 2009; W3C, 2020), como *Hypertext Markup Language* (HTML), em 1995 e evolução em 1997; *Extensible Markup Language* (XML), em 1996; em 1999 publicam a primeira especificação da tecnologia que se tornaria a base da *web* semântica, o *Resource Description Framework* (RDF); e em 2000 surge o *Uniform Resource Identifier* (URI) (W3C, 2020). Em 2001 é publicado o artigo "*The Semantic Web*" na *Scientific American*, de autoria: do então diretor do W3C e reconhecido como o inventor da *web*, Tim Berners-Lee, também pesquisador do Laboratório de Ciência da Computação do Instituto de Tecnologia da Universidade de Massachusetts; James Hendler, professor de ciência da computação da Universidade de Maryland, onde desenvolvia pesquisas voltadas à representação do conhecimento no contexto da *web*; e Ora Lassila, pesquisador do Centro de Pesquisas Nokia em Boston e cientista chefe da Nokia Venture Partners, além de membro do conselho do W3C e co-autor da primeira especificação RDF (Berners-Lee; Hendler; Lassila, 2001).

No estudo, os autores abordam a expressão de significado, a representação do conhecimento, apresentam um glossário com os principais conceitos da *web* semântica, explicam o uso de taxonomias como parte das ontologias, citam brevemente a evolução das pesquisas sobre Inteligência Artificial (IA), falam da atuação dos agentes inteligentes e contextualizam a *web* semântica no âmbito da evolução do conhecimento (Berners-Lee; Hendler; Lassila, 2001). O foco principal se colocava em conferir semântica por meio da capacidade computacional dos objetos, conforme afirmam que:

> Ao aumentar as páginas da Web com dados direcionados a computadores e ao adicionar documentos exclusivamente para

> computadores, transformaremos a Web na Web Semântica. Os computadores encontrarão o significado dos dados semânticos seguindo hiperlinks para definições de termos-chave e regras para raciocinar sobre eles logicamente. A infraestrutura resultante estimulará o desenvolvimento de serviços da Web automatizados, como agentes altamente funcionais. Os usuários comuns comporão páginas da Web semântica e adicionarão novas definições e regras usando software de prateleira que ajudará na marcação semântica (Berners-Lee; Hendler; Lassila, 2001, p. 36, tradução nossa).

Nesse contexto, cabem as definições conferidas aos principais conceitos da então recém-criada *web* semântica, conforme glossário apresentado no artigo:

> HTML: Linguagem de Marcação de Hipertexto. A linguagem usada para codificar formatação, links e outros recursos em páginas da Web. Usa “tags” padronizados como <H1> e <BODY> cujo significado e interpretação são definidos universalmente pelo World Wide Web Consortium.
>
> XML: eXtensible Markup Language. Uma linguagem de marcação como HTML que permite que indivíduos definam e usem suas próprias tags. O XML não possui nenhum mecanismo interno para transmitir o significado das novas tags do usuário para outros usuários.
>
> RECURSO: Jargão da Web para qualquer entidade. Inclui páginas da Web, partes de uma página da Web, dispositivos, pessoas e muito mais.
>
> URL: Localizador Uniforme de Recursos. Os códigos familiares (como http://www.sciam.com/index.html) usados em hiperlinks.
>
> URI: Identificador de recurso universal. URLs são o tipo mais conhecido de URI. Um URI define ou especifica uma entidade, não necessariamente nomeando sua localização na Web.
>
> RDF: Estrutura de descrição de recursos. Um esquema para definir informações na Web. O RDF fornece a tecnologia para expressar o significado de termos e conceitos de uma forma que os computadores possam processar prontamente. RDF pode usar XML para sua sintaxe e URIs para especificar entidades, conceitos, propriedades e relações.
>
> ONTOLOGIAS: Coleções de declarações escritas em uma linguagem como RDF que definem as relações entre conceitos e especificam regras lógicas para raciocinar sobre eles. Os computadores “entenderão” o significado dos dados semânticos em uma página da Web seguindo links para ontologias específicas.
>
> AGENTE: Um pedaço de software que é executado sem controle humano direto ou supervisão constante para atingir os objetivos

> fornecidos por um usuário. Os agentes geralmente coletam, filtram e processam informações encontradas na Web, às vezes com a ajuda de outros agentes.
>
> DESCOBERTA DE SERVIÇO: O processo de localização de um agente ou serviço automatizado baseado na Web que executará uma função necessária. A semântica permitirá que os agentes descrevam uns aos outros precisamente qual função eles executam e quais dados de entrada são necessários (Berners-Lee; Hendler; Lassila, 2001, p. 38).

Com base nos principais aspectos da *web* 3.0 ou *web* semântica, cabe a compreensão dos três estágios desde a criação da *web*: *web* 1.0, *web* 2.0 e *web* 3.0. A primeira geração de serviços e produtos na *web* possibilitou a organização e visualização de informações em um novo ambiente e de forma compartilhada, porém sem que houvesse meios para amplo uso e manipulação pelos usuários, que não detinham os conhecimentos técnicos para tanto. Assim, o uso limitava-se majoritariamente à leitura. A segunda geração da *web* proporcionou aos usuários tanto ler quanto produzir informações em sites e aplicações sem que para isso houvesse a necessidade de conhecimento técnico, o mercado proporcionava acesso a funcionalidades que facilitavam o uso. E a terceira, tem como enfoque a interação homem-máquina e propicia leitura, gravação e processamento em tempo real e de forma vinculada a outros dados disponíveis na *web*.

O conceito de Web Semântica, conferido pelo W3C e estabelecido a partir de 2004 com a atualização de uma série de especificações, dentre elas a do RDF, envolve diferentes aspectos. Trata-se (Berners-Lee; Hendler; Lassila, 2001; Catarino eSouza, 2012; Ferreira eSantos, 2013; W3C, 2013, 2015):

(a) do nome dado ao esforço colaborativo liderado pelo W3C para maior interação entre computadores e pessoas;

(b) de uma nova geração da *web*, que envolve o desenvolvimento de tecnologias, linguagens, padrões e recomendações que tornem as informações legíveis por máquinas;

(c) da estruturação de páginas da *web* adicionando a elas significado e desenvolvendo um ambiente onde agentes inteligentes criem tarefas sofisticadas a partir do conteúdo das páginas e de bases de dados.

2.3.4 RDF e SKOS: relações entre web semântica e taxonomias

A compreensão do arcabouço que envolve a *web* semântica e também sua relação com as taxonomias, pode ser melhor compreendida por meio da conceituação de seus elementos, iniciando pela principal tecnologia, o *Resource Description Framework* (RDF), em seguida introduzindo o *Simple Knowledge Organization System* (SKOS), o padrão atual de representação do conhecimento na *web*.

O RDF foi criado para expressar informações sobre qualquer recurso para a *web*, esteja ele disponível ou não neste ambiente. Por recurso podemos entender desde conceitos abstratos a objetos físicos (Schreiber eRaimond, 2014). Essa capacidade encontra um dos principais objetivos do W3C, que é a melhoria de funcionalidades e interoperabilidade do ambiente *web*, e a concepção de *web* semântica, onde se localizam conceitos como dados lincados, aplicações verticais e recuperação da informação (Catarino; Cervantes; Almeida, 2015). Tais objetivos só podem ser alcançados por meio de informações estruturadas de contexto, os chamados metadados, que facilitam a recuperação, o (re)uso e a gestão dos dados. Quanto mais detalhados ou enriquecidos são os metadados, maior será a capacidade de compreensão dos recursos, sua origem, proveniência, propósito, autoria, história, licenças e detalhes técnicos (Rocha; Sales; Sayão, 2017). Porém, conforme Ferreira e Santos (2013) sinalizam, para o uso efetivo de metadados entre aplicativos (interoperabilidade), é necessário o estabelecimento de padrões semânticos, sintáticos e estruturais. Com RDF é possível obter uma estrutura padrão para o intercâmbio de informações entre sistemas sem perda de significado (Catarino; Cervantes; Almeida, 2015).

O RDF permite descrever cada recurso a partir de declarações sobre eles e sobre as relações que guardam com outros recursos, independente do ambiente em que estejam. É composto basicamente por triplas e pode ser expressado em grafos e serializações, sendo a *Extensible Markup Language* (XML) a serialização mais comum; seu principal enfoque é o processamento da informação por máquinas. Funciona como um modelo de dados, onde as triplas dão o formato adotado no modelo. Assim, sempre haverá numa declaração RDF a descrição: (1) do sujeito, isto é, aquilo que a declaração RDF descreve; (2) do predicado, que identifica a propriedade, atributo ou característica do sujeito; e (3) do objeto, que identifica o valor

ou conteúdo do predicado (Ferreira; Santos, 2013). Quando expresso em grafo, o RDF é composto por nós em arcos direcionados sempre do sujeito para o objeto, sendo a relação entre eles o predicado. Segundo Ferreira e Santos (2013), “esse modo de representar os dados em RDF facilita o aprendizado do modelo e a leitura por humanos, além de servir [de maneira precisa] para a modelagem conceitual de um domínio”. Já uma serialização RDF pode ser realizada a partir da estruturação sistemática e hierarquizada de elementos de dados através de uma sintaxe, isto é, por meio da estruturação do RDF em uma cadeia de caracteres lineares facilmente legíveis por máquinas.

Mendes e Bentes (2019, p. 38) afirmam que “com o avanço do sistema *web* a taxonomia ganha maior importância para a representação, organização e recuperação da informação”. Já para Campos e Gomes (2007, p. 2) as taxonomias permitem que “se estabeleçam padrões de alto nível para a ordenação e classificação de informação por meio do uso de mecanismos de herança. O conceito de herança é um dos mais poderosos no desenvolvimento de software” e diz respeito à possibilidade de uma classe herdar características e métodos de outra classe.

A estrutura RDF se assemelha à estrutura de uma taxonomia, podendo ser utilizada para a representação dos termos de uma taxonomia em um formato estruturado, através da criação de relações semânticas entre eles. A combinação do RDF com taxonomias poderia permitir maior interoperabilidade entre sistemas de informação, já que as informações representadas em RDF independem do uso de um software ou aplicação específica e podem ser usadas para conectar diferentes fontes de informação. Nesse sentido, o RDF constitui a base do *Simple Knowledge Organization System* (SKOS), o modelo padrão para representação de Sistemas de Organização do Conhecimento (SOC) na *web* semântica.

A expressão de dados em RDF é um dos princípios do *Linked Data* (LD) ou Dados Conectados, conceito de Tim Berners-Lee, que surge como parte do desenvolvimento da *web* semântica e se traduz num conjunto de boas práticas para criação e publicação de dados estruturados com a finalidade de conectá-los e enriquece-los através da vinculação entre suas relações. Segundo o W3C, tecnologias como RDF, SPARQL, OWL e SKOS possibilitam os dados conectados (Isotani eBittencourt, 2015; W3C, 2015).

Smith (2022, n.p.) afirma que:

> O modelo de dados SKOS é centrado em conceitos, que podem ter rótulos preferenciais e alternativos em qualquer idioma, bem como outros metadados, e são identificados por endereços na *World Wide Web* (URIs[17]). Os conceitos são agrupados em hierarquias por meio de relações mais amplas e mais estreitas, com os principais conceitos no nível conceitual mais amplo. Os conceitos também são organizados em esquemas de conceitos, também identificados por URIs. Outras relações, mapeamentos e agrupamentos também são suportados (Smith, 2022, n.p.).

O SKOS é o modelo recomendado pelo W3C para a publicação de SOC de todos os tipos seguindo o princípio de interoperabilidade que acompanha a proposição da *web* semântica e que mais tarde origina conceitos complementares como *Linked Open Data* (LOD) e princípios FAIR (*Findable, Accessible, Interoperable, Reusable*), os quais não serão trabalhados na presente pesquisa por envolverem um arcabouço conceitual mais amplo.

O desenvolvimento da modelagem SKOS se deu de maneira conjunta pelos grupos de trabalho das normas internacionais BS 8723 (britânica) e ISO 25964 e por isso guardam elevado grau de compatibilidade. Os modelos de dados da ISO 25964 e o SKOS não são idênticos, já que a primeira se dedica aos tesauros, independente do ambiente de aplicação, enquanto o último atende a qualquer tipo de SOC, mesmo àqueles que não estejam em conformidade com a ISO 25964 ou outra norma. Devido a este alinhamento há disponível uma série de correspondências documentadas, incluindo esquemas RDF que fornecem uma versão legível por máquina a esses mapeamentos (Dextre Clarke, 2019).

Isotani e Bittencourt (2015) comparam a Web de Documentos à Web de Dados, ambos como ambientes onde os recursos se unem por meio de URIs, porém nesta última, metadados são disponibilizados para permitir que as máquinas possam compreender melhor os dados contidos nestes recursos e o significado de determinada relação entre dois ou mais recursos, enquanto links RDF são usados para acessar dados de diferentes fontes. Conforme Gomes e Campos (2019, p. 48)

---

[17] Uniform Resource Identifiers (URI), também chamados de identificadores persistentes, são cadeias de caracteres lineares que tornam os recursos únicos e não confundíveis, possibilitando que se identifique recursos dentro e fora do ambiente *web* (Catarino; Souza, 2012b).

destacam:

> O aumento considerável das tecnologias que promoveram interfaces abertas e a criação de padrões de metadados abrangentes foram grandes passos no sentido de avançar o processo de interoperabilidade. O problema está no tratamento desses dados em nível não somente descritivo, mas semântico. Neste sentido, um dos grandes problemas que a Web enfrenta está relacionada a compatibilidade semântica entre essas fontes de informação, garantindo a manipulação do conjunto sem forçar a conversão das partes em um formato único (Gomes; Campos, 2019, p. 48).

Nas palavras dos idealizadores da *web* semântica:

> A taxonomia define classes de objetos e relações entre eles. [...] Classes, subclasses e relações entre entidades são uma ferramenta muito poderosa para uso na Web. Podemos expressar um grande número de relações entre entidades atribuindo propriedades a classes e permitindo que subclasses herdem tais propriedades (Berners-Lee; Hendler; Lassila, 2001, p. 40).

Em resumo, Gomes e Campos destacam e concordam com Isotani e Bittencourt, que a interoperabilidade entre diferentes sistemas e fontes de dados é um dos principais desafios da *web* semântica, que exige o tratamento semântico adequado dos dados para que possam ser utilizados de forma eficiente e efetiva. Nesse contexto, há que se considerar o desenvolvimento de taxonomias para o enriquecimento semântico dos objetos informacionais.

3 METODOLOGIA

Este capítulo se destina ao desenho da pesquisa e metodologia aplicada. Segundo Deslandes e Gomes (2007, p.7), a metodologia “ocupa um lugar central no interior das teorias e está referida a elas”, inclui ao mesmo tempo o método, as técnicas e o pesquisador, este último constituído de sua experiência, sua capacidade pessoal e sua sensibilidade.

3.1 DESENHO DA PESQUISA E PROCEDIMENTOS

Para atender ao **objetivo** da pesquisa, atualizar e sistematizar o estado da arte das taxonomias na ciência da informação no contexto da *web* semântica, delineou-se uma pesquisa do **tipo** qualitativa, teórica, descritiva e documental empreendida em duas **etapas**: (1) uma revisão narrativa nos domínios da organização e da recuperação da informação e do conhecimento e da *web* semântica, e (2) uma revisão integrativa sobre as relações entre taxonomias na ciência da informação e *web* semântica. A **coleta de dados** em ambas as etapas consistiu na recuperação de materiais bibliográficos obtidos junto a bases de dados selecionadas; especificidades se colocam em relação a cada tipo de revisão apresentada, às quais serão descritas e diferenciadas adiante.

A busca pela compreensão profunda de um fenômeno social específico demanda a realização de uma pesquisa qualitativa, cujo processo de trabalho científico é iniciado na fase exploratória, onde é definido e delimitado o objeto a ser desenvolvido teorica e metodologicamente e construído o projeto de pesquisa. Assim, nessa fase foram coletados os primeiros materiais bibliográficos que delinearam o objeto aqui proposto e a espinha dorsal do referencial teórico, bem como estruturado o projeto de pesquisa, seu cronograma e etapas. O passo seguinte se deu pelo desenvolvimento do *constructo* da revisão de literatura, quando o referencial teórico se consolida; para esta etapa foi adotada a revisão narrativa de literatura. Já a segunda etapa da pesquisa se delineou por meio de uma revisão integrativa, que buscou atualizar e sistematizar o estado da arte das taxonomias na ciência da informação no contexto da *web* semântica, a partir da produção científica disponível, descrevendo seus elementos constituintes, relações interdisciplinares e aplicações.

Segundo Booth, Papaioannou e Sutton (2012) e Rother (2007), uma revisão de

literatura busca sintetizar e comunicar à comunidade científica o que se sabe sobre determinado assunto, podendo ser conduzida de três maneiras: narrativa, integrativa e sistemática. A Tabela 1 apresenta as diferenças entre essas três formas, consideradas as mais comuns.

Ao nos referimos aos elementos e atributos de uma revisão de literatura teremos que a narrativa confere ao pesquisador maior liberdade para localizar e analisar estudos e documentos sem que as escolhas e os passos realizados necessitem seguir cirtérios preestabelecidos e relatados. Diferentemente, às revisões integrativa e sistemática é exigido o cumprimento de determinadas etapas que garantam a confiabilidade dos achados e a reprodutibilidade do estudo. As formas narrativa e integrativa, indicadas a estudos qualitativos, são adotadas na presente pesquisa e detalhadas adiante, bem como aprofundadas as diferenças entre as revisões integrativa e sistemática da literatura (Booth; Papaioannou; Sutton, 2012; Rother, 2007).

Os primeiros levantamentos bibliográficos realizados ainda na fase exploratória da pesquisa indicaram haver poucos estudos de revisão sistemática no campo da Ciência da Informação, o que foi sinalizado por Cerrao, Castro e Jesus (2018) em estudo que abordou a adoção do método de revisão sistemática por pesquisadores brasileiros na área da CI. Os autores destacaram a importância desse método, especialmente numa área caracterizada pela interdisciplinaridade. Outros autores também destacaram a escassez de estudos e a baixa frequência de uso de taxonomias em comparação com o cenário internacional (Lopes; Aganette; Maculan, 2020; Mendes; Bentes, 2019).

Tabela 1. Diferenças entre os tipos de revisão mais comuns

Diferença	Narrativa	Integrativa	Sistemática
Proposta	visão geral sobre um tópico de investigação para um estudo de pesquisa, dissertação ou revisão independente	Análise crítica da literatura empírica, metodológica ou teórica, que chama a atenção para futuras necessidades de pesquisa	Responde a uma pergunta objetiva
Equipe	Um ou mais revisores	Recomenda-se o envolvimento de dois ou mais revisores e bibliotecário	Recomenda-se o envolvimento de três ou mais revisores, incluindo um bibliotecário ou especialista em informação e um estatístico se a metanálise for realizada
Protocolo de Revisão	Não	Não necessita	Recomendado o registro do protocolo
Questão de Revisão	Não	Um propósito amplo definido e/ou pergunta(s) de revisão	Questão objetiva única geralmente no formato PICO (P = população; I = intervenção; C = comparação; O = resultados (*outcomes*).

Diferença	Narrativa	Integrativa	Sistemática
Diretrizes para relatoria	Não	Pode utilizar o modelo PRISMA	Deve utilizar o modelo PRISMA
Duração estimada	2-6 meses	6-12 meses	12-24 meses
Uso de metodologia sistemática e reprodutível de pesquisa	Não	Sim	Sim
Uso de critérios de inclusão e exclusão	Não	Sim	Sim
Uso de diagrama com fluxo de pesquisa	Não	Sim	Sim, PRISMA
Avaliação crítica de estudos	Não	Sim	Sim
Extração de dados	Não	Sim	Sim
Análise e Síntese	Análise narrativa	Análise narrativa e/ou temática com síntese descritiva e qualitativa	Análise narrativa e/ou temática com síntese descritiva e qualitativa, podendo incluir síntese quantitativa (metanálise)
Rigor científico	Não	Sim	Sim

Fonte: Adaptado de Toronto e Remington (2020, p.3)

No Guia *Cochrane handbook for systematic reviews of interventions*, reconhecido por fornecer diretrizes para a realização de revisões sistemáticas de alta qualidade, Higgins et al. (2019) apontam fatores que refletem a importância desse tipo de estudo, conforme elencado a seguir:

- evitar a duplicidade de estudos ao se projetar ou encomendar novas pesquisas;
- identificar estudos atuais e em andamento, bem como lacunas específicas de conhecimento ou de evidências;
- revelar limitações na condução de estudos anteriores que podem ser abordadas em novos estudos;
- possibilitar aos autores declarações confiáveis acerca da extensão do conhecimento humano em áreas importantes, identificando prioridades para pesquisas futuras.

Booth, Papaioannou e Sutton (2012) completam:

- contextualizar trabalhos, tornando-os mais aplicáveis;
- revelar pontos comuns e as relações entre estudos, sistematizando seus pontos mais relevantes;
- identificar e solucionar conflitos entre estudos;
- revelar conceitos, técnicas, ferramentas e métodos;
- criar fundamentos robustos para o avanço do conhecimento.

Além disso, afirmam que as revisões sistemáticas sintetizam o conhecimento, formulando estimativas mais confiáveis da situação de determinada temática, podendo ser parte ou até mesmo a totalidade de qualquer trabalho acadêmico, já que informa de forma confiável sobre a natureza ou extensão de um problema, de um potencial benefício, de uma preocupação científica, de incertezas e de custos de intervenções e políticas (Booth; Papaioannou; Sutton, 2012).

Os pontos elencados por Higgins et al. (2019) e Booth, Papaioannou e Sutton (2012), com a sinalização de Cerrao, Castro e Jesus (2018), parecem indicar que a frequência de revisões sistemáticas de um campo pode ter relação com o seu nível de maturidade científica. Nesse sentido, torna-se premente a realização desse tipo de

estudo no campo da Ciência da Informação.

3.2 REVISÃO NARRATIVA

A revisão narrativa foi conduzida nas áreas de organização e representação da informação e do conhecimento e da *web* semântica. A seleção desses domínios se deu pela visão inicial de que o objeto taxonomias se constitui, de maneira geral, no campo da Ciência da Informação, como um SOC, mapa do conhecimento, esquema de classificação, instrumento de representação semântica, o que o localiza tanto na OC quanto na RC e nos indica a necessidade de compreensão também dos elementos constituintes da OI e da RI, bem como suas semelhanças e diferenças. Do mesmo modo, a *web* semântica é o terreno onde a pesquisa deseja aplicar suas análises e contribuições, necessitando, portanto, explorar esse conceito e seu desenvolvimento, bem como suas relações com as taxonomias. Nesta etapa foram consultadas bases de dados como o Google Acadêmico, a Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI), o Portal de Periódicos Capes e a base Scopus, além de *sites* e livros de referência nas temáticas pesquisadas.

Alguns autores apontam as revisões narrativas como mais apropriadas para descrever o estado da arte de determinado assunto, explorar um tópico ou abordar uma questão de pesquisa mais ampla, uma vez que se constituem basicamente da interpretação e análise crítica subjetiva do pesquisador, não sendo exigida a aplicação de estratégias de busca sofisticadas e exaustivas e preocupação com a redução de vieses de diferentes tipos; em outras palavras, uma revisão narrativa não exige um protocolo rígido para sua condução, as escolhas metodológicas em todas as etapas e procedimentos são realizadas livremente pelo pesquisador com base nos objetivos da pesquisa. Desse modo, em uma revisão narrativa há maior facilidade para o pesquisador adquirir e atualizar o conhecimento sobre uma temática em um curto espaço de tempo (Camargo Júnior et al., 2023; Casarin et al., 2020; Rother, 2007; Sant’anna Ramos Vosgeraue Paulin Romanowski, 2014).

No trabalho *Estudos de revisão: implicações conceituais e metodológicas*, Sant’anna Ramos Vosgerau e Paulin Romanowski (2014) analisaram revisões selecionadas por meio da base de periódicos Capes e identificaram o uso de uma diversidade de termos na literatura para denominar os diferentes tipos de estudos de revisão. Com a finalidade de relacioná-los e distingui-los os autores os agruparam em

estudos que mapeiam e estudos que avaliam e sintetizam o conhecimento. Dentre os estudos que mapeiam estão a revisão de literatura ou revisão bibliográfica, pesquisa do tipo estado da arte ou estado do conhecimento, revisão narrativa e estudos bibliométricos. Já dentre os estudos que avaliam e sintetizam o conhecimento, os autores classificaram: a revisão sistemática, a revisão integrativa, a síntese de evidências qualitativas, a metanálise e a metassíntese qualitativa ou metasumarização.

A principal diferença entre os dois grupos está nos procedimentos de pesquisa, conforme demonstrado no Quadro 2. Outras diferenças significativas podem ser encontradas entre os tipos de estudo, especialmente em relação ao uso de ferramentas e métricas e à visualização dos dados, pontos que se distanciam do objeto discutido no presente trabalho, por isso não são abordados, assim como outros tipos de estudos de revisão não incluídos no estudo de Ramos Vosgerau e Paulin Romanowski (2014), como por exemplo as metaetnografias e revisões de escopo. Além disso, é possível identificar na literatura diferentes formas de classificação dos estudos de revisão, especialmente quando localizados na área da saúde, onde as formas sistematizadas prevalecem e cujas etapas serão aprofundadas adiante.

Também, é possível identificar na literatura o uso do termo “estado da arte” de diferentes formas, por vezes como um tipo de estudo, outras vezes como um aspecto, característica ou objetivo de uma pesquisa, sendo frequentemente associada às revisões narrativas pela semelhança metodológica existente entre elas (Casarin et al., 2020; Goes; Fernandez, 2018; Sant’anna Ramos Vosgerau e Paulin Romanowski, 2014).

Quadro 2. Comparativo de tipos de estudos de revisão

Grupo	Tipo de Estudo	Proposta	Questão de pesquisa	Critérios predefinidos	Relatoria transparente	Controle de viés
M	Revisão Bibliográfica	Coleta e análise de uma variedade de fontes de informação, como artigos científicos, livros e documentos acadêmicos para contextualização, problematização ou primeira validação de quadro teórico.	Ampla	Não	Não	Não
M	Estado da Arte	Análise aprofundada de produções bibliográficas em determinada área com enfoque na problematização e na metodologia. Revela enfoques e perspectivas e contribui com teoria e prática de uma área do conhecimento, tem como finalidade principal o mapeamento.	Ampla	Não	Não	Não

Grupo	Tipo de Estudo	Proposta	Questão de pesquisa	Critérios predefinidos	Relatoria transparente	Controle de viés
M	Revisão Narrativa	Abordagem descritiva e subjetiva, narra a evolução das ideias e teorias em torno de um tópico e seu contexto. Aprofundamento de análise em relação à revisão bibliográfica, estabelece relações entre produções, identifica temáticas recorrentes, aponta novas perspectivas e/ou consolida uma área de conhecimento.	Ampla	Não	Não	Não
M	Estudos Bibliométricos	Utiliza métodos estatísticos para realizar análise quantitativa de publicações científicas com o objetivo de compreender as características e tendências da literatura acadêmica em uma determinada área.	Ampla	Sim, especialmente em relação à seleção de fontes de dados e à definição de métricas.	Sim, no que diz respeito à compreensão do estudo, sem exigência de reprodutibilidade.	Sim

Grupo	Tipo de Estudo	Proposta	Questão de pesquisa	Critérios predefinidos	Relatoria transparente	Controle de viés
AS	Revisão Sistemática	Sintetiza de forma objetiva e imparcial todas as evidências disponíveis sobre uma pergunta de pesquisa específica.	Única e específica	Sim, exigido	Sim, exigido	Sim, exigido
AS	Revisão Integrativa	Combina elementos da revisão sistemática e da revisão narrativa, de modo a integrar e sintetizar estudos de uma variedade de fontes.	Ampla	Sim, exigido	Sim, exigido	Sim
AS	Síntese de Evidências Qualitativas	Sintetiza estudos primários para uma compreensão profunda e holística de um tópico, com base em evidências qualitativas, numa abordagem mais flexível e interpretativa.	Ampla ou específica conforme objetivo do estudo.	Sim	Não exigida	Sim

Grupo	Tipo de Estudo	Proposta	Questão de pesquisa	Critérios predefinidos	Relatoria transparente	Controle de viés
AS	Metanálise	Utiliza métodos estatísticos para combinar e analisar dados quantitativos de estudos primários. A análise avalia, dentre outros fatores, a heterogeneidade entre os estudos incluídos.	Única e específica.	Sim, exigido	Sim, exigido	Sim, exigido
AS	Metasumarização	Sintetiza estudos primários sobre um tópico específico, com base em evidências qualitativas. É frequentemente indicada para explorar aspectos particulares de um fenômeno.	Com foco mais restrito, específica.	Sim, exigido	Sim, exigido	Sim, exigido

Legenda: M = estudos que mapeiam o conhecimento; AS = estudos que avaliam e sintetizam o conhecimento.
Fonte: elaborado pelo autor com base em Sant'anna Ramos Vosgerau e Paulin Romanowski (2014).

Por fim, o objetivo da revisão narrativa foi explorar o conhecimento acerca das temáticas taxonomias na CI e *web* semântica, com especial atenção ao histórico e evolução desta desde a sua origem. As já citadas características deste tipo de revisão permitiram o desenvolvimento das ideias e teorias em torno dos tópicos de interesse e seu contexto, facilitando explorar a área e organizar um conhecimento abrangente. Referindo-se às revisões integrativas, Rodrigues, Sachinski e Martins (2022) afirmam que o pesquisador compreende em profundidade sua área de pesquisa e pode verificar como a abordagem qualitativa se expressa nos estudos encontrados. O mesmo pode ser dito de uma revisão narrativa planejada e bem conduzida. Neste sentido e diante da heterogeneidade e da interdisciplinaridade inerentes às temáticas das taxonomias e da *web* semântica, acredita-se que as revisões narrativa e integrativa possam atuar de forma complementar, atendendo aos objetivos da pesquisa.

3.3 REVISÃO INTEGRATIVA

A Revisão Integrativa (RI), objeto da presente pesquisa, busca atualizar e sistematizar o estado da arte das taxonomias na ciência da informação e sua relação com a *web* semântica; recupera a produção científica disponível nos últimos cinco anos, de modo a captar pesquisas recentes e traçar as possíveis convergências entre as temáticas. Uma combinação de fatores motivou a escolha da RI para essa etapa da pesquisa, são eles: (a) a confiabilidade conferida pelo rigor metodológico deste tipo de revisão; (b) a amplitude do tema da pesquisa; (c) a possibilidade de analisar qualitativamente um *corpus* científico suficientemente abrangente para comportar o tema pretendido; e (d) a complementaridade existente entre duas abordagens descritivas, sendo uma majoritariamente subjetiva (revisão narrativa) e outra sistemática (revisão integrativa), o que possibilita o mapeamento de uma área e a atualização de seu estado da arte (Casarin et al., 2020; Deslandes; Gomes, 2007; Rother, 2007; Sant’anna Ramos Vosgerau; Paulin Romanowski, 2014; Saunders; Lewis; Thornhill, 2019; Souza; Silva; Carvalho, 2010).

Saunders, Lewis e Thornhill (2019, p. 79) definem a RI como aquela “que critica e sintetiza a literatura representativa sobre um tópico de forma integrativa para gerar novas estruturas e perspectivas”. Souza, Silva e Carvalho (2010) apontam que as RI permitem:

> incorporar um vasto leque de propósitos: definição de conceitos, revisão de teorias e evidências, e análise de problemas metodológicos de um tópico particular. A ampla amostra, em conjunto com a multiplicidade de propostas, deve gerar um panorama consistente e compreensível de conceitos complexos, teorias ou problemas (Souza; Silva; Carvalho, 2010, p. 103)

O conhecimento existente sobre as características de uma RI e suas diferenças em relação a outros tipos de pesquisas bibliográficas ainda é escasso, apesar do vasto material encontrado no âmbito das Revisões Sistemáticas de Literatura (RSL). Frequentemente utilizadas no campo da saúde, especialmente em estudos da área da enfermagem, as RI tem como propósito o profundo entendimento de um determinado fenômeno a partir de uma questão mais ampla de pesquisa. Numa RI é permitida a combinação de literatura teórica e empírica, uma vez que a qualidade de cada um dos estudos é avaliada e os selecionados são individualmente interpretados e sintetizados (Mendes; Silveira; Galvão, 2008; Mota de Sousa et al., 2018; Sant'anna Ramos Vosgerau; Paulin Romanowski, 2014).

Segundo Casarin et al. (2020), em trabalho denominado *Tipos de revisão de literatura: considerações das editoras do Journal of Nursing and Health*, a abordagem mais adequada à análise de fontes exclusivamente primárias é a RSL, cujos resultados são considerados o nível de evidência científica de maior grandeza. São muito comuns na medicina e partem de uma pergunta clara e específica, o que torna as RSL úteis à tomada de decisão na prática clínica. No entanto, Camargo Júnior et al. (2023, p. 4) afirmam que uma RI "busca investigar minuciosamente e combinar estudos de diversas metodologias, [...] pode ser utilizada em diversas áreas do conhecimento, mantendo o rigor metodológico de revisões sistemáticas". A síntese do conhecimento realizada por uma RI e a variedade na composição dos estudos incluídos, no contexto de um processo de pesquisa rigorosamente conduzido, reduz incertezas sobre conceitos, fenômenos e recomendações resultantes desse tipo de estudo (Souza; Silva; Carvalho, 2010; Toronto; Remington, 2020).

As RI seguem um método de pesquisa bem definido e com etapas reprodutíveis, similares aos das RSL e das metanálises, porém não incluem métodos estatísticos de síntese e apresentam importantes diferenças "em proposta e escopo, nos tipos de literatura incluídos e no tempo e recursos necessários para sua execução" (Toronto; Remington, 2020, p. 2). Segundo Toronto e Remington (2020), possuir tais etapas é o principal critério para que uma revisão seja considerada rigorosa.

Segundo Higgins et al. (2019), ao analisarmos as semelhanças existentes entre a RI e a RSL é possível encontrar: (a) uma questão de pesquisa; (b) clareza de escopo e critérios de elegibilidade dos estudos; (c) esforço para seleção de estudos relevantes e garantia de contabilização dos vieses encontrados; e (d) análise dos estudos incluídos para traçar conclusões baseadas nos achados de forma imparcial e objetiva. Toronto e Remington (2020) afirmam ainda que ambas exigem uma abordagem transparente e rigorosa, apesar de ainda não haver para as RI um conjunto consistente de padrões e guias sobre a melhor maneira de conduzi-las.

Neste sentido, o presente trabalho adota como guia a obra A Step-by-Step Guide to Conducting an Integrative Review (2020) e utiliza o modelo recomendado pela Cochrane Library, Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA), para registro, representação e relatoria do processo de pesquisa, conferindo maior rigor. No APÊNDICE A **-** Relatório Checklist PRISMA, é possível verificar se cada componente da presente pesquisa foi completa e devidamente relatado, sendo, portanto, reprodutível.

O guia de Toronto e Remington (2020) adota a abordagem metodológica de Cooper, amplamente utilizada para condução de RI, que consiste em cinco etapas evoluídas para seis ao longo dos anos, uma vez que o último passo abordava, dentre outras coisas, a disseminação dos resultados, desta vez compreendida como um passo independente. A Figura 6 apresenta as seis etapas.

Figura 6. As seis etapas do processo de revisão integrativa, Toronto e Remington (2020)

Fonte: elaborado pelo autor (2023)

No passo 1 são declaradas e formuladas a proposta da pesquisa de forma ampla e/ou questão(ões) de revisão; o passo 2 consiste na recuperação da literatura a partir de estratégia de busca assertiva e replicável e na seleção de estudos a partir de critérios de inclusão e exclusão predefinidos utilizando a técnica de *screening*, onde as citações e os resumos recuperados do processo de busca são avaliados para seleção daqueles considerados elegíveis à análise do texto completo; já no passo 3, cada estudo selecionado é então avaliado de forma completa; o passo 4 consiste na análise narrativa dos estudos e na integração de dados para identificação de novas perspectivas sobre o assunto; no passo 5 os achados são então apresentados pelo revisor e comparados com a revisão narrativa de literatura realizada como base para o planejamento da revisão integrativa, além de sinalizadas as limitações do estudo e realizada análise concisa dos resultados e principais contribuições do estudo; por fim,

no passo final, o estudo é disseminado em diferentes meios, a fim de alcançar o maior número possível de interlocutores da comunidade científica (Toronto; Remington, 2020).

Neste contexto, realiza-se pesquisa bibliográfica qualitativa para recuperação de estudos primários em meio eletrônico, publicados nos últimos cinco anos. A RI adota, portanto, a seguinte questão de revisão: "Qual é o estado da arte das taxonomias na ciência da informação no contexto da *web* semântica?"

3.3.1 Critérios de elegibilidade

Foram incluídos estudos primários publicados a partir de 2019, nos idiomas inglês e português, que abordem a relação entre taxonomias e *web* semântica na ciência da informação e áreas correlatas.

A opção pela avaliação de trabalhos publicados nos últimos cinco anos objetiva garantir a atualidade dos estudos incluídos na revisão, o que é especialmente importante em áreas de rápido desenvolvimento como na tecnologia e nas abordagens relacionadas a taxonomias e *web* semântica.

Nesse sentido, acredita-se que a análise de estudos primários finalizados - excluindo-se, portanto, estudos de revisão, livros, teses e dissertações e materiais semelhantes - trará maior confiabilidade e robustez aos achados, bem como delimitará a pesquisa adequando-a aos recursos disponíveis à sua condução.

Também, são considerados os artigos publicados no contexto de eventos científicos, uma vez que nesses espaços são discutidas novas descobertas e inovações. Da mesma forma, são consideradas as áreas que guardam interfaces com a CI, uma vez que algumas discussões possam estar sendo travadas em espaços ainda não ocupados pelos profissionais da CI, mas que são fundamentais para a análise proposta no presente trabalho.

Devido ao amplo uso das taxonomias em biologia, zoologia e na botânica foram excluídos estudos destas áreas e afins.

O Quadro 3 resume os critérios de inclusão e exclusão selecionados para a revisão integrativa, conforme exposto.

Quadro 3. Critérios de elegibilidade adotados na revisão integrativa

Critérios de Inclusão	Critérios de Exclusão
• estudos primários finalizados; • publicados de 2019 a 2023; • nos idiomas inglês e português; • que abordem taxonomias na ciência da informação ou área relacionada; • no contexto da *web* semântica.	• estudos da área da biologia; • estudos da área da zoologia; • estudos da área da botânica;

Fonte: elaborado pelo autor (2023)

3.3.2 Fontes de informação

Foram empreendidas buscas em provedores de serviços e dados especializados como Oasisbr e Scopus. Tais escolhas objetivam garantir o enfoque em assuntos da ciência da informação e áreas correlatas de forma precisa.

O *Open Access and Scholarly Information System* (OASIS.BR) ou *Portal Brasileiro de Publicações e Dados Científicos em Acesso Aberto* (Oasisbr), mantido pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), reúne mais de quatro milhões de documentos com a produção científica publicada em revistas científicas, repositórios digitais de publicações científicas, repositórios digitais de dados de pesquisa e bibliotecas digitais de teses e dissertações do Brasil e de Portugal. Destes, 1.641.704 são artigos científicos e ao todo são indexadas 1.307 revistas científicas, dentre as quais estão a *Revista Ibero-americana de Ciência da Informação*, a *Perspectivas em Ciência da Informação* (*Online*), a *Revista Brasileira de Educação em Ciência da Informação* e a *Revista Digital de Biblioteconomia e Ciência da Informação*. O Oasisbr abrange os seguintes campos: ciência da computação; linguística, letras e artes; ciências sociais aplicadas, especialmente a ciência da informação; engenharia; e multidisciplinares.

Já a base de dados internacional *Scopus*, da editora *Reed Elsevier*, é do tipo referencial com resumos e abrange as áreas de ciência, tecnologia, medicina, ciências sociais e artes e humanidades. Contém mais de 55 milhões de registros e cerca de 22 mil títulos de mais de 5.000 editoras internacionais, incluindo a cobertura de 16.500 revistas. A Scopus indexa diversas revistas científicas, dentre elas as mais importantes relacionadas às temáticas abordadas, a saber: *Cognitive semantics*; *International Journal of Metadata*; *Semantics and Ontologies*; *International Journal of Semantic Computing*; *Journal of Biomedical Semantics*; *Journal of Literary Semantics*; *Journal of Semantics*; *Journal on Data Semantics*; *Natural Language Semantics*; *RUDN Journal of Language Studies Semiotics and Semantics*; *Semantic Web – Interoperability, Usability, Applicability*; *Semantic Web and Information Systems*; *Syntax and Semantics*; *Synthesis Lectures on the Semantic Web: Theory and Technology*; e *Web Semantics*. Além disso, atualmente 343 revistas relacionadas ao campo da Ciência da Informação são indexadas.

3.3.3 Estratégias de busca

A Tabela 2 apresenta a estratégia de busca adotada para cada base de dados, utilizando os termos mapeados para os assuntos taxonomia e *web* semântica, seguindo os critérios de exclusão definidos.

Tabela 2. Estratégia de busca por base de dados

Base de Dados	**Estratégia de Busca**	**Data de execução**	**Resultados**
Oasisbr	taxonom* AND semantic* AND *web** Filtros: “todos os campos”; idiomas inglês e português; tipo de documento artigo, artigo (*working paper*) e artigo de conferência; ilustrado “sem preferência”; e ano da publicação 2019 a 2023.	16/09/2023	98

Base de Dados	Estratégia de Busca	Data de execução	Resultados
Scopus	TITLE-ABS-KEY (taxonom* AND semantic* AND *web**) AND PUBYEAR > 2018 AND PUBYEAR < 2024 AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "MATH") OR LIMIT-TO (SUBJAREA , "ENGI") OR LIMIT-TO (SUBJAREA , "SOCI") OR LIMIT-TO (SUBJAREA , "DECI") OR LIMIT-TO (SUBJAREA , "BUSI") OR LIMIT-TO (SUBJAREA , "MULT")) AND (LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English") OR LIMIT-TO (LANGUAGE , "Portuguese")) AND (LIMIT-TO (SRCTYPE , "p") OR LIMIT-TO (SRCTYPE , "j")) AND (LIMIT-TO (PUBSTAGE , "final"))	16/09/2023	143 + 3 (secundários[18]) = 146

Fonte: elaborado pelo autor (2023)

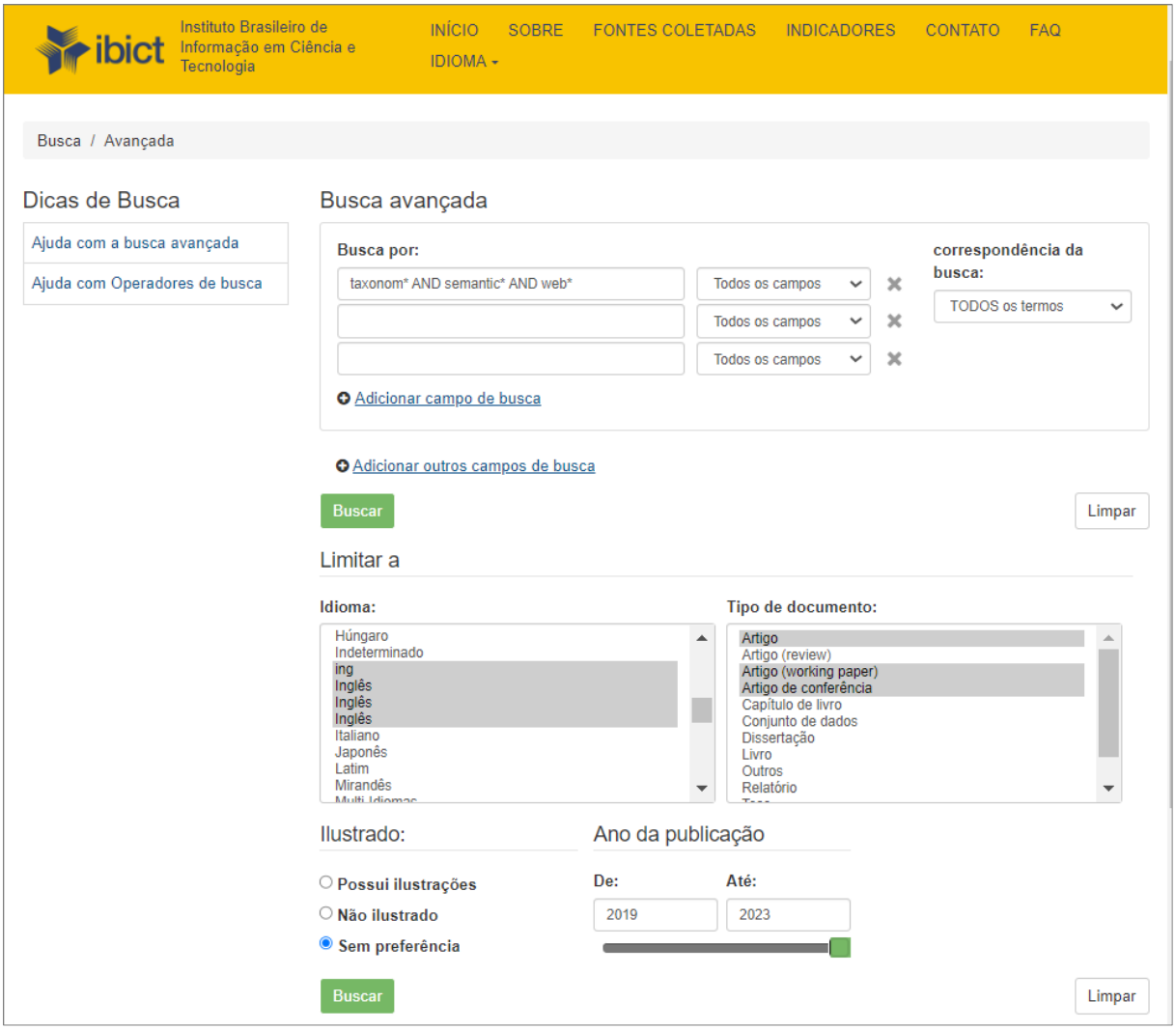
3.3.3.1 Base de dados Oasisbr

Em 16 de setembro de 2023 foi realizado acesso ao Oasisbr por meio do Portal de Periódicos Capes e de acesso ao proxy da UFRJ via Cafe. Foi utilizada inicialmente estratégia de busca simplificada - “taxonom* AND semantic* AND *web**” - com a opção “todos os campos” e filtrados os resultados conforme os critérios de elegibilidade estabelecidos para o presente trabalho. Sendo assim, no campo idioma foram selecionados todos os termos alusivos aos idiomas inglês (por exemplo, “eng” e “ing”)

[18] Registros relacionados à estratégia de busca adotada, extraídos pelos algoritmos da Scopus a partir de uma lista de referências, mas que não estão indexados ou disponíveis nesta base.

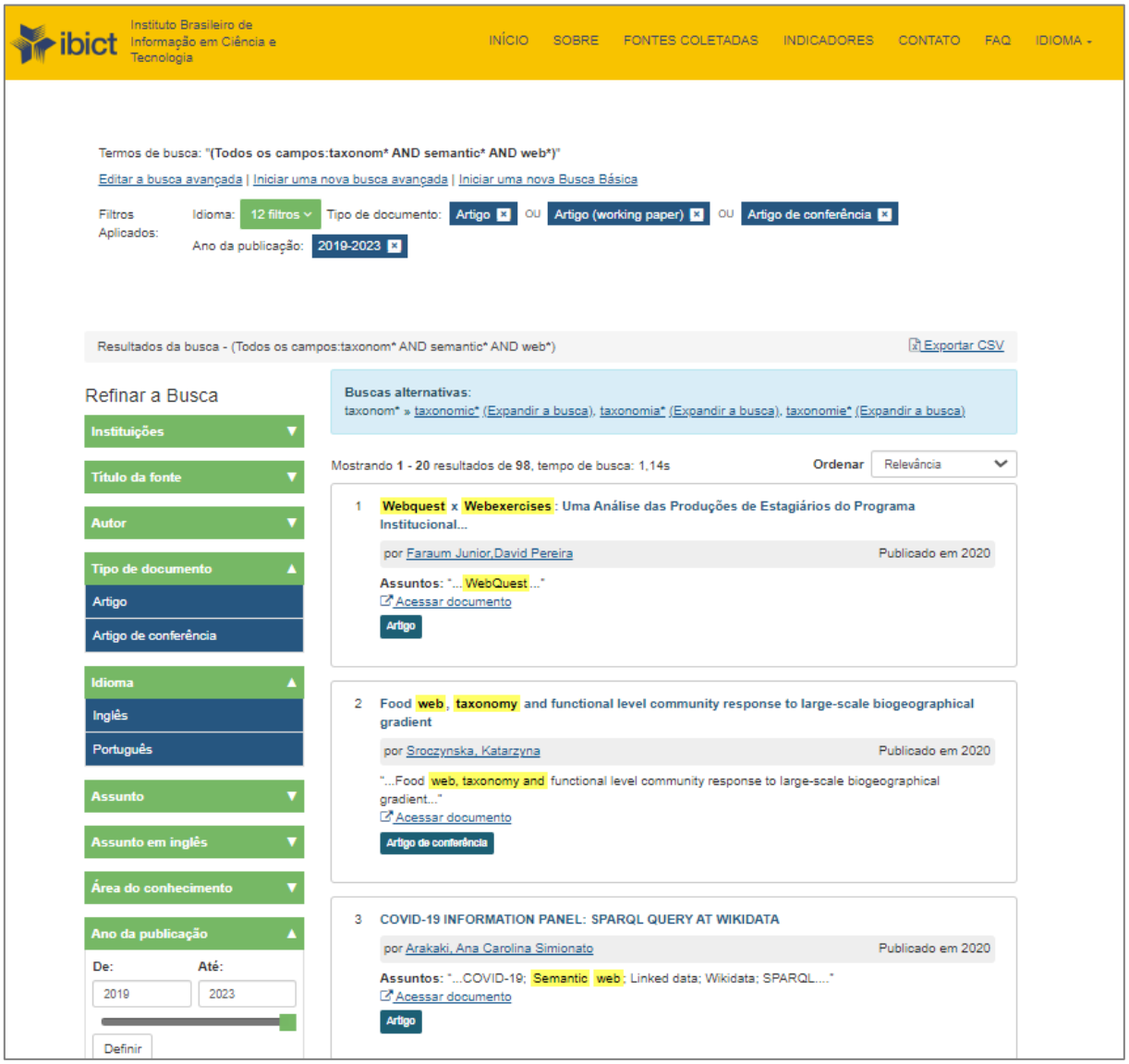
e português, inclusive repetições; no campo tipo de documento foram selecionados artigo, artigo (*working paper*) e artigo de conferência; o campo ilustrado foi marcado como “sem preferência”; e o ano da publicação foi delimitado de 2019 a 2023. A Figura 7 ilustra todos os filtros aplicados para consulta na base Oasisbr, enquanto na Figura 8 pode ser encontrada a visão parcial da tela de consulta com o quantitativo de documentos encontrados.

Figura 7. Tela de refinamento da estratégia de busca na base Oasisbr



Fonte: IBICT (2023)

Figura 8. Visão parcial da tela de consulta na Oasisbr: quantitativo de documentos encontrados



Fonte: IBICT (2023)

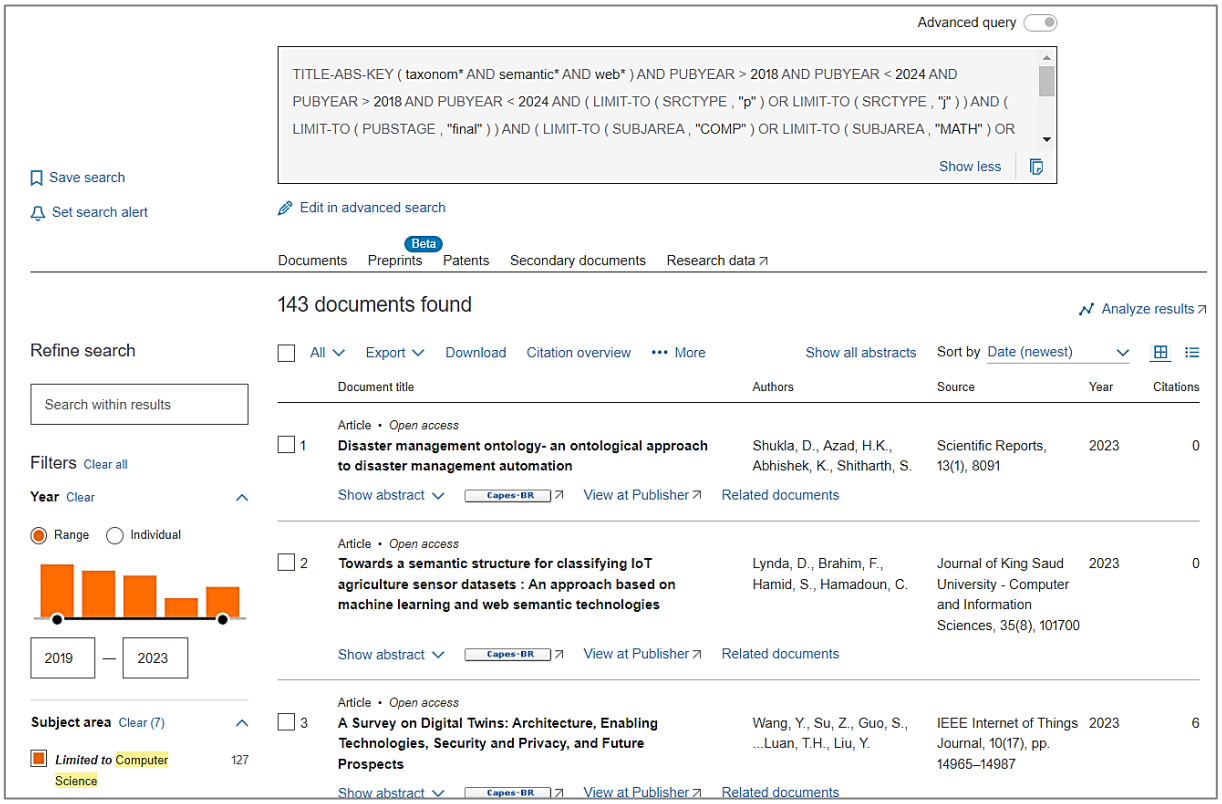
3.3.3.2 Base de dados Scopus

Em 16 de setembro de 2023 foi realizado acesso à base de dados Scopus por meio do Portal de Periódicos Capes e de acesso ao proxy da UFRJ via Cafe. Foi utilizada inicialmente estratégia de busca simplificada - “TITLE-ABS-KEY (taxonom* AND semantic* AND *web**)” - e filtrados os resultados conforme os critérios de elegibilidade estabelecidos para o presente trabalho. Sendo assim, os resultados de busca limitaram-se aos materiais do período de 2019 a 2023 e no campo área temática (*subject area*) foram incluídas somente aquelas afins à temática abordada, são elas:

Computer Science; *Mathematics*; *Engineering*; *Social Sciences*; *Decision Sciences*; *Business, Management and Accounting;* e *Multidisciplinary*. Como consequência foram excluidas todas as demais áreas apresentadas na interface, consideradas áreas que se afastam dos objetivos da revisão proposta, são elas: *Arts and Humanities*; *Biochemistry, Genetics and Molecular Biology*; *Medicine*; *Agricultural and Biological Sciences*; *Earth and Planetary Sciences*; *Materials Science*; *Environmental Science*; *Physics and Astronomy*; *Psychology*; *Economics, Econometrics and Finance*; *Neuroscience*; *Health Professions*; *Pharmacology, Toxicology and Pharmaceutics*; *Energy*; *Chemical Engineering*; *Nursing*; *Immunology and Microbiology*; *Chemistry*; *Physics and Astronomy*; *Neuroscience*; *Health Professions*. Já no campo tipo de documento (*Document type*), foram selecionadas somente as alternativas *Conference paper* e *Article*, de modo a limitar a consulta também a estes filtros, excuindo-se consequentemente os estudos secundários como revisões e capítulos de livros. No campo idioma foram selecionados inglês e português. No campo *Source Type* selecionados *Conference proceeding* e *Journal*, excluindo-se assim as demais alternativas: *Book series* e *Trade journal*. Por fim, no campo *Publication stage* foi selecionada a alternativa *Final*, limitando os resultados a este critério. A Figura 9 apresenta uma visão da tela de consulta na base de dados Scopus, demonstrando a aplicação dos filtros relatados e os resultados obtidos. Enquanto a Figura 10 apresenta uma visão geral dos filtros aplicados, seguindo os critérios de elegibilidade do presente estudo. Fornecer evidências sobre todo o processo de pesquisa, conforme já citado, é apontado pela literatura como fundamental para a confiabilidade do estudo (HIGGINS et al., 2019).

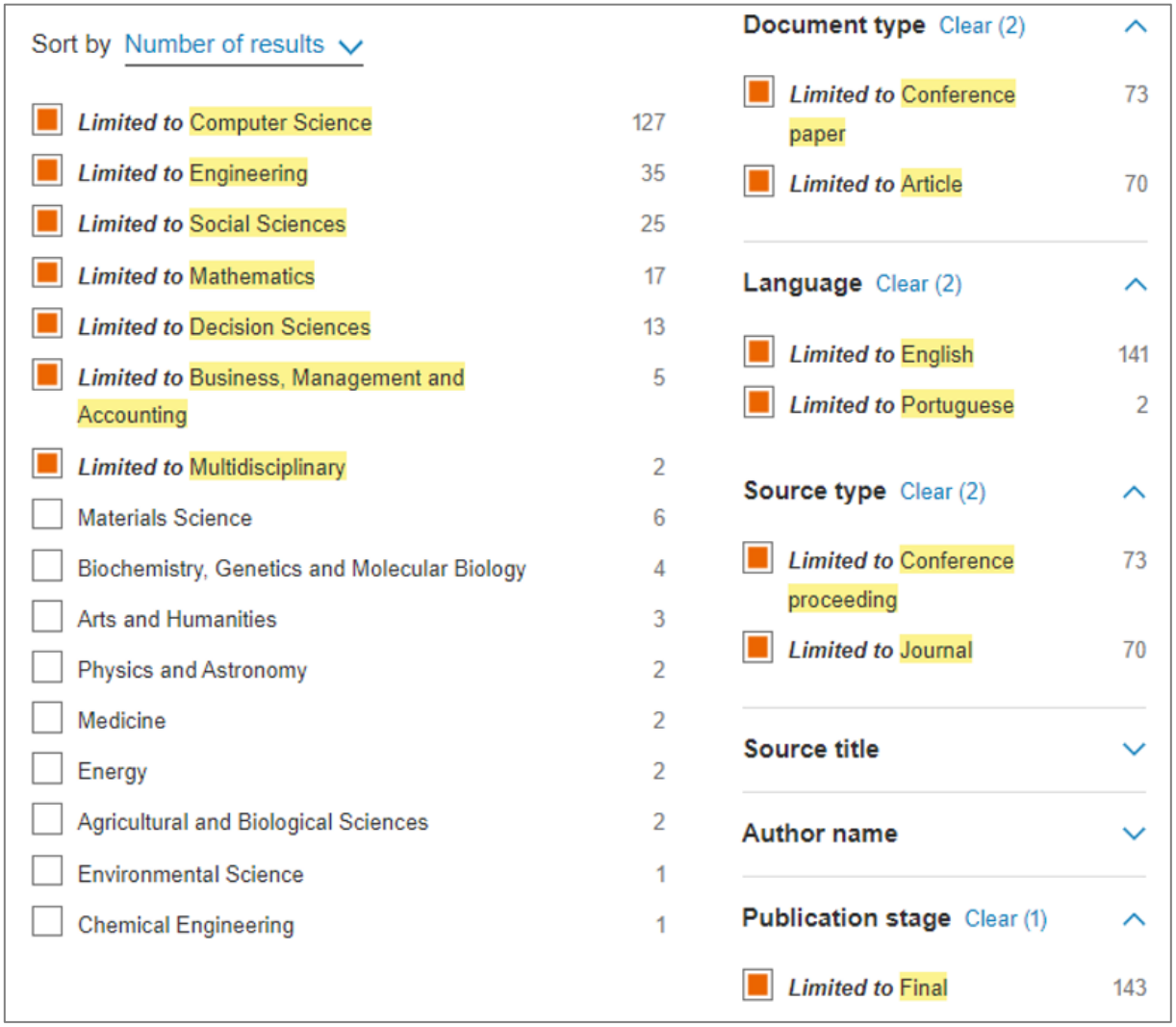
A Scopus oferece a possibilidade de obter registros adicionais relacionados à estratégia de busca adotada que não estão indexados ou disponíveis nesta base. No processo de consulta relatado foram encontrados 3 registros relacionados, denominados secundários, todos foram incluídos para a etapa de *screening*. Há atualmente cerca de 227 milhões de documentos desse tipo na base, extraídos de uma lista de referências.

Figura 9. Visão parcial da tela de consulta na Scopus: quantitativo de documentos encontrados



Fonte: Elsevier (2023)

Figura 10. Visão dos filtros aplicados na tela de consulta na Scopus



Fonte: Elsevier (2023)

Partindo da funcionalidade oferecida pela interface da base de dados Scopus para visualização da ocorrência de palavras-chave relacionadas à consulta executada foi possível criar uma nuvem de palavras, conforme Figura 11. Na nuvem destacam-se a existência de poucos termos não relacionados à presente pesquisa e a presença de termos intimamente relacionados como, traduzidos para o idioma português: semântica, *web* semântica, ontologia, taxonomias, grafos do conhecimento, *sites*, sistemas de processamento de linguagem natural, representação do conhecimento, incorporações, taxonomia, entre outros. Tal análise possibilitou uma rápida avaliação da assertividade da estratégia de busca adotada, especialmente no que diz respeito à escolha dos termos utilizados.

Figura 11. Nuvem de palavras-chave obtidas pela estratégia de busca adotada na base de dados Scopus

Fonte: Worldclouds (2023)

3.3.4 Aspectos metodológicos e resultados preliminares dos processos de avaliação de estudos

A partir deste ponto serão descritos tanto os procedimentos e escolhas metodológicas referentes aos processos de seleção de estudos, coleta de dados, avaliação de riscos e síntese, quanto os resultados preliminares relacionados a tais passos. Conforme já citado, o presente trabalho baseia sua relatoria no modelo recomendado pela *Cochrane Library*, cujo *Relatório Checklist PRISMA* (apêndice A)

orienta sobre as informações que devem compor cada um dos itens de uma pesquisa de revisão para garantia da reprodutibilidade e da confiança nos resultados. O referido *checklist* subdivide-se em sete partes, dentre as quais estão os métodos, os resultados e a discussão, e foi desenvolvido para estudos de RSL do campo da saúde, que frequentemente exigem a elaboração e o registro de um protocolo de revisão que indique a *priori* todas as etapas planejadas para a revisão. Dessa forma, todos os passos metodológicos são relatados de forma separada dos resultados relacionados a estes processos. Tal exigência não se aplica a revisões integrativas, motivo pelo qual é proposta a apresentação dos aspectos metodológicos e dos resultados preliminares dos processos de avaliação dos estudos de forma integrada. O item que trata do processo de seleção, por exemplo, é avaliado no *checklist* PRISMA em dois momentos diferentes: nos métodos (item 9), orienta que sejam descritas as decisões de seleção de estudos sob a ótica dos critérios de inclusão, assim como informações sobre revisores e detalhes das ferramentas de automação usadas no processo; já nos resultados (itens 16a e 16b), orienta que sejam informados os quantitativos de estudos identificados desde a etapa de busca até a seleção dos estudos incluídos na revisão, de preferência com a apresentação do fluxograma PRISMA, conforme protocolo da *Cochrane Library*. Com a nova proposta, adiante neste capítulo são ao mesmo tempo relatados os passos metodológicos e os resultados associados a eles, reservando-se a descrição da análise dos achados para o capítulo dedicado aos resultados da revisão integrativa, que antecede a discussão.

3.3.4.1 Processo de seleção

Após a execução das estratégias de busca adotadas nas bases Oasisbr e Scopus foram identificados 244 estudos no total, os quais foram inseridos no *software web* Covidence, cujo *framework* segue o protocolo Cochrane e por isso é especializado na estruturação de diferentes tipos de revisões sistemáticas. O uso do software possibilitou maior confiabilidade e rapidez na análise dos estudos, (a) com uma interface para estruturação das características do estudo, evidenciando os critérios de inclusão e exclusão, (b) possibilitando o uso de palavras como *highlights* que são destacadas positiva ou negativamente no título e no resumo da lista de registros para facilitar a análise na fase de *screening*, (c) removendo duplicatas automaticamente e (d) fornecendo uma interface *web* de *screening* onde é possível clicar para incluir, excluir ou separar estudos como “talvez”, de modo que já sejam

automaticamente computados no fluxograma PRISMA, também fornecido pelo software. Outras funcionalidades são oferecidas, como a realização da fase de *screening* por dois ou mais revisores independentes gerido pelo sistema, o mesmo para as demais etapas de qualquer tipo de revisão. O Covidence é, assim como o EPPI-Reviewer, ferramenta preferencial da Cochrane para os processos de seleção de estudos. Na presente revisão integrativa foi utilizada a versão gratuita do Covidence apenas para a fase inicial, conforme descrito, que contou com um revisor e duas etapas de análise, de modo a reduzir o risco de vieses.

Na primeira etapa foram lidos título e resumo de cada um dos 237 estudos para *screening*, uma vez que 5 estudos foram excluídos pelo Covidence e 2 identificados durante a análise, assim como observados os critérios de inclusão e exclusão para cada registro, resultando na exclusão de 128 estudos que foram considerados irrelevantes para a presente pesquisa e registrados em planilha no *software* Excel com anotações. Já na segunda etapa a análise se voltou aos 109 estudos selecionados na fase de *screening* no Covidence. Neste ponto cada um dos 109 estudos foi recuperado para uma dupla checagem acerca de sua inclusão na revisão, etapa considerada necessária devido à diversidade de recortes temáticos existentes no objeto taxonomias na *web* semântica e ao desafio de seleção de estudos que contribuam para a pesquisa proposta. Além do título e do resumo, ao recuperar os estudos foi possível realizar rápida leitura para escaneamento da introdução, objetivos, metodologia, resultados e discussão de cada um; computar no fluxograma os de difícil recuperação; realizar tentativas adicionais de recuperação, como solicitações de cópia a autores, instituições de apoio à pesquisa e profissionais da informação; e, principalmente, identificar categorias de análise para posterior elaboração das matrizes de síntese dos achados. Mais uma vez foram considerados os critérios de inclusão e exclusão, que muitas vezes são de difícil identificação apenas por meio do título e do resumo, que nem sempre seguem um padrão científico. O elemento mais prejudicado costuma ser a metodologia, que muitas vezes não é descrita nos resumos dificultando a diferenciação entre estudos primários, secundários e mistos.

Uma segunda etapa de análise ainda na fase de *screening* possibilitou também diferenciar estudos que abordavam, por exemplo, taxonomias no contexto da construção ou uso de ontologias dos estudos que abordavam somente aspectos referentes à complexidade das ontologias (sem abordar taxonomias) ou à

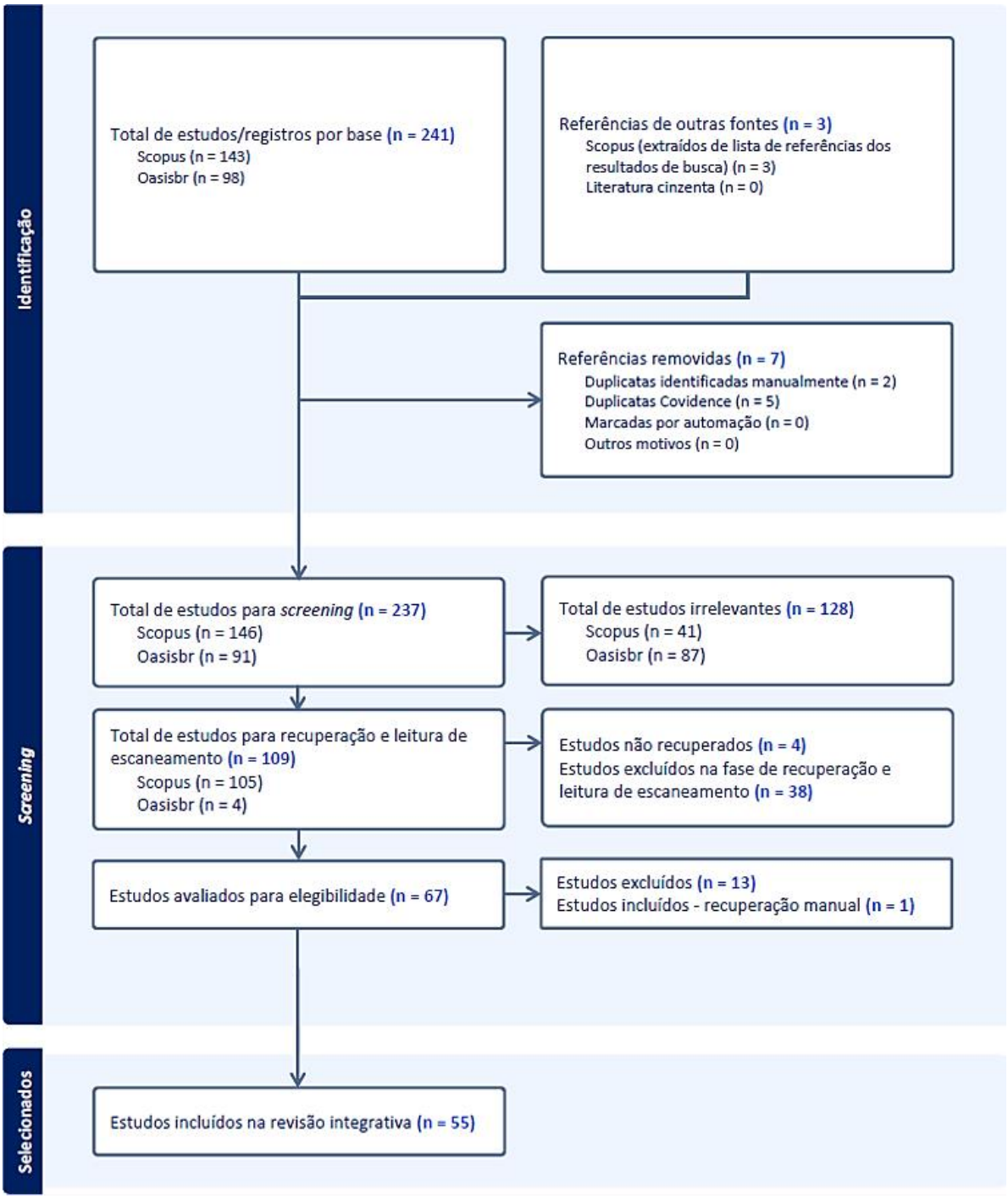
identificação de relações não-taxonômicas. Desse modo, foram excluídos estudos que abordaram exclusivamente relações não-taxonômicas - aquelas que não se encaixam em uma estrutura hierárquica ou de classificação, característica das taxonomias - por estarem mais relacionadas à modelagem de ontologias na *web* semântica, numa abordagem semântica mais profunda do que as encontradas nas taxonomias. No entanto, para evitar a perda de estudos com potencial de contribuição para a pesquisa, foram incluídos todos os estudos sobre ontologias que também abordaram aspectos relativos às taxonomias.

Da mesma forma, o presente trabalho, a fim de se ater à investigação sobre o papel das taxonomias na *web* semântica, não considerou aqueles estudos que apenas citaram a utilização de uma ou mais taxonomias sem que a temática fosse abordada na perspectiva da CI ou áreas correlatas, como a ciência da computação. Exemplos disso são os estudos que utilizaram taxonomias educacionais, como a taxonomia de Bloom; taxonomias do campo da biologia; e taxonomias prontas disponíveis como ferramentas *web* como WordNet[19] e DBPedia[20]. Por outro lado, se fossem encontrados estudos que analisassem tais taxonomias eles seriam então incluídos na presente revisão. Foram selecionados 67 estudos para leitura completa, cujo conteúdo indica potencial contribuição para a discussão proposta e cuja lista completa é apresentada no APÊNDICE B. Estudos avaliados para elegibilidade. A Figura 12 apresenta o fluxograma PRISMA, que resume cada uma das etapas descritas.

[19] WordNet é base de dados lexical para a língua inglesa disponível na *web* sob o endereço “wordnet.princeton.edu”, onde é possível encontrar conjuntos de sinônimos e suas relações semânticas. A base funciona a partir de uma taxonomia.

[20] A DBpedia é uma ferramenta *web* de dados abertos conectados ou *Linked Open Data* (LOD), que extrai dados estruturados da enciclopédia online Wikipedia e os disponibiliza em formato aberto legível por máquinas, como grafo do conhecimento ou *Knowledge Graph* (KG).

Figura 12. Fluxograma PRISMA da Revisão Integrativa sobre taxonomias na *web* semântica



Fonte: elaborado pelo autor (2023)

3.3.4.2 Processos de coleta e análise de dados

A etapa de coleta de dados teve sua primeira etapa realizada ainda na fase de *screening*, na dupla checagem dos estudos selecionados feita por meio da leitura de

objetivos, metodologia, resultados e discussão. A análise prévia realizada possibilitou reunir os conhecimentos obtidos durante a revisão narrativa com os primeiros achados da pesquisa, o que deu origem a uma taxonomia construída com base no conteúdo encontrado nos 67 estudos avaliados para elegibilidade. A taxonomia tem como objetivo ser um primeiro instrumento de organização intelectual do conteúdo encontrado face a um *corpus* de conhecimento revelado pela revisão narrativa. De modo a obter uma taxonomia concisa, preservando a facilidade leitura, a praticidade de utilização para a finalidade acadêmica aqui pretendida e a clareza dos conceitos, foram selecionados apenas os conceitos mais amplos, sendo encontrados poucos níveis de profundidade em sua estrutura. Devido à já referida contribuição que a presente pesquisa espera conferir à comunidade acadêmica, com a produção científica em língua portuguesa sobre a temática das taxonomias e da *web* semântica, cada conceito é apresentado nos idiomas português e inglês na taxonomia apresentada na Figura 13.

Sabe-se que é recomendado o uso de letras minúsculas nos termos de uma taxonomia, que deve seguir diretrizes descritas nas normas ANSI/NISO Z39.19-2010R e ISO 25964-2:2013. No entanto, no presente contexto acadêmico serão privilegiados os seguintes aspectos facilitadores: (a) uso de siglas comuns nos assuntos abordados na revisão integrativa, apresentando sua tradução; (b) visualização estética com o auxílio de barras e espaços entre os termos em português e em inglês; (c) uso de mais de duas palavras em um mesmo termo para preservar a ideia geral contida nos termos subordinados; (d) ordenação e estrutura hierárquica selecionada conforme o objetivo de melhor compreensão da forma como os estudos apontam o relacionamento entre taxonomias e *web* semântica.

Conforme apontado por Toronto e Remington (2020), é fundamental numa revisão integrativa que cada estudo recuperado seja avaliado quanto à qualidade metodológica. Desafios importantes podem se colocar no sentido da avaliação de uma diversidade de estudos. Um exemplo são os classificados como teóricos, que se concentram na articulação de ideias e na demonstração do ponto de vista dos autores. Nesses casos, é recomendado que sejam verificados outros fatores como a adequação lógica, ou a estrutura da teoria para identificar se as conclusões e as evidências nas quais se baseiam fazem sentido. Nesse sentido, uma ferramenta de avaliação crítica pode ser escolhida como instrumento de avaliação crítica ou

adaptada conforme as características da pesquisa. A ferramenta adaptada no presente estudo foi a *CASP Qualitative Checklist* (2018), da *Critical Appraisal Skills Program* (CASP), específica para estudos qualitativos, conforme indicado por Toronto e Remington (2020). Desse modo, as dez perguntas verificadas pelo *checklist* foram adaptadas como modelo de relatoria para a coluna “metodologia” na matriz de análise, tendo sido observados: (a) se o estudo possui uma declaração clara dos objetivos; (b) se a metodologia estava adequada à proposta; (c) se há inconsistência nas medidas adotadas; (d) se há descrição metodológica adequada e transparência nos relatos/ética; (e) se os questionários que por ventura foram elaborados continham em sua maioria questões que encaminhavam aos objetivos do estudo; (f) se as referências citadas eram suficientes para abordar o objeto proposto; e (g) possíveis vieses. Outras colunas da matriz de análise também contribuíram para a coleta com base no instrumento original, que foi impresso e auxiliou a coleta de dados de cada estudo recuperado; a adaptação das informações do instrumento para a matriz de análise de dados é apresentada no anexo 1.

A coleta de dados dos estudos foi realizada por um revisor, que extraiu os objetivos de cada estudo, os aspectos metodológicos, os resultados, os conceitos encontrados, o nome de cada *framework* ou ferramenta citada e os modelos ou métodos analisados ou propostos, assim como avaliou a aplicabilidade dos estudos, organizou um glossário e classificou cada registro quanto à qualidade do estudo (razoável, boa, muito boa ou excelente), quanto às contribuições para a compreensão do papel das taxonomias na *web* semântica e quanto a abordagem utilizada para criar, gerir e/ou utilizar taxonomias na *web* semântica. Os estudos excluídos na etapa de extração de dados foram justificados na mesma matriz de análise. Versões da matriz de análise utilizada podem ser encontradas no APÊNDICE B. Estudos avaliados para elegibilidade e qualidade, no APÊNDICE C. Contextualização conceitual dos estudos incluídos na revisão integrativa, no APÊNDICE D. Resultados e aplicabilidade dos estudos incluídos na revisão integrativa e no APÊNDICE E. Abordagens utilizadas para criação, gestão e/ou uso de taxonomias na *web* semântica nos últimos cinco anos.

a) Coleta de dados com apoio de inteligência artificial: limitações e potencialidades para a pesquisa qualitativa

O uso de inteligência artificial (IA) já se tornou comum em atividades de diversos

setores, não é diferente com a área acadêmica, que vem observando o crescimento de aplicações com o propósito de auxiliar processos de pesquisa, organização e construção de conhecimento a partir da produção científica disponível. Assim, a presente pesquisa contou no processo de coleta de dados com o apoio da ferramenta *web* SciSpace, uma aplicação específica de IA para revisões de literatura que se difere de outras mais generalistas como os conhecidos ChatGPT da OpenAI e o Bard da Google. Apenas a funcionalidade denominada *Copilot - Read with AI* foi utilizada no auxílio à análise dos estudos em PDF de leitura completa. A ferramenta possibilita a realização de perguntas, que são respondidas por uma IA com base no artigo que está sendo analisado, no conjunto de artigos organizados na pasta pessoal do pesquisador na ferramenta, com base em toda a literatura presente no *corpus* SciSpace, ou em todas essas fontes simultaneamente. Além disso, uma matriz de análise é preenchida pela inteligência artificial conforme colunas são propostas pelo pesquisador, tais como resumo do estudo, principais achados, metodologia adotada, entre outras possibilidades. O relato de uso da ferramenta é apresentado abaixo, baseado na condução da presente pesquisa de característica qualitativa, teórica, descritiva e documental, uma revisão integrativa de estudos primários. Experiências de uso de outras IA disponíveis no mercado podem apresentar importantes diferenças pelo fato de possuírem níveis de desenvolvimento, escopo, propósitos e funcionalidades distintas.

A experiência de uso da SciSpace na análise do conteúdo dos estudos demonstrou riscos e potencialidades dessa abordagem para pesquisas em diferentes áreas. Em consultas que consideraram toda a base de dados da ferramenta foi possível aprender ou esclarecer dúvidas e diferenças sobre conceitos bem estabelecidos do campo da ciência da computação, com informações passíveis de verificação em literatura da área e outras fontes. As respostas geradas forneceram também sinônimos úteis à redação das sínteses construídas para preenchimento da matriz de análise e advindas das anotações feitas durante a leitura completa dos estudos impressos. Por outro lado, o preenchimento da matriz pela máquina, em comparação com a matriz de análise do pesquisador (obtida a partir da leitura completa dos estudos), demonstrou que: (a) algumas das colunas são indicadas com “N/A” ou não disponível, mesmo nos casos em que a resposta se revela no estudo pela união de duas ou mais afirmações; (b) há limitações quanto à disponibilidade da

máquina, já que algumas colunas são preenchidas com a mensagem "*o copiloto não conseguiu gerar a resposta. Por favor, tente novamente depois de algum tempo*"; (c) algumas traduções feitas não condizem com o termo mais comumente adotado para determinado conceito; (d) a máquina não foi capaz de captar a metodologia, os procedimentos ou as camadas metodológicas, que são de difícil identificação até mesmo para o pesquisador quando não são claramente dispostas no texto; e, principalmente, que (e) algumas respostas dadas não refletem um ou mais aspectos do estudo, tais como objeto, assunto, abordagem utilizada, métodos aplicados *versus* analisados pelo estudo, sendo comum o fornecimento de respostas geradas pela mistura de informações extraídas de diferentes partes do documento, como descrição de trabalhos relacionados, embasamento teórico e desenho do estudo.

Cabe esclarecer que não há atualmente uma forma automatizada de extração da matriz gerada pelo SciSpace e que a construção de colunas apresenta limitações quanto ao que a máquina é capaz de compreender sobre o que se pretende responder com a coluna escolhida. Dessa forma, a matriz gerada pela ferramenta não é capaz de substituir de nenhuma forma a matriz de análise de uma pesquisa. A ferramenta é capaz apenas de apoiar timidamente o processo de compreensão daquilo que é abordado no estudo e somente será descoberto pela leitura completa e cuidadosa feita pelo pesquisador, que é também o principal responsável por propor os campos da matriz de acordo com a pesquisa empreendida. Um exemplo prático vivenciado na presente pesquisa se deu no âmbito do surgimento do conceito em inglês *embeddings*, do domínio da mineração de dados, que é comumente traduzido pela ferramenta como incorporações, termo que não reflete o verdadeiro conceito no contexto da temática pesquisada e apresenta nuances. Na verdade, o termo se refere ao conceito de vetores, uma tradução mais adequada e que se difere do conceito de incorporação, que é comum no mesmo domínio e diz respeito à acoplagem de um objeto digital de qualquer tipo num ambiente ou documento na *web*. Por outro lado, quando diretamente associados aos grafos do conhecimento é possível encontrar o termo *embeddings vectors* ou incorporação de vetores, que reúne as duas ideias.

Nesse sentido, uma importante questão que se coloca em relação ao uso de inteligências artificiais em diferentes contextos, seja a SciSpace ou concorrentes, é a nítida incapacidade de compreensão semântica em nível aprofundado e que atualmente está disponível apenas ao poder da mente humana. Não são raros os

casos de inferências equivocadas, que levam a conclusões que não refletem o que está sendo abordado de fato pelo estudo. É fundamental compreender que toda e qualquer resposta dada precisa ser questionada pelo agente humano que a interpreta, o que pode ser feito por meio da checagem em fontes confiáveis combinadas com a realização de perguntas de controle, com vistas a confirmar se numa relação de oposição, por exemplo, a coerência da resposta será mantida. O mesmo pode ser dito em relação à tradução obtida por estes meios, que nem sempre apresenta os melhores termos para representar determinados conceitos. Não faz parte do escopo do presente estudo o aprofundamento de tal análise, porém considerou-se ao relatar essas experiências as possíveis contribuições a uma nova perspectiva metodológica em desenvolvimento, além da transparência do trabalho.

b) Método de Análise de dados

Segundo Toronto e Remington (2020), a forma como os dados serão analisados deve ser a primeira questão para um pesquisador que opta pela condução de uma revisão do tipo integrativa, isto porque as decisões metodológicas acerca da análise contribuirão expressivamente para o rigor adotado na pesquisa. A análise demanda a ordenação, codificação e categorização de dados de múltiplas fontes e não há um método recomendado para isso. Uma abordagem comumente utilizada em pesquisas qualitativas é a análise temática, adotada na presente revisão e considerada um método flexível e adequado à identificação de padrões nos dados (Toronto eRemington, 2020). Rosa e Mackedanz (2021) apontam que a análise temática se associa ao texto e não necessariamente ao referencial teórico para definição dos temas, que podem ser predefinidos pelo pesquisador, possibilitando uma análise ampla e ao mesmo tempo comparativa, gerando um conjunto rico e detalhado de informações. Os autores concordam com Braun e Clarke (2006) ao considerarem que a análise temática contém as habilidades essenciais à realização de outras formas de análise qualitativa, devido à abrangência de recursos envolvidos nos processos de organização, análise e relatoria dos dados. Seis fases compõem a análise temática, conforme descrito no Quadro 4. O quadro tem por objetivo localizar o momento da pesquisa onde foi realizada cada fase da análise temática, que conforme citado se distribui de forma abrangente por diferentes etapas demandando ampla gama de recursos (Braun; Clarke, 2006; Rosa; Mackedanz, 2021; Toronto; Remington, 2020).

Quadro 4. Processo de análise temática por etapa da revisão integrativa

Fase da análise temática	Etapa do estudo
1. Familiarização com os dados	Processo de seleção > *screening* em duas etapas: (1) classificação com base na leitura de título e resumo; (2) recuperação de estudos selecionados e leitura de escaneamento para definição dos estudos elegíveis.
2. Geração dos códigos iniciais	Etapas 1 e 2 do *screening.*
3. Busca por temas	Planejamento da coleta de dados
4. Revisão dos temas	Coleta (extração) e análise de dados
5. Definição e denominação dos temas	Análise de dados, relatoria do processo de pesquisa e elaboração da taxonomia da revisão integrativa
6. Produção do relatório	Síntese dos estudos, finalização da relatoria do processo de pesquisa e apresentação dos resultados

Fonte: elaborado pelo autor (2023)

Cabe esclarecer que as seis fases apresentadas para a análise temática não são exclusivas desta abordagem. Da mesma forma, não há um guia que indique o modelo a ser adotado para a estruturação da relatoria do processo de pesquisa e da apresentação dos resultados (Toronto; Remington, 2020).

Na presente revisão a análise temática se apresenta como um método essencialista, isto é, relata as experiências e os significados com base na verdade objetiva descrita nos estudos, a realidade. A escolha dos temas se deu por indução ou guiada pelos dados, numa abordagem semântica, onde os temas são identificados conforme explicitados pelos dados contidos nos estudos incluídos na revisão, não por meio de suposições. Buscou-se encontrar padrões frequentes de significado (Braun; Clarke, 2006; Rosa; Mackedanz, 2021; Toronto; Remington, 2020).

3.3.4.3 Avaliação de riscos de vieses e confiabilidade dos estudos

Quanto à avaliação de vieses, considerando que a revisão integrativa incluiu somente estudos primários, todo o processo de pesquisa buscou avaliar a qualidade metodológica e a validade dos estudos de forma abrangente e cuidadosa, de modo a garantir a inclusão de estudos confiáveis e representativos, evitando distorções nos

resultados ou conclusões da revisão. Apesar de a revisão ter sido conduzida por apenas um revisor, entende-se que as decisões tomadas e forma de relatoria contribuíram para mitigar tais riscos, dentre elas o uso de um software de apoio, a inclusão de uma etapa adicional de análise na fase de *screening* e a transparência dos métodos e procedimentos.

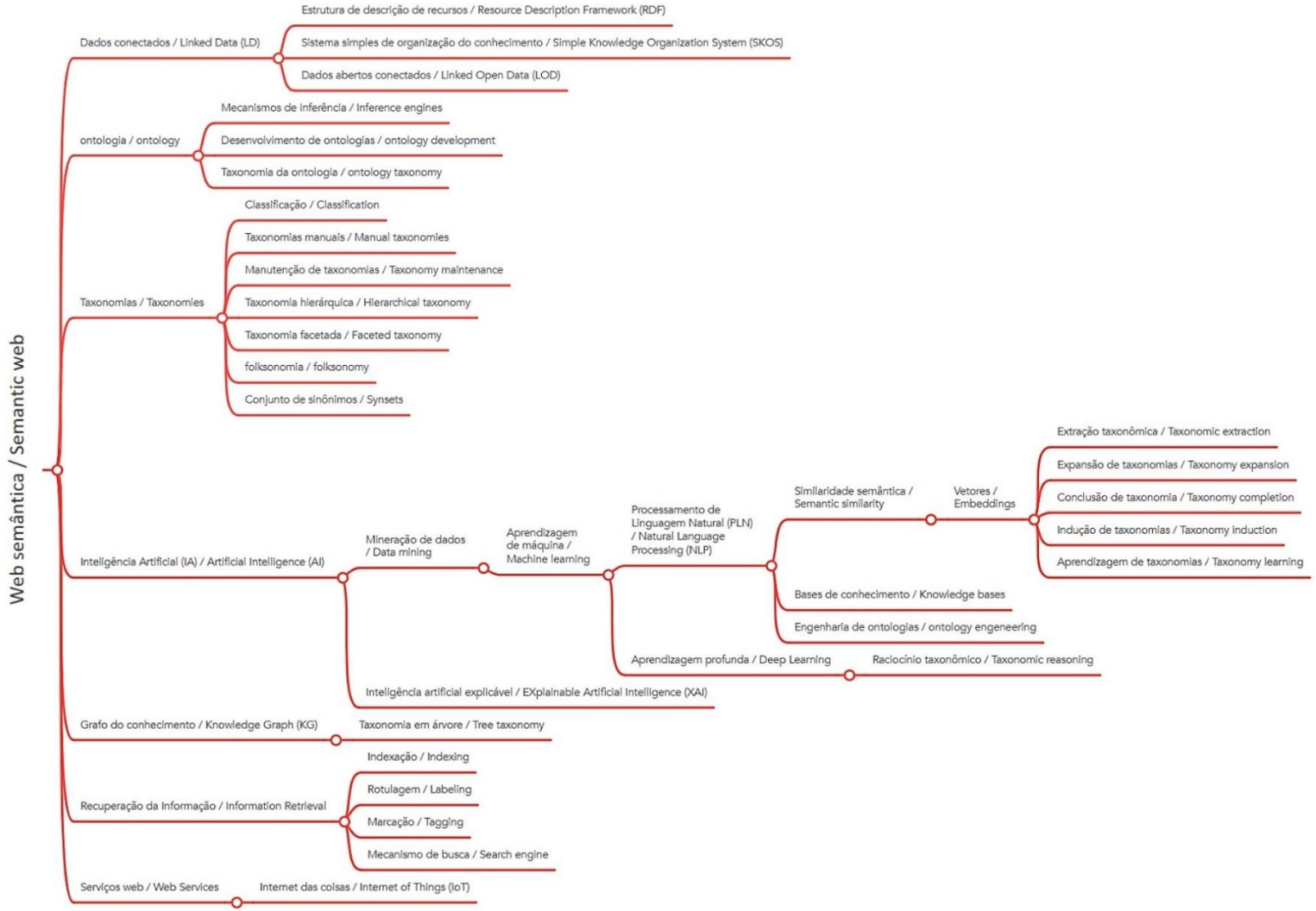
3.3.4.4 Métodos de síntese

Seguindo o método de análise e a taxonomia desenvolvida com base na revisão narrativa e nos achados da RI, foi adotada a síntese temática para a presente pesquisa. Os conceitos da taxonomia foram agregados e sintetizadas em tópicos relacionados e mais abrangentes conforme Quadro 5, do menor grau de incidência do poder computacional ao mais elevado, com vistas a explicitar novos conceitos surgidos do *corpus* da pesquisa e estabelecer uma análise de pretensão evolutiva.

Quadro 5. Categorias selecionadas para apresentação das sínteses da RI

Classificação para contextualizar, compreender e recuperar informações: categoria que agrega os conceitos classificação, taxonomia da ontologia, taxonomia facetada, taxonomia hierárquica, taxonomias manuais, ontologia e similaridade semântica.
Bases de conhecimento (*knowledge bases*) e vetores de palavras (*embeddings*)
Grafos do conhecimento (*knowledge graphs*)
Linked (Open) Data: categoria que congrega os conceitos de LD e LOD, selecionados como termos desta categoria ao invés de sua forma em português pelo impacto de sua representação enquanto tecnologia e objetivo fim da *web* semântica.
Inteligência Artificial: categoria que compreende os conceitos de aprendizagem de máquina; aprendizagem profunda; aprendizagem de taxonomias; conclusão, expansão e manutenção de taxonomias; extração taxonômica; engenharia de ontologias; PLN e mineração de dados.
Serviços *Web* e IoT

Figura 13. Taxonomia da revisão integrativa sobre taxonomias na *web* semântica (2019-2023)



Fonte: elaborado pelo autor (2023)

4 RESULTADOS: REVISÃO INTEGRATIVA

Este capítulo se destina à apresentação dos resultados da revisão integrativa, que recuperou a produção científica disponível nos últimos cinco anos para atualizar e sistematizar o estado da arte das taxonomias na ciência da informação e sua relação com a *web* semântica.

4.1 CARACTERÍSTICAS DOS ESTUDOS

Conforme apresentado no fluxograma PRISMA, 109 estudos foram selecionados na etapa de *screening* para uma primeira leitura de escaneamento e seleção de elegíveis à leitura completa. Destes, 105 eram da base Scopus e quatro da Oasisbr. Apenas quatro não foram recuperados e 38 estudos não atendiam a pelo menos um dos critérios de elegibilidade. Dessa forma, 67 estudos foram avaliados por meio de leitura completa, tendo sido excluídos 13, devidamente justificados no APÊNDICE B, e incluído um estudo manualmente que não integrava os registros obtidos na execução das estratégias de busca. Destaca-se o fato de todos os estudos incluídos na RI pertencerem à base Scopus, o que pode estar relacionado ao uso de algoritmos avançados de classificação nesta base. A Tabela 3 resume esses totais e Tabela 4 apresenta o total por idioma e fonte de informação com enfoque específico nos 55 estudos finalmente incluídos na RI; somente um estudo do total no idioma Português e todos os demais Inglês. Da mesma forma, o Gráfico 1 aponta os totais por ano, onde aparece 2020, início da pandemia COVID-19, com o maior número de estudos, o que pode ser explicado pela publicação de estudos redigidos nos anos anteriores. Observa-se queda acentuada nos anos de 2021 e 2022, quando ainda havia medidas sanitárias restritivas, e retomada no ano 2023 quando é decretado o fim da pandemia pelas autoridades mundiais.

Quanto à distribuição dos estudos no mundo, considerou-se o país de atuação dos autores ou da instituição ao qual são vinculados, indicando possíveis cooperações conforme apresentado na Tabela 5. O Quadro 6 resume a frequência em que cada país aparece organizados por continente. Destaca-se a concentração de autores na Europa, seguida da América do Norte e da Ásia. Apesar disso, EUA e China aparecem com maior frequência, o que corresponde à forma como observamos a corrida tecnológica no mundo, com o domínio destes dois países.

Tabela 3. Total de estudos avaliados por fonte de informação

Avaliação	Base Oasisbr	Base Scopus	Total
Não atende	2	36	38
Não recuperado	0	4	4
Incluído	0	54	54
Excluído	2	11	13
Incluído (adicionado)	0	0	1
Total	**4**	**105**	**110**

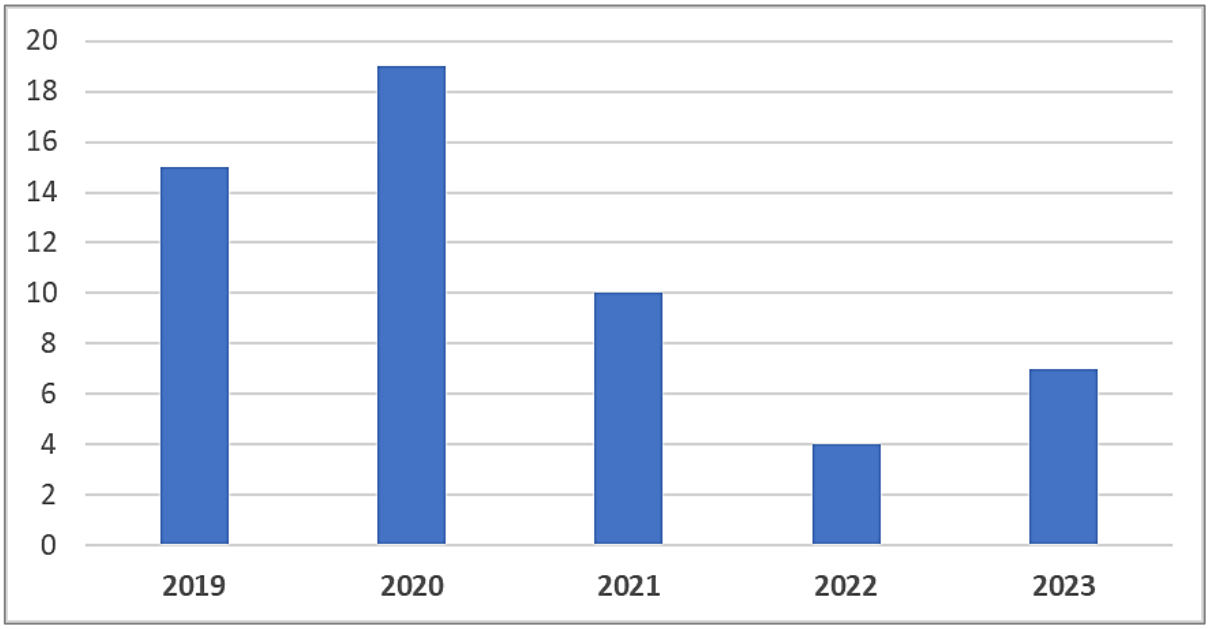
Fonte: elaborado pelo autor (2023)

Tabela 4. Total de estudos incluídos por fonte de informação e idioma

Idioma	Base Oasisbr	Base Scopus	Total
Inglês	**0**	**53**	**54**
Incluído	0	53	53
Incluído (adicionado)	0	0	1
Português (BR)	**0**	**1**	**1**
Incluído	0	1	1
Total	**0**	**54**	**55**

Fonte: elaborado pelo autor (2023)

Gráfico 1. Total de estudos incluídos na revisão integrativa por ano



Fonte: elaborado pelo autor (2023)

Tabela 5. Total de estudos incluídos na RI por ano e país(es)

Ano/País	Total de estudos
2019	**15**
Belarus	1
Chile; Alemanha	1
China	5
EUA	3
India	1
índia	1
Irlanda	1
Irlanda e China	1
Malásia/Paquistão	1
2020	**19**
Alemanha	2
Brasil	3
Canadá	1
EUA	5
EUA e China	1
Irlanda	1
Itália	3
Reino Unido e Alemanha	1
Tailândia	1
Ucrânia	1
2021	**10**
Austrália	1
Canadá	1
China	1
EUA	3
França	1
Irlanda	1
Itália	1
Noruega	1
2022	**4**
China	2
EUA	1
EUA; Canadá	1
2023	**7**
Austrália	2
índia	1
Irlanda	1
Irlanda, Holanda e França	1
Itália	1
Suíça	1
Total	**55**

Fonte: elaborado pelo autor (2023)

Quadro 6. Localização dos autores dos estudos incluídos na RI no mundo

América do Norte	**17**	**América do Sul**	**4**
EUA	14	Brasil	3
Canadá	3	Chile	1
Europa	**23**	**Ásia**	**16**
Irlanda	6	China	10
Itália	5	Índia	3
Alemanha	4	Malásia	1
França	2	Paquistão	1
Belarus	1	Tailândia	1
Holanda	1		
Reino Unido	1	**Oceania**	**3**
Suíça	1	Austrália	3
Noruega	1		
Ucrânia	1		

Fonte: elaborado pelo autor (2023)

Por fim, cabe observar a distribuição conceitual dos estudos incluídos por ano e área identificada nos estudos, conforme demonstrado na Tabela 6. Quando não especificada no estudo, mesmo em caso de variação terminológica, buscou-se padronizar a terminologia e estimar aqueles não informados por meio do currículo do autor. Já os conceitos se referem aos adotados na taxonomia da RI.

Tabela 6. Distribuição conceitual dos estudos incluídos na RI por ano e área identificada

Ano/Área/Conceito na taxonomia da RI	**Total de estudos incluídos**
2019	**15**
CC	**6**
Aprendizagem de máquina	2
KG	1
Mecanismo de busca	1
ontologia	1
Vetores	1
CI	**1**
Mecanismo de busca	1
Ciência de Dados	**1**
KB	1
Comunicação	**2**
Classificação	1
ontologia	1
Economia e CC	**1**
PLN	1
Engenharia	**2**

Ano/Área/Conceito na taxonomia da RI	Total de estudos incluídos
ontologia	1
PLN	1
Engenharia; Informática	**1**
ontologia	1
Linguística computacional	**1**
KG	1
2020	**19**
Automação industrial; Engenharia elétrica e eletrônica	**1**
ontologia	1
Biblioteconomia e CI	**1**
Taxonomia facetada	1
CC	**11**
Engenharia de ontologias	1
Expansão de taxonomias	2
Extração taxonômica	1
KG	1
LD	1
Mecanismo de busca	1
ontologia	1
Serviços *web*	1
Taxonomia da ontologia	1
Taxonomia hierárquica	1
Ciência de Dados	**1**
Mineração de dados	1
Comunicação	**1**
Aprendizagem profunda	1
Content *web*; Social media	**1**
Aprendizagem de máquina	1
Engenharia Elétrica e Informática	**1**
LD	1
Estatística; Informática; CC	**2**
Aprendizagem profunda	1
Vetores	1
2021	**10**
Bioinformática	**1**
Aprendizagem de máquina	1
CC	**4**
Aprendizagem de taxonomias	1
IoT	1
Taxonomia da ontologia	1
Taxonomias manuais	1
CC; Engenharia	**1**
Manutenção de taxonomias	1
Ciência de Dados	**1**
LD	1
Engenharia do Conhecimento	**1**
KG	1

Ano/Área/Conceito na taxonomia da RI	Total de estudos incluídos
Estatística; Informática; CC	**1**
Vetores	1
Linguística	**1**
Aprendizagem profunda	1
2022	**4**
CC	**1**
Conclusão de taxonomia	1
Engenharia da Informação	**1**
Aprendizagem profunda	1
Engenharia e Informática	**1**
Similaridade semântica	1
Medicina e Estatística	**1**
ontologia	1
2023	**7**
CC	**5**
LD	1
LOD	1
ontologia	1
PLN	1
Similaridade semântica	1
Ciência de Dados	**1**
Aprendizagem profunda	1
Engenharia	**1**
Conclusão de taxonomia	1
Total	**55**

Fonte: elaborado pelo autor (2023)

4.2 SÍNTESES

A seguir é apresentada a síntese de cada estudo incluído na RI numa ordenação temática seguida de cronológica. Busca-se, com isso, apresentar evidências sobre o estado da arte das taxonomias na *web* semântica nos últimos cinco anos, indicando a trajetória dessa relação, as convergências entre as temáticas e os aspectos interdisciplinares.

4.2.1 Classificação para contextualizar, compreender e recuperar informações

Do ponto de vista teórico destacam-se dois estudos que tem como enfoque a representação do conhecimento. São eles, Patel et al.(2019), que objetiva explicitar a visão de dados da *web* semântica como unidades de conhecimento, e Poirier (2019), que discute como os projetistas de infraestruturas semióticas – muitas vezes treinados

pelas lentes da lógica formal – lidam com os desafios crescentes que se colocam à representação do conhecimento em meio à diversidade e às diferenças socioculturais existentes no mundo e refletidas na *web*.

Patel et al.(2019), utiliza um estudo de situações emergenciais como base para explicar sua teoria e analisa as várias facetas de um conceito à medida que o reduz a uma unidade atômica de conhecimento, propondo uma discussão profunda sobre os critérios que um dado precisa atender para exercer efetivamente seu papel na *web* semântica ou *web* inteligente, ou seja, ser lido e compreendido por humanos e máquinas. Busca demonstrar a eficácia das ontologias na organização e na representação do conhecimento, exaltando aspectos de **atribuição de conceitos, estruturação hierárquica e mecanismos de herança**, característicos das taxonomias. Os autores apontam dois principais benefícios do conhecimento taxonômico: **redução de redundâncias** e a **redução de complexidade** nos mecanismos de busca. Tais funções se estendem aos domínios e aos conjuntos de dados de diferentes naturezas e tamanhos, conforme será exposto adiante.

Patel et al. (2019) dialoga com Poirier (2019) ao afirmar que a base de conhecimento da *web* semântica deve ser o mais generalizada possível para comportar a natureza dinâmica das informações globais. Ao mesmo tempo propõe, como Poirier (2019), aprofundar a **capacidade de conceituação das diferenças com maior riqueza de detalhes**. A autora compreende que a sensibilidade humana é a chave para os desafios impostos pelas especificidades de diferentes naturezas no mundo.

O artigo de Poirier (2019) apresenta os resultados de um projeto de pesquisa etnográfica de três anos, que examinou o discurso e a prática de trabalho de projetistas de infraestrutura semiótica de três comunidades: *Semantic Web*, ligada ao W3C; *Research Data Alliance* (RDA); e uma comunidade de profissionais que projetam linguagens e modelos para descrever e ordenar dados sobre acesso ao serviço social e os critérios pelos quais os indivíduos se tornam elegíveis a eles. Nesse contexto, analisou: 18 entrevistas semiestruturadas com projetistas das 3 comunidades; dados qualitativos de anos de observação participante em eventos científicos dessas comunidades; e comunicações eletrônicas com profissionais da área. A autora denomina infraestruturas semióticas as taxonomias, esquemas e ontologias, uma vez que os compreende como um conjunto de palavras ou frases

padronizadas usadas para definir, descrever ou indexar conteúdo digital. Já o grupo denominado "**projetistas de infraestrutura semiótica**" é contido por teóricos e praticantes, dentre os quais estão especialistas em representação do conhecimento; ontologistas, cientistas de dados e engenheiros de dados. Frequentemente possuem formação em filosofia analítica, ciência da computação ou ciência da informação, com conhecimentos em linguística teórica. Trata-se de um grupo que compartilha de uma meta comum, que é a construção de modelos de dados para preservação de significado conforme se movem por diferentes domínios. A autora considera que a prática da semiótica se constitui basicamente pela **promulgação de significados e mediação de diferenças**, à medida que formalizam **padrões de dados**. Dessa forma, os projetistas de infraestrutura semiótica não têm escolha a não ser considerar compensações quando se trata de decidir sobre a complexidade/restritividade, flexibilidade/dureza, desleixo/limpeza de seus padrões e estruturas com vistas à tomada de decisões éticas e justas. O estudo, de boa qualidade metodológica, traz contribuições fundamentais para a discussão sobre o papel dos **SOC em meio à diversidade e às diferenças socioculturais**. Além disso, aborda o aspecto da informação taxonômica como informação de contexto, essencial na representação de dados e fonte-chave para a aprendizagem de máquina, conforme será abordado adiante.

Informação de contexto remete aos metadados, abordados no estudo de Kasenchak (2019), que avaliou técnicas atualmente descritas como pesquisa semântica situando-as no contexto do estado da arte da indústria editorial acadêmica. O autor afirma que as abordagens tradicionais de pesquisa falham principalmente por manipular termos ao invés de conceitos, os quais carregam significado fundamental para a correta compreensão do que está sendo pesquisado. Algumas das questões apontadas por editores de grandes *corporas* e conteúdo especializado em relação a isso dizem respeito à existência de ambiguidade na linguagem e à dificuldade de manter vocabulários controlados devido às rápidas mudanças ocorridas ao longo do tempo.

O conceito de **pesquisa semântica** abrange uma variedade de estratégias e técnicas desenvolvidas para aperfeiçoar funcionalidades de buscas tradicionais, tais como: **variantes lexicais**, que extrai diferentes formas de uma palavra, como formas singulares e plurais ou diferentes tempos verbais; correspondência difusa ou ***fuzzy***

matching, que encontra correspondências próximas em vez de correspondências exatas para melhorar a relevância da pesquisa; análise de consulta ou ***query parsing***, que tenta identificar o significado contextual de uma pesquisa em linguagem natural; **pesquisa contextual**, que abrange uma série de metodologias que utilizam informações de usuários para melhorar os resultados de pesquisa; e o **aproveitamento de metadados semânticos por meio de taxonomias**, que fornecem informações adicionais para aprimorar a experiência de pesquisa e melhorar a recuperação de informações relevantes.

O estudo descreve o uso de plataformas de publicação científica no tratamento das questões de busca sinalizadas; a análise de consultas para identificar o significado contextual das pesquisas em linguagem natural; o uso de metadados semânticos por meio de taxonomias e marcação; além da contabilização de variantes lexicais e correspondência difusa para melhorar a relevância da pesquisa. Dessa forma, revela a importância das taxonomias para o aprimoramento da experiência de busca e recuperação da informação.

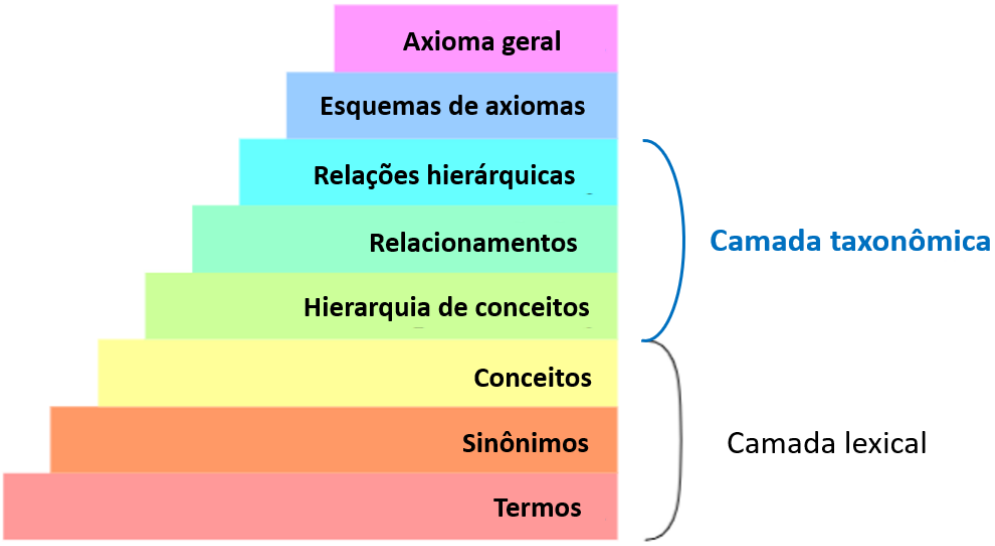
Musgrove e Walsh (2019), cujo estudo não apresenta um desenho claro, abordaram o **relacionamento entre ontologias, taxonomias e folksonomias**. O problema de pesquisa se situa na Callisto Media, uma editora que mapeia suas ontologias para várias folksonomias na *web*, incluindo as taxonomias da Amazon, conforme apontado no estudo. Com isso, compilam esboços para livros que atendem às necessidades informativas do consumidor. Nas taxonomias da Amazon os proprietários de produtos podem declarar relacionamentos e adicionar metadados, o que resulta numa enorme folksonomia com relacionamentos muitas vezes enviesados e duvidosos, prejudicando a recuperação de informações de forma assertiva. O estudo busca identificar e eliminar tais conexões duvidosas por meio de duas abordagens: uma extensional, baseada na frequência de coocorrências de categorias semelhantes de livros competitivos, e uma intencional, baseada na semelhança semântica entre o título/subtítulo de um livro e sua colocação verificadas por meio de regras e definições formais ontológicas.

Lourdusamy e John (2019) não apresentam claramente o objetivo do estudo, porém é possível verificar a proposição de um modelo de avaliação denominado *Metric based Ontology Quality Analyzer* (MOQA), que verifica a sintaxe, a estrutura e a semântica de ontologias enviadas em formato OWL. O modelo é também uma

aplicação *web* amigável que mapeia as camadas ontológicas e avalia especificamente a estrutura taxonômica para entender as relações e a organização dos conceitos. A análise é feita por **similaridade semântica** e considera o conhecimento contido na taxonomia como o mais confiável para tal análise. Assim, uma **avaliação sintática da taxonomia da ontologia** é empreendida para verificação de inconsistências e redundâncias.

Yang et al. (2019) propuseram uma abordagem para aprendizagem de termos, conceitos, relações taxonômicas e não taxonômicas no âmbito de uma ontologia. Tiveram como objetivo a **detecção automática de termos de domínios complexos** e a conceituação do domínio da engenharia de sistemas, com uso de técnicas de processamento de linguagem natural baseadas em linguística e estatística. O estudo revelou especialmente a **importância da camada taxonômica**, que será tema específico de estudos futuros dos autores, no processo de aprendizagem de ontologias. A Figura 14 foi extraída do estudo e traduzida. Demonstra as camadas da aprendizagem ontológica na visão dos autores, enfatizando a camada lexical, que consiste na extração de termos, aquisição de sinônimos e formação de conceitos; e a camada taxonômica, que se concentra na criação da hierarquia de conceitos (extração de relações taxonômicas) e da hierarquia de relações (extração de relações não taxonômicas).

Figura 14. Camadas de aprendizagem ontológica, traduzida de Yang et al., 2019



Fonte: traduzido de Yang et al., 2019, n.p.

No último estudo do ano de 2019, Raza et al. (2019) trazem o emprego de uma **taxonomia para organizar abordagens semânticas de expansão de consulta**, *Query Expansion* (QE). Utilizam relações taxonômicas para permitir a recuperação de conceitos mais gerais e específicos relacionados aos termos da pesquisa. A ideia é que ao percorrer a taxonomia para cima ou para baixo o processo de QE possa expandir a consulta a partir dos conceitos estreitamente relacionados. No entanto, um desafio se coloca no sentido de determinar a distância hierárquica apropriada para obter conceitos de expansão candidatos. Em outras palavras, é crucial decidir quantos níveis de distância do conceito original devem ser considerados ao selecionar conceitos de expansão da estrutura de conhecimento; a recomendação é de uma distância conceitual de menos de cinco níveis. São apontadas diferentes tipos de abordagens semânticas de QE, tais como: linguística, que usa um tesauro como base para extração do sentido das palavras utilizadas numa consulta; baseada em ontologia, que explora relações de generalização, especialização e outras para extrair o significado das palavras utilizadas numa consulta; híbrida, que obtém a expansão de termos a partir de múltiplas estruturas de conhecimento; ou tradicionais, que utilizam um *corpus* de conhecimento como base para a expansão de consultas. Ao final, os autores apresentam como principal resultado de experimentos de consulta a técnica QE como a mais efetiva para lidar com a escassez de termos, problema que prevalece em consultas de pesquisa na *web*.

Abre o ano de 2020 o estudo de Alonso et al.(2020), ainda com enfoque nos mecanismos de busca. O artigo propõe uma solução que **combina registros de consultas de pesquisa com dados de redes sociais** baseadas em localização - *Location-Based Social Networks* (LBSN) - para atender necessidades de usuários e plataformas. A solução apresentada, orientada a dados, mostrou resultados relevantes para consultas recreativas usando um conjunto de dados do mundo real. Foi introduzida uma taxonomia que usa informações sociais, geográficas e temporais como uma primeira etapa para modelar a relevância de pontos de interesse - *Points Of Interest* (POI) - com base em dados de *check-in* do LBSN, em *logs* de consulta e outras anotações, como dados do Twitter. O estudo descreve como classificar POI e como classificar diferentes aspectos dentro de um POI. As etapas da pesquisa contaram com: (a) uma avaliação *off-line* usando um sistema interno de *crowdsourcing,* que coletou rótulos para 10 mil POI de 2.300 cidades e onde cada POI

foi classificado como bom, regular ou ruim, em termos de sua atratividade para um visitante da cidade; (b) uma relatoria da avaliação de qualidade e taxa de defeitos da abordagem proposta; (c) um protótipo que mostrou POI e informações semânticas associadas, como dicas e outras mídias sociais; e (d) integração em um mecanismo de busca comercial para pesquisa na *web*. Uma importante limitação declarada no estudo diz respeito à confidencialidade dos dados, que não permitiu experimentos mais profundos e maior transparência.

O artigo de Salatino et al. (2020) é uma versão estendida de um estudo apresentado em 2018 na 17ª Conferência Internacional de Web Semântica, destacando a revisada *Computer Science Ontology* (CSO), ou ontologia da Ciência da Computação, e a geração de relacionamentos com várias bases de conhecimento. A CSO é desenvolvida desde 2012, foi lançada oficialmente em 2019 e em 2020 foi disseminada por meio do artigo. Trata-se de uma ontologia de grande escala, gerada automaticamente, que fornece uma caracterização mais abrangente e granular dos tópicos de pesquisa em Ciência da Computação do que o que estava disponível em taxonomias do campo e de áreas interdisciplinares até então. O estudo apresenta também o CSO *Classifier,* uma ferramenta para classificar artigos automaticamente, e o Portal CSO, um aplicativo Web que permite aos usuários baixar, explorar e fornecer *feedback* sobre CSO. A CSO é automaticamente gerada pelo algoritmo Klink-2, que analisa metadados acadêmicos para inferir relações semânticas entre palavras-chave e criar uma taxonomia de áreas de pesquisa. Utiliza métricas como intensidade de relacionamento hierárquico, informações temporais e similaridade de tópicos para determinar os relacionamentos entre as áreas de pesquisa. Por fim, uma avaliação é empreendida para verificar a eficácia e a utilidade do CSO em diferentes tarefas como classificar publicações de pesquisa, detectar comunidades de pesquisa e prever tendências. Os resultados apontaram que o novo CSO superou as abordagens existentes de evolução de ontologia, demonstrou maior precisão e recordação na previsão de novas áreas de pesquisa em comparação com métodos baseados em palavras-chave e alcançou a maior média de concordância em comparação com classificações feitas por anotadores especialistas de domínio. O estudo é apontado pelos autores como o primeiro passo na avaliação de aplicações do CSO, indicando como objetivo final a criação de um conjunto abrangente de **taxonomias de tópicos em grande escala** que descrevam os diferentes ramos da ciência.

No artigo de Shang et al. (2020) aparecem o conceito de **vetores (*embeddings*)** e o termo **construção automática de taxonomias**. O estudo apresentou a NetTaxo, uma plataforma de construção automática de **taxonomia de tópicos,** que treina vetores do texto e da rede (de conhecimento) capturando-os como informação de contexto e permitindo que haja **troca de informações entre os dados e a estrutura**. É feita uma seleção de *motifs*, que refinam os vetores de termos de acordo com a granularidade e a semântica de cada nó na taxonomia e os agrupa em subtópicos. Experimentos em dois conjuntos de dados avaliaram a plataforma em relação a outras metodologias e a importância da seleção de *motifs* a nível de instâncias. Os autores valorizam as taxonomias na organização do conhecimento de um domínio ao afirmar que "uma das maiores vantagens da construção automática de taxonomias é a habilidade de capturar informações de *corpus* específicos e adotar para diferentes cenários" (Shang et al., 2020, n.p., tradução nossa). Ressaltam também sua aplicação acadêmica ao apontar que a função de organizar artigos científicos dentro de uma "estrutura de taxonomia dá aos pesquisadores uma **visão de águia sobre o campo**, e então eles podem rapidamente identificar seus interesses, e facilmente adquirir a informação que desejam" (Shang et al., 2020, n.p., tradução nossa).

Em Soares et al. (2020) é realizada uma demonstração que tem como objetivo apresentar a arquitetura de múltiplas visualizações do sistema de exploração de conhecimento *Knowledge Explorer System* (KES), que permite a modelagem contextualizada de ontologias sobre uma representação híbrida de conhecimento, facilitando a criação e a **visualização de taxonomias e relacionamentos definidos em um contexto específico**. Apresentam também o *Hyper Knowledge* (HK), que suporta a integração de conteúdo simbólico e não simbólico. O conteúdo simbólico se refere a conceitos de alto nível e relações semânticas dentro de ontologias, representando o conhecimento do domínio de forma estruturada. Já o conteúdo não simbólico é constituído de conteúdo multimodal, como modelos de aprendizagem de máquina, imagens, vídeos, texto e áudio. Trata-se de uma representação híbrida de conhecimento que preenche a lacuna semântica entre o que significa um conteúdo não simbólico e sua representação conceitual simbólica, fornecendo construções de primeira classe que podem ser usadas para descrever relacionamentos entre esses dois tipos de entidades. O estudo aborda o mecanismo de visualização de taxonomias e sua importância na organização do conhecimento, especialmente o contextual.

Com Weser et al. (2020) a abordagem das taxonomias de domínio se manteve ao utilizarem uma taxonomia amplamente conhecida no contexto de aplicação do estudo. Diante de mudanças no setor fabril provocadas pela passagem da produção em massa para os mercados de nicho, que exigem alta flexibilidade e disponibilidade de recursos sob demanda, foi proposta por uma arquitetura de ontologia hierárquica denominada *Capability for industry* (C41), feita para representar, publicar e ampliar o conhecimento sobre *capabilities* ou efeitos que os recursos de manufatura podem causar em diferentes domínios de aplicação e contextos. Além disso, o artigo descreve padrões de como a lógica formal subjacente pode ser usada na **modelagem taxonômica** e na **inferência** de fatos de capacidade implícita. O uso de uma conhecida taxonomia de domínio garantiu que houvesse um entendimento comum sobre a classificação e a categorização das capacidades fabris, culminando em inferências para descoberta de novas informações.

Encerra o ano de 2020 nesta temática o artigo de Teixeira e Sousa (2020), um dos três estudos brasileiros da RI e o único do campo da CI. Apresenta uma **taxonomia facetada do domínio** da lesbianidade. A taxonomia é proposta como forma de explorar os termos metafóricos empregados na representação temática de uma amostra de fotografias presentes no catálogo de assuntos do *Lesbian Herstory Archives* (LHA). É empregada como sistema de representação e contextualização da comunidade LGBTQIA+ enquanto comunidade discursiva, detentora dos significados reais dos conceitos aplicados nesse tema. A coleção do LHA conta com 12 mil fotografias, das quais 664 disponíveis foram analisadas no estudo. Os termos do *corpus* da pesquisa foram provenientes do metadado *Subject-Topics* do LHA, campo que geralmente recebe o termo ou frase que melhor representa o assunto/tema de um documento na base. A *Encyclopedia of sex and gender* auxiliou a conceituação de 100 figuras de linguagem selecionadas, que foram então submetidas a um processo de categorização voltado à construção de facetas. O processo de análise dos **metadados para a construção da taxonomia** revelou inconsistências terminológicas ou incongruências na estrutura de **representação temática**, potenciais ruídos para a recuperação da informação. Uma taxonomia facetada estruturada em 4 termos metafóricos foi desenvolvida e apresentada.

Wen et al.(2021) apresentaram o desenho e a implementação de uma ontologia, cujo objetivo foi identificar aspectos da modelagem de cenários de testes que fossem

relevantes para a prática de exercícios de segurança cibernética, ou *Cyber Security Exercises* (CSE), e para a construção de uma **taxonomia do conhecimento como vocabulário controlado** para esses cenários. Os CSE dividem-se em exercícios baseados em discussões e exercícios baseados em operações e constituem-se como uma oportunidade de aplicação de conceitos hipotéticos em um ambiente de treinamento controlado, sem efeitos no mundo real. A modelagem da ontologia contou com uma revisão de literatura nos domínios dos exercícios de segurança cibernética e na modelagem do conhecimento nessa área. Então, três submodelos foram constituídos, sendo eles: o modelo do cenário informacional, o modelo do cenário operacional e o modelo de segurança do conhecimento. A implementação da ontologia contou com o *software* Protégé e padrões e ontologias de segurança cibernética existentes foram pesquisados para incorporar as **taxonomias mais comumente usadas** na ontologia e para fornecer um entendimento comum do domínio da segurança cibernética. A pesquisa resultou no uso das seguintes taxonomias como base para a padronização: para padrões de ataques foi adotada a *Common Attack Pattern Enumeration and Classification* (CAPEC), a taxonomia considerada de maior desempenho em relação às demais do domínio, que organiza padrões de ataque hierarquicamente com base em mecanismos que são frequentemente empregados na exploração de uma vulnerabilidade; para ameaças, a *European Union Agency for Cybersecurity* (ENISA), que resume as ameaças cibernéticas que foram acessadas através da coleta de informações disponíveis publicamente; para fraquezas foram adotadas a *Software Engineering Institution* (SEI) no âmbito dos fatores humanos e organizacionais, e sob a perspectiva de sistemas e tecnologias, a *Common Weakness Enumeration* (CWE). A taxonomia final foi estruturada em cinco funções de alto nível: identificar, proteger, detectar, responder e recuperar.

Zhang et al. (2021c), no estudo denominado *CrawlPhish: Large-scale analysis of client-side cloaking techniques in phishing*, analisou técnicas de camuflagem do lado do cliente (usuário) em *sites* de *phishing* para entender seus métodos de evasão por meio de uma taxonomia construída manualmente e empregada na organização e compreensão do conhecimento do domínio. O sistema de análise CrawlPhish apresentado no estudo, detecta e categoriza as técnicas de camuflagem em *sites* e robôs no ambiente *web* para melhorar as estratégias de defesa e a monitorização de

potenciais ataques. O CrawlPhish coleta o código-fonte de *sites* de *phishing* em tempo real, detecta e categoriza automaticamente as técnicas de ameaça adotadas. O sistema foi executado por mais de 14 meses entre 2018 e 2019 e analisou 112.005 sites de *phishing*, identificando 1.128 implementações de técnicas de camuflagem distintas. Com base nos resultados, foi empreendida análise semântica das técnicas para a proposição de uma taxonomia de oito tipos de técnicas de evasão usadas nos ataques em três categorias de alto nível, sendo elas: interação do usuário, impressão digital e comportamento do robô. Por meio de 150 *sites* de *phishing* artificiais o estudo demonstrou a eficácia de cada categoria de técnica de evasão ao evitar a detecção de *phishing* baseada em navegadores de internet. Objetivamente, o artigo demonstrou o uso de **taxonomia na categorização** das técnicas de ameaça com alta precisão, o que apresenta importante potencial de aplicação para a segurança em diferentes domínios.

O artigo de Smith et al. (2022) discutiu a necessidade de criação de uma ontologia para classificar as profissões, **atualizando populares taxonomias** de ocupação disponíveis em língua inglesa. Para tanto, utiliza como dados de entrada quatro taxonomias que refletem décadas de desenvolvimento e adoção, mas apresentam diferentes padrões de classificação e, na visão dos autores, se baseiam em um **sistema restritivo de codificação digital**, são elas: a *International Standard Classification of Occupations* (ISCO), do português Classificação Internacional Padrão de Ocupações, que serviu de base para a construções de outras taxonomias populares em inglês, sendo portanto a referência internacional oficial; a *US Bureau of Labor Statistics Standard Occupational Classification* (US SOC), do português Classificação Ocupacional Padrão da Secretaria de Estatísticas Trabalhistas dos EUA, disponível no sistema O*Net; a *The UK National Statistics Standard Occupational Classification* (UK SOC), do português Classificação Ocupacional Padrão Nacional de Estatísticas do Reino Unido; e a *The European Skills, Competences, Qualifications and Occupations* (ESCO), do português Habilidades, Competências, Qualificações e Profissões Europeias. O estudo destaca diferenças de **granularidade** e **categorizaçã**o entre as quatro taxonomias, a falta de **consenso** global sobre determinadas métricas e sugere a padronização da forma como atributos ocupacionais, como habilidades, são organizados ontologicamente em diferentes taxonomias. A abordagem proposta pelos autores emprega métodos bem-sucedidos

da *Open Biological and Biomedical* (OBO), que é utilizada como base para a transformação da US SOC em um protótipo de ontologia, nomeada então como *Occupational Classification Compliant Ontology* (OccO). A OccO é comparada à Wikidata, base de representação de ocupações da *web* semântica, onde as ocupações não são organizadas segundo uma taxonomia padrão ou uma estrutura ontológica consistente de nível superior. Apresenta-se então os benefícios do desenvolvimento de uma estrutura ocupacional orientada pela ontologia. Além dos motivos já explicitados, o desenvolvimento da OccO busca o apoio de organizações que possam cooperar para formular e demonstrar a utilidade de uma ontologia bem concebida e amplamente utilizada para ocupações. A OccO é desenvolvida seguindo o *Basic Formal Ontology* (BFO). Dentre os principais resultados afirma-se que: é recomendado o uso dos princípios da OBO *Foundry* no desenvolvimento de ontologias em geral; a implementação da ISCO na ESCO da UE tem melhorias significativas e é sugerida como uma melhor base para começar um trabalho semelhante, seguida pela incorporação de elementos adicionais de outras taxonomias padrão; competências, habilidades e áreas de conhecimento estão representadas entre as taxonomias padrão, mas a harmonização entre estas referências é necessária e seria uma realização benéfica.

Zhang e Sun (2022) trazem como objetivo estimar a similaridade existente nas **relações hiperônimo/hipônimo**. Para tanto propõem um método híbrido que utiliza a WordNet como base e envolve duas abordagens comumente utilizadas para esta finalidade: *IC-based*, que considera o conteúdo da informação nos pares; e *feature-based*, que estima a similaridade por meio da hierarquia da taxonomia. Foram utilizados dados artificiais no conjunto de dados M&C, que incluiu 30 pares de palavras determinados por profissionais. Cada palavra correspondia a uma série de conceitos na WordNet, entendidos no âmbito do estudo como palavras de busca. Para validar a proposta os autores compararam o desempenho do método, o teste artificial e mais quatro métodos típicos na WordNet. Os resultados experimentais mostraram que o método proposto obteve melhor desempenho no conjunto de dados padrão M&C. A qualidade metodológica do estudo foi considerada muito boa e avaliou-se que os autores trouxeram a taxonomia como elemento fundamental em métodos de estimativa de **similaridade semântica**, que nesse estudo foi denominada similaridade computacional.

Este ano aparecem Sharifpour et al. (2023) com um estudo em grande escala baseado em *logs* de comportamento de usuários em pesquisas realizadas em 2019 na *Research Data Australia* (RDA), o portal nacional australiano de descoberta de dados de pesquisa, que faz parte do *Australian Research Data Commons* (ARDC). Uma análise de *cluster* (agrupamento) foi aplicada para diferenciar perfis de usuários com base nos comportamentos de pesquisa. A análise considerou os seguintes critérios: (a) o número de páginas visitadas em uma sessão; (b) o número de consultas emitidas em uma sessão; (c) o tempo total de duração da sessão; (d) o número de *tokens* exclusivos em uma sessão para identificar usuários únicos removendo duplicidades; (e) o número de palavras exclusivas que um usuário utilizou para especificar sua consulta de pesquisa em uma sessão; e (f) o número médio de *tokens* usados para formular consultas em uma sessão. A semelhança semântica entre as consultas de pesquisa e os títulos dos metadados foi calculada para avaliar a experiência dos usuários e identificou seis perfis ou classes: *Expert Research, Expert Search, Expert Explore, Novice Research, Novice Search* e *Novice Explore*. Nesse contexto, os autores propuseram uma taxonomia de perfis de usuários de pesquisa de nível de especialização que compartilha semelhanças com taxonomias previamente identificadas na área de pesquisa na *web* e no comércio eletrônico. A taxonomia proposta contempla tanto a experiência dos usuários no domínio, quanto o objetivo ou a motivação da busca no portal de dados de pesquisa, contribuindo para identificar uma classificação mais significativa dos utilizadores, tendo em conta não só o seu **comportamento de pesquisa**, mas também as motivações subjacentes às suas características de pesquisa. É feita distinção entre *expertise* de domínio e *expertise* de pesquisa, o que contribui para a compreensão do estudo. Segundo os autores, o conhecimento do domínio está relacionado ao conhecimento do pesquisador sobre o tópico de pesquisa, enquanto a experiência em pesquisa está relacionada ao conhecimento do processo de pesquisa e à capacidade de construir uma consulta que resulte em alta precisão dos resultados da pesquisa. Segundo os autores, trata-se da primeira **taxonomia baseada em *logs* de usuários em um *dataset* de dados de pesquisa**.

Já Jain (2023), com vistas a superar desafios do mercado de comércio eletrônico, propõe o desenvolvimento de um portal de pesquisa e comparação de preços para ofertas de produtos fornecidos por *eShops*, OEMs (fabricantes originais

dos equipamentos) e varejistas locais. A arquitetura do portal segue elementos da *web* semântica - como metadados, taxonomias, ontologias e **grafos do conhecimento** - e oferece duas interfaces: uma para novos produtos que serão incorporados à base de conhecimento e outra para os usuários finais obterem sugestões sobre quais produtos comprar. O usuário final pode inserir especificações técnicas brutas do produto a ser adquirido e obter as principais recomendações (em ordem decrescente de similaridade) sobre os produtos disponíveis por meio de recursos de venda. Cada produto recomendado é acompanhado de especificações detalhadas. A abordagem proposta consiste em aproveitar os dados do portal para fornecer inteligência à experiência de uso, agregando valor por meio do que a *web* semântica pode oferecer. Assim, ontologias são formalizadas em OWL e representadas por grafos do conhecimento; uma representação do catálogo de produtos é transformada na ontologia de equipamentos de engenharia (EEO), de modo que seja interoperável e possa ser combinada com ofertas de outros varejistas; e um vocabulário EXIF é utilizado para mapear os metadados da imagem do produto com os metadados do produto, permitindo aos usuários pesquisar por meio de imagens. O portal se utiliza de **inteligência artificial e mecanismos de inferência para recomendações direcionadas** e de alta qualidade. Os **esquemas de metadados** GoodRelations, schema.org e Dublin Core descrevem a taxonomia central relacionada aos produtos, porém apesar de conterem detalhes de fabricação dos produtos, não contam com outras especificações relacionadas à busca, como cor, tamanho, peso, entre outras, o que demonstra a falta de uma taxonomia com essas características para agregar valor ao sistema. Princípios de ***Linked Data*** **(LD)** também foram adotados pelo portal, que faz uso de **identificadores persistentes** para conectar diferentes **ecossistemas** e **bases de conhecimento** no suporte aos **mecanismos de busca e recomendação**. Resultados de diversos experimentos comparativos com portais de *e-commerce* semelhantes demonstraram principalmente que a aplicação de tecnologias da *web* semântica e de inteligência artificial aprimoram a **integração** e a **interoperabilidade** de dados para a descoberta de produtos.

4.2.2 Bases de conhecimento (knowledge bases) e vetores de palavras (embeddings)

Em Maldonado et al. (2019) surge pela primeira vez o termo *embeddings* no conjunto de estudos recuperados. Abordaram o uso de **bases de conhecimento** e de

enriquecimento taxonômico para melhorar a compreensão das informações existentes nessas bases, geralmente volumosas. O estudo descreve experimentos que combinam *corpus* natural com vetores taxonômicos treinados em dados de diferentes tamanhos. A base de conhecimento para **vetores de *corpus* natural** é a Wikipedia e para **vetores taxonômicos** são realizadas caminhadas randomizadas pela estrutura hierárquica do WordNet (um *synset*), que geram dados de treinamento (*pseudo-corpus*) posteriormente analisados no estudo. Os autores analisaram em profundidade o impacto que **vetores taxonomicamente enriquecidos** causam em *corpus* naturais, com vistas à descoberta da quantidade ideal de dados taxonômicos e de *corpus* natural para ganhos de desempenho no treinamento de vetores de palavras para **aprendizagem de máquina** em bases de conhecimento. Para tanto, dois modelos de combinação de vetores foram propostos: concatenação, que cria um vetor com as dimensões obtidas pela soma da caminhada radomizada (*random walk*) do vetor taxonômico por cada palavra do *corpus* natural; e ajuste fino (*fine-tunning*), quando são embutidas informações taxonômicas no *corpus* natural para formação de um *pseudo-corpus*. Os modelos propostos pelo estudo foram avaliados em três diferentes *datasets* de **similaridade lexical**: SemEval-17, WordSim-353 e SimLex-999.

Wang (2019b), no artigo *Set-based Noise Elimination for Is-a Relations in a Large-Scale Lexical Taxonomy*, propõem, descrevem e avaliam uma abordagem baseada em conjuntos para eliminação de ruídos (erros na marcação de relacionamentos, erros de grafia ou ausência de contexto que impossibilite a compreensão de determinado termo ou relacionamento) em bases de conhecimento por meio de um **algoritmo**, que decide quais relacionamentos devem ser eliminados para melhorar a acurácia da base. Conclui que a baixa similaridade existente entre entidades próximas pode ser um importante indicador para localizar **relações tipo-instância** incorretas em **taxonomias lexicais de grande escala**. O estudo emprega como base a taxonomia probabilística Probase da Microsoft, considerada um **grafo do conhecimento**, onde cada nó é uma entidade, que por sua vez subdivide-se em conceito geral ou hiperônimos (com arestas de entrada) e instância específica ou hipônimos (subordinados a conceitos amplos).

Malandri et al.(2020) inauguram 2020 nesta temática com o experimento de implementação do *Method for Embeddings Evaluation for Taxonomic Data* (MEET),

uma metodologia para gerar e avaliar vetores de palavras a partir de um *corpus* de texto, preservando as relações de similaridade semântica derivadas de uma taxonomia. É introduzida uma nova medida chamada Similaridade Semântica Hierárquica ou *Hierarchical Semantic Similarity* (HSS), para calcular a semelhança semântica entre elementos taxonômicos. São tomadas como base de conhecimento para avaliação das propostas a WordNet, que provê uma hierarquia de significados e um conjunto de sinônimos, e a Wikipedia no idioma inglês. O modelo vetorial é testado na biblioteca FastText usando Skipgram e CBOW, modelos de treinamento de redes neurais. Conforme demonstra o estudo, as taxonomias são essenciais para a compreensão da máquina no **Processamento de Linguagem Natural (PLN)**. No entanto, manter essas hierarquias atualizadas e capazes de representar o domínio do qual foram extraídas ainda é uma atividade demorada e propensa a erros, ponto onde a aprendizagem de máquina atua.

Em 2021 os autores publicam o artigo *MEET-LM: A method for embeddings evaluation for taxonomic data in the labour market*, que aborda o enriquecimento de taxonomias (Malandri et al., 2021). O MEET-LM é uma nova metodologia de geração e avaliação de vetores de palavras a partir de um *corpus* de texto ou um método não supervisionado para gerar e selecionar vetores considerando relações de similaridade codificadas em uma taxonomia de domínio. O MEET-LM é aplicado a um caso real de um projeto da União Européia (UE) sobre inteligência do mercado de trabalho, cuja taxonomia base denominada ESCO apresenta mais de dois milhões de vagas *online* e é utilizada como um padrão de classificação. No estudo são consideradas as vagas pertencentes ao setor de tecnologia. Os resultados experimentais mostraram que o MEET-LM é capaz de captar, incorporar e representar de forma eficaz as semelhanças entre esses empregos dentro da taxonomia padrão da ESCO à medida que emergem da procura real do mercado de trabalho. O MEET-LM alcança uma precisão de 99,4%.

4.2.3 Grafos do conhecimento (knowledge graphs)

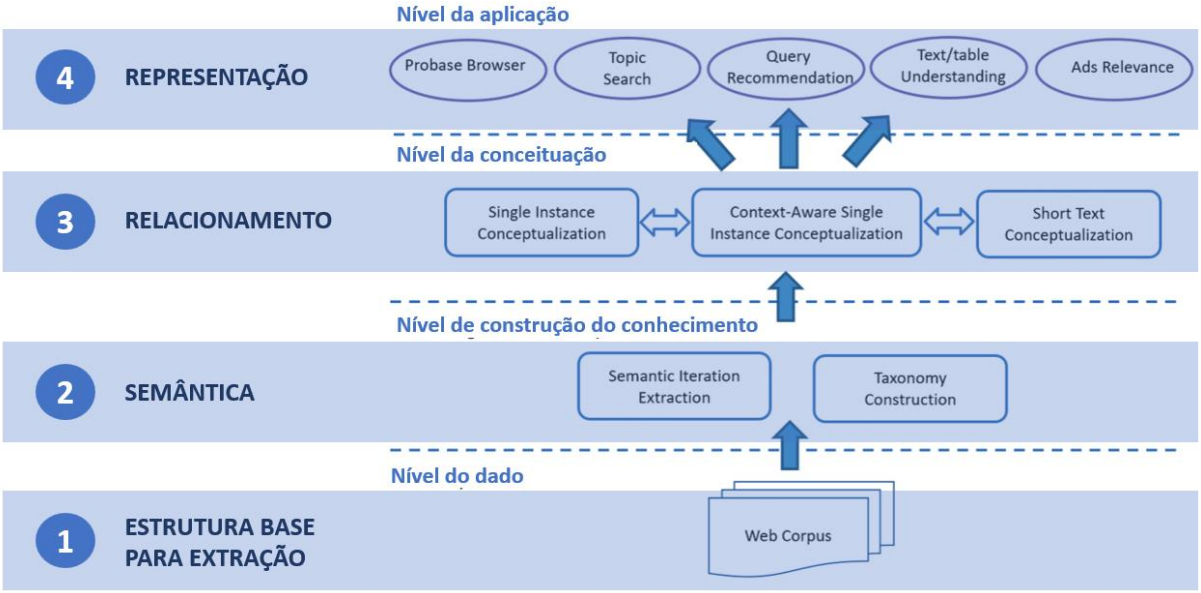
Fauceglia et al. (2019) demonstraram como o serviço de indução de grafos do conhecimento ou *Knowledge Graph Induction Service* (KGIS) realiza a **indução automática de taxonomias** a partir de conjuntos de documentos por meio da análise de **padrões linguísticos**, de **similaridade semântica** e de **redes neurais**. Descrevem também funcionalidades para **curadoria e expansão de taxonomias** utilizando um componente KGIS denominado *Smart Spreadsheet* (SSS), uma planilha

editável com preenchimento inteligente para autopreenchimento de termos relacionados e maior controle da taxonomia. No estudo, as taxonomias são a estrutura e o produto principal gerado pelo KGIS e armazenado como SSS para manipulação. A consistência dos relacionamentos gerados promove melhor compreensão dos dados para a geração de informação a partir deles e permite que a semântica seja expandida e aprofundada por outros conjuntos de dados advindos de bases de conhecimento - também estruturadas como taxonomias - ou por usuários.

Ji et al. (2019) é um dos estudos mais relevantes desta temática, uma vez que apresenta a arquitetura da nova aplicação de marcação de conceitos Microsoft Concept Graph, desenvolvida para facilitar o PLN, construída com base na taxonomia probabilística universal Probase e composta por conceitos e instâncias extraídos da *web*. O estudo compara a Probase com outras taxonomias indicando um volume expressivamente superior às concorrentes e a inclusão de um tratamento probabilístico para aprofundamento semântico. Os autores tinham como objetivo construir automaticamente a rede semântica de fatos *isA* (relações tipo-instância verificadas), já que acreditavam que seus concorrentes apresentavam baixa cobertura de conceitos e não possuíam taxonomias bem organizadas. Múltiplos cenários de avaliação das ferramentas foram apresentados, bem como os resultados, que demonstraram uma precisão de 92,8% na média de todos os conceitos existentes (cerca de 5,4 milhões). O processo de conceituação do Microsoft Concept Graph foi considerado efetivo e robusto. Foi extraída e parcialmente traduzida do estudo a Figura 15, que buscou explicitar as funções das taxonomias na arquitetura do Microsoft Concept Graph. Nos níveis mais básicos as taxonomias atuam no suporte aos algoritmos de extração de dados para criação de *corpus* de conhecimento, fornecendo a estrutura hierárquica necessária à distinção entre conceitos e suas instâncias; no nível 2 ocorrem as tarefas de efetiva mineração dos relacionamentos tipo-instância (extração de conceitos) e construção de taxonomias que serão a base para a geração de grafos do conhecimento (KG) (redes semânticas); já no nível 3 são analisados os KG obtidos, o nível de confiança dos relacionamentos identificados e inferidos os conceitos finais com auxílio de bases de conhecimento *online*; o nível 4 é a camada em que os KG fornecem API de marcação de conceitos para facilitar a compreensão das linguagens humanas. As taxonomias geradas no nível 2 são mescladas vertical e horizontalmente, momento em que são aplicadas análises

probabilísticas para construção de uma taxonomia global.

Figura 15. Funções das taxonomias na arquitetura do Microsoft Concept Graph, adaptado de Ji et al., 2019



Fonte: parcialmente traduzido de Yang et al., 2019, n.p.

Zouaq e Martel (2020) aparecem em 2020 com um estudo experimental que investiga uma série de modelos de incorporação de KG e avalia sua adequação para a identificação de uma **taxonomia atual de grafos de conhecimento**. Um método de agrupamento automático de entidades de KG é proposto e conta com a distribuição estatística de classes e propriedades para rotular cada **agrupamento** e gerar uma **taxonomia expressiva** que represente o conteúdo atual de um KG.

No ano seguinte, Zhang et al.(2021b) publicam o *AliCG: Fine-grained and Evolvable Conceptual Graph Construction for Semantic Search at Alibaba.* Estudo experimental que introduz o *Alibaba Conceptual Graph* (AliCG), um grafo com mais de cinco milhões de conceitos refinados extraídos de *logs* de pesquisa ruidosos. O AliCG possui quatro níveis: (1) conceitos de domínio e suas instâncias; (2) conceitos que se referem a um tipo ou subclasse de instância; (3) conceitos que são a conceituação refinada de instâncias que expressam as intenções implícitas do usuário; (4) e a camada que inclui todas as instâncias, inclusive entidades e frases que não são de entidades. Dentre as etapas de construção do grafo estão o refinamento de conceitos, a **mineração de conceitos de cauda longa** e a construção de uma taxonomia a partir da distribuição de conceitos baseados no comportamento

dos usuários e minerados nas etapas anteriores. O estudo também compara o método proposto com outros semelhantes. Destaca-se o papel das **taxonomias como etapa de construção de um grafo do conhecimento**.

4.2.4 Linked (Open) Data

Nesta categoria aparecem estudos somente a partir de 2020, com Crotti Junior et al. (2020), que apresentam resultados preliminares e propõem uma abordagem para gerenciamento de riscos por meio de uma plataforma denominada *Acess Risk Knowledge* (ARK), que utiliza o arcabouço de tecnologias da *web* semântica e se baseia na metodologia *Cube* de análise de riscos e coleção de evidências científicas. O objetivo do trabalho é demonstrar o potencial de uso do *Linked Data* (LD) para integrar dados operacionais e de segurança de sistemas dentro de um **espaço de informação unificada** que apoie o gerenciamento de riscos clínicos de forma mais **contínua**, **adaptativa** e **preditiva**. Para tanto, são empregadas ontologia, taxonomia representada em SKOS, RDF e LD para modelar, integrar e classificar riscos e informações de análise de sistemas sociotécnicos a partir de fontes de dados quantitativas e qualitativas posteriormente unificadas em um gráfico de risco. A metodologia Cube é aplicada há vinte anos em sistemas de segurança para os domínios da saúde, aviação e finanças, como guia para análise estruturada de especialistas e síntese multidimensional de aspectos sociotécnicos baseados na psicologia organizacional. O estudo contou com apenas dois participantes, um especialista em gerenciamento de riscos e um não especialista, que utilizaram a plataforma ARK para analisar o impacto da liderança coletiva na integração de um sistema de prontuário eletrônico (saúde). Uma taxonomia de gerenciamento de riscos clínicos foi manualmente desenvolvida e posteriormente expressa em SKOS para **análise automatizada de dados qualitativos**. A plataforma ARK foi considerada útil e de fácil aprendizado e fornece uma visão abrangente dos casos clínicos. Funcionalidades como interligação de fontes de dados como evidências, definições da taxonomia disponíveis na interface e possibilidade de ocultar funções se destacaram entre os usuários.

Já Dilorenzo et al. (2020) desenvolveram uma pesquisa experimental qualitativa, que incluiu uma taxonomia de histórias de usuários e a avaliação de seu potencial para promover a **engenharia de software orientada a reuso**, no contexto de uma abordagem de **desenvolvimento ágil** de software denominada *Scrum*. A histórias de

usuários são anotações semiestruturadas que especificam os requisitos do projeto compartilhados entre os envolvidos. A **taxonomia facetada** proposta em dois níveis foi adaptada para sistemas de informação *web* e foi validada usando *backlogs* de produtos de projetos de *software* de quatro empresas desse ramo. O processo de validação contou com um estudo *offline* para investigação da viabilidade e possíveis problemas, onde foi aplicada a taxonomia, e com a realização de um estudo de caso em dois projetos reais. Os resultados indicaram, num primeiro momento, que a taxonomia foi eficaz no reuso de histórias de usuários e seus ativos relacionados. Então, embasado pelo Modelo de Aceitação de Tecnologia ou *Technology Acceptance Model* (TAM), foram avaliadas a facilidade de uso da taxonomia e sua utilidade percebida, apontando a pronta adoção da taxonomia pela equipe, pela percepção positiva que tiveram. O estudo contribui para a aplicação de taxonomias voltadas para o **reuso de dados no ambiente *web***, uma vez que os autores propõem o uso da taxonomia como estrutura de um sistema corporativo que possa atender diferentes projetos e enriquecer as histórias a partir de informações de fornecedores e outros sistemas.

Em 2021 Yaman et al.(2021) apresentaram uma taxonomia de licença aberta documentada para os temas de dados das Nações Unidas sobre Gestão Global de Informação Geoespacial - *The United Nations initiative on Global Geospatial Information Management* (UN-GGIM), que segue as melhores práticas de *Linked Data*, como o padrão de representação do conhecimento na *web* semântica SKOS e o padrão para catalogação de metadados DCAT. Além disso, fornece uma visão geral dos temas da UN-GGIM e um exemplo de aplicação da taxonomia para implantação em *Ordnance Survey Ireland* (OSi) para geração de metadados DCAT. A OSi é a agência nacional de mapeamento da Irlanda, que produz dados geoespaciais. O principal objetivo do estudo foi promover interoperabilidade de dados entre organizações das Nações Unidas, facilitar a integração de dados descentralizados e melhorar a recuperação por meio de temáticas específicas de dados. Para tanto, uma taxonomia foi implementada como um vocabulário controlado LD para padronizar definições que envolvessem a gestão dos conjuntos de dados. A modelagem SKOS foi escolhida e a construção em si seguiu três etapas metodológicas: codificação RDF para gerar a **taxonomia em formato LD**; validação de dados em busca de erros; e publicação dos dados na *web*. Cada tema UN-CGIM foi definido como uma classe de conceitos.

Recentemente, Sha et al. (2023) propuseram uma nova ferramenta de classificação de tópicos baseada em regras para perguntas em plataformas de perguntas e respostas (Q&A) mediadas pela Wikidata, uma base de conhecimento multilíngue aberta e acessível, que possui uma hierarquia formal, mas permite aos usuários adicionar e alterar informações. Duas populares plataformas de perguntas e respostas, a chinesa Zhihu e a americana Quora, são tomadas como fonte de dados. A Wikidata é adotada com o objetivo de alcançar interoperabilidade de dados e classificação hierárquica de tópicos entre as múltiplas plataformas. Assim, na ferramenta proposta as **relações taxonômicas da Wikidata servem como referência para a categorização do conteúdo *web* de maneira unificada, em diferentes níveis de refinamento e conforme a cobertura do domínio**. Diferentes casos de uso e experimentos são realizados a fim de validar a abordagem e demonstrar a aplicabilidade da ferramenta, que também é comparada com a saída gerada pelo classificador GPT-3 da OpenAI para maior aprofundamento da análise. De forma geral, a abordagem proposta consiste em: extração de entidade de cima para baixo dos itens que possuem URI nas três plataformas; estruturação hierárquica de acordo com suas relações taxonômicas; rastreamento complementar de tópico de baixo para cima; mapeamento com base em regras de inferência; e definição das categorias com cobertura semântica semelhante formando uma taxonomia única mais confiável e que pode ser útil à classificação de conteúdo nas plataformas Zhihu e Quora no contexto de uma ontologia.

Somente este ano surge nos estudos o conceito de LOD, acompanhado de princípios FAIR, com o trabalho de Golpayegani et al. (2023). O estudo se coloca no contexto da lei de Inteligência Artifical (IA) da União Européia, que tem como enfoque o impacto dos sistemas de IA considerados de “alto risco” para a saúde, a segurança e os direitos. Analisa as cláusulas que definem os critérios para IA de alto risco e uma série de taxonomias da área como *AI, Algorithmic and Automation Incidents And Controversies* (AIAAIC), *AI Incident Database* (AIID), estrutura de classificação da Organização para a Cooperação e o Desenvolvimento Econômico (OCDE) para sistemas de IA e *AI risk ontology* (AIRO), para mapeamento da estrutura conceitual da temática. Os critérios para determinar se uma IA é de alto risco incluem o domínio, a finalidade, a capacidade, o usuário e o assunto da IA. Um sistema de IA precisa atender a apenas uma das condições para ser considerado de alto risco, a menos que

tenha uma saída puramente acessória ou seja usado por um provedor de pequena escala para fins específicos. Porém, os conceitos identificados não são suficientes para anotar casos de utilização de IA, representar e documentar a gestão de riscos, estabelecer regras para identificação de IA de alto risco e partilhar conhecimentos e melhores práticas sobre riscos de IA, o que demanda enriquecimento do modelo com instâncias de conceitos representados formalmente e organizados em hierarquias. Assim, o estudo discute o desenvolvimento do VAIR, uma **taxonomia formal para representar hierarquias de IA** e conceitos de risco. O VAIR integra conceitos da Lei de IA da UE, da ISO/IEC 22989 e da taxonomia do *AI Watch*, com o objetivo de resolver inconsistências e conflitos para uma taxonomia de riscos coerente. Além disso, adota os **princípios FAIR** de encontrabilidade, acessibilidade, interoperabilidade e reuso. Reutiliza conceitos centrais da taxonomia AIRO e representa instâncias usando o modelo SKOS recomendado pelo W3C para a *web* semântica.

4.2.5 Inteligência Artificial

Danilevitch (2019) apresenta um estudo experimental de cunho linguístico, realizado no campo da computação e aplicado ao universo econômico-financeiro, que investiga diferenças na morfologia e na **semântica das linguagens naturais**, analíticas *versus* sintéticas, e na forma como essas diferenças podem influenciar a modelagem de dados nas taxonomias *eXtensible Business Reporting Language* (XBRL). As taxonomias XBRL são a modelagem padrão de dados que digitaliza a estrutura regulatória e a cadeia de suprimentos de relatórios financeiros, permitindo a classificação de demonstrações contábeis e documentos de relatórios comerciais. A tecnologia XBRL é padronizada fora do W3C pela XBRL International, uma organização independente, e faz uso de variadas tecnologias XML para troca automatizada de metadados, que são marcados de acordo com taxonomias que contém elementos do esquema XBRL vinculados a um objeto do mundo real representativo de um conceito contábil, econômico ou comercial definido em um documento regulatório oficial. O estudo tem como principal objetivo formalizar um **protocolo para modelagem de taxonomias** XBRL bem-sucedidas, que seja útil à transição da semântica de dados de uma linguagem humana para uma linguagem formal de máquina. A autora coloca que as **linguagens naturais sintéticas** são mais desafiadoras para o *design* de taxonomia. Para domínios de conhecimento complexos

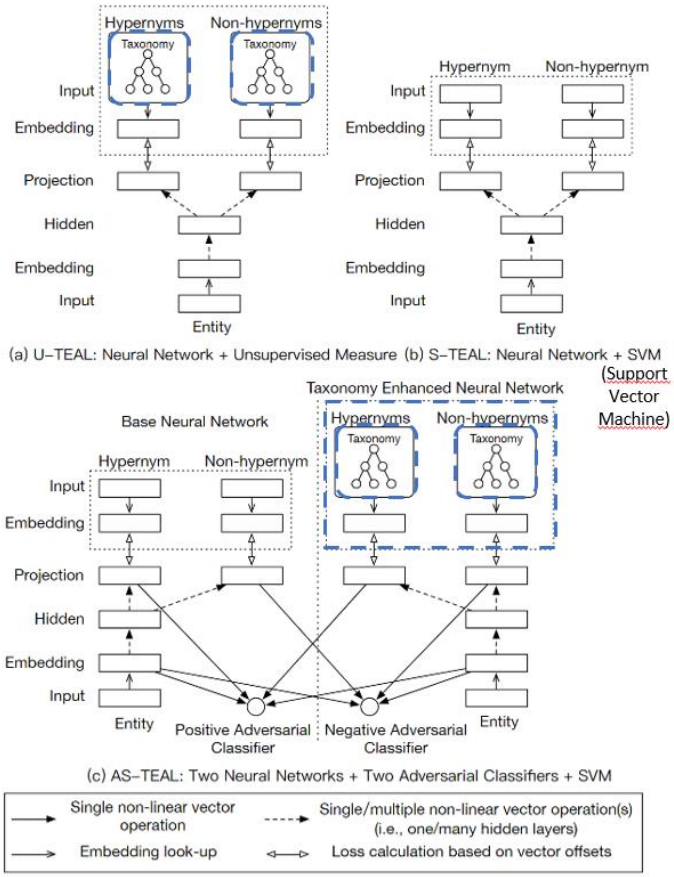
diretamente ligados a pronunciamentos, regras ou leis regulatórias, a classificação deve ser projetada por pessoas que entendam a natureza técnica dessas regras de negócios. A modelagem de dados em taxonomias XBRL representa um exemplo de **abordagem baseada em humanos para classificação de informações legíveis por humanos em um formato legível por máquina**. Nesse sentido, a pesquisa proposta contribui no campo da *web* semântica para a transformação de dados não estruturados em relatórios estruturados processáveis por IA, capazes de configurações e comparações complexas.

Figueroa et al.(2019) apresentam o resultado da **fusão de taxonomias** obtidas em trabalhos anteriores dos autores, nos anos de 2012 e 2016, gerando um *corpus* de 6.683 perguntas em inglês do Yahoo! Respostas. Trata-se de um *corpus* anotado com facetas temporais por meio de *crowdsourcing*, com vistas a facilitar que a máquina identifique a obsolescência de perguntas ao longo do tempo e a regularidade temporal do interesse dos usuários, aprimorando tanto as respostas, quanto as perguntas relacionadas a elas. O estudo também analisa de forma sistemática a complexidade e o grau de dificuldade da anotação humana em questões oriundas de diferentes temas. O estudo demonstrou que pode ser difícil alcançar a **tomada de decisão por máquina** no contexto do estudo, uma vez que os experimentos realizados indicaram que **um tópico facilmente rotulado por um ser humano pode ser difícil de rotular por uma máquina e vice-versa**. O estudo apontou possibilidades de aplicação de taxonomias para auxílio às análises linguísticas por meio de máquinas.

Ainda em 2019, Wang et al. (2019) conduziram uma pesquisa experimental com o objetivo de confirmar a efetividade do modelo *Taxonomy Enhanced Adversarial Learning* (TEAL). A avaliação foi feita por meio de três padrões de tarefas de PLN: classificação de hiperonímia não supervisionada (U-TEAL); detecção de hiperonímia supervisionada (S-TEAL) e inferência de grades lexicais. Para os autores, não está claro o quanto as taxonomias podem beneficiar essas tarefas numa estrutura unificada. Por isso propõem uma taxonomia aperfeiçoada por aprendizagem adversarial supervisionada, o modelo AS-TEAL de **predição de hiperonímia**. A ideia é **empregar grandes taxonomias para aprender como um termo se projeta para seus hiperônimos e não-hiperônimos em vetores de palavras**. O experimento utilizou a taxonomia Probase por meio do Microsoft Concep Graph, o modelo Glovec

treinado em *corpus* da Wikipedia e Gigaword e a dimensionalidade adotada para os vetores foi de 100 em todos os testes. A Figura 16 foi extraída do estudo e adaptada apenas para fins de demonstração das taxonomias em um sistema de aprendizagem. Elas fornecem um grande número de relações de hiperonímia para aprimorar o processo de aprendizagem adversarial em modelos preditivos. A figura ressalta o uso de taxonomias no treinamento de vetores de palavras no modelo proposto, que: (a) com uso de redes neurais e medida não supervisionada U-TEAL distingue a hiperonímia de outras relações semânticas; e (b) na medida supervisionada AS-TEAL, com apoio de redes neurais e do algoritmo de aprendizagem de máquina *Support Vector Machine* (SVM), utiliza taxonomias como forma de transferência de conhecimento hierárquico para os modelos de predição de hiperonímia.

Figura 16. Uso de taxonomias na arquitetura do modelo TEAL proposto por Wang et al., 2019



Fonte: adaptado de Wang et al., 2019

Já no artigo denominado *A family of fuzzy orthogonal projection models for monolingual and cross-lingual hypernymy prediction,* Wang et al. (2019a) propõem

uma família de modelos de projeção ortogonal difusa para abordagem de tarefas de **predição de hiperonímia monolíngue e multilíngue**, são elas: projeção Multi-Wahba, *Multi-Wahba Projection* (MWP), para distinguir relações entre hiperonímia monolíngue e não-hiperonímia; Projeção Multi-Wahba de Transferência, *Transfer Multi-Wahaba Projection* (TMWP), para predição de hiperonímia multilíngue; e para aumento de dados de treinamento, o *Iterative Transfer Multi-Wahba Projection* (ITMWP). Para a tarefa monolíngue, é modelado explicitamente como os vetores de um termo são mapeados para os de seus hiperônimos ou não-hiperônimos, respectivamente, a fim de evitar o que os autores chamam de “memorização lexical”. Para lidar com irregularidades linguísticas complicadas, para cada componente latente das relações de hiperonímia e não-hiperonímia, um mapeamento ortogonal difuso é estabelecido no espaço vetorial. O processo de aprendizagem dos mapeamentos está intimamente relacionado a uma versão generalizada do problema de Wahba em matemática, denominada Projeção Multi-Wahba ou WMP. O estudo deriva uma solução de forma fechada para MWP e propõe uma rede neural baseada em MWP para distinguir relações de hiperonímia versus não-hiperonímia.

O ano de 2020 inicia com Daraio et al. (2020) e a introdução do conceito de engenharia da ontologia. O artigo propõe uma abordagem automatizada para mapeamento de taxonomia de diretórios da *web* com base na categorização de texto. Introduz o conceito de **Inteligência Artificial Explicável** ou ***EXplainable Artificial Intelligence* (XAI)**, que visa combater a **falta de transparência dos modelos populares de aprendizagem de máquina**, e no contexto do estudo busca tornar os **modelos interpretáveis por especialistas de domínio**. O estudo explorou duas representações de texto complementares baseadas em palavras para lidar com hierarquias complexas e de múltiplos níveis: uma baseada em ocorrência, que captura semelhanças sintáticas de texto, e outra vetorial, que destaca as relações semânticas subjacentes entre as palavras. O mapeamento utiliza **aprendizagem profunda para combinação de duas taxonomias**, uma externa que enriquece outra interna. Nesse sentido é realizada uma limpeza nas descrições textuais dos elementos da taxonomia; eliminação de palavras irrelevantes; modelagem de dados; categorização de texto, objetivo final; e explicação do modelo local e global, por meio de evidências das razões por trás de uma atribuição de categoria específica, bem como uma descrição do modelo de previsão baseado em palavras. Foi utilizado um modelo de

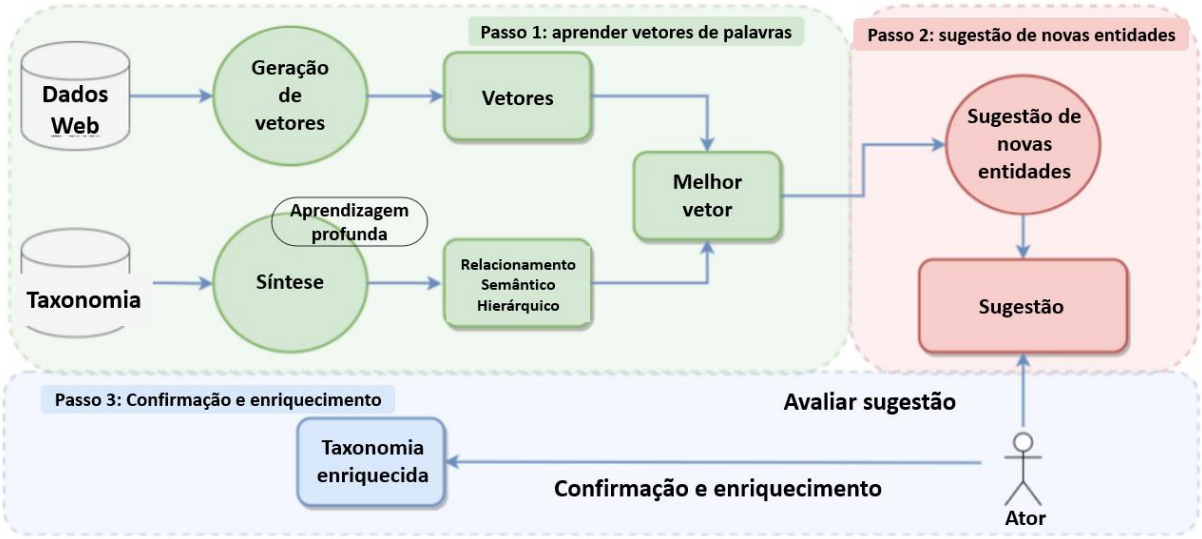
processamento profundo de linguagem natural para explicar as correlações entre descrições de serviços/produtos e categorias de taxonomia. O principal objetivo da etapa de categorização de texto foi a construção de um **modelo preditivo** capaz de atribuir automaticamente as categorias mais prováveis a um novo elemento de taxonomia (para o qual o rótulo é desconhecido). Como resultado dos experimentos realizados destacou-se a representação de texto baseada em frequência, que desempenhou positivamente no procedimento automatizado de mapeamento de taxonomia.

Doan et al. (2020) tiveram como enfoque a extração taxonômica a partir de um *corpus* textual por meio de uma abordagem de dois estágios principais: **extração de conceitos e extração de relações taxonômicas**. O modelo completo contou com o uso de processamento de linguagem natural e envolveu especialmente: filtragem de termos pela medição de sua frequência e relevância, além da **padronização lexical** das palavras que aparecem em um termo multipalavras; determinação automática do significado correto das palavras ambíguas; consultas à taxonomia WordNet para incremento semântico; agrupamento e rotulagem de termos de maior ocorrência, que constituem os dados de entrada para a extração das relações taxonômicas; obtenção de candidatos a hiperônimos de um conceito; consultas *web* com base na combinação de padrões lexicais e informações semânticas ou contextuais. Por fim, o modelo foi testado através de uma aplicação com *corpora* de entrada em documentos de texto nos domínios do turismo e do desporto. Os resultados demonstraram que o modelo proposto foi capaz de extrair eficientemente as relações taxonômicas de *corporas* de texto, mesmo quando os dados são insuficientes ou carecem de treinamento. Na avaliação do modelo foi demonstrada alta precisão e recordação na extração de termos-chave dos conjuntos de dados, superando um método de última geração em conjuntos de dados com altos resultados nas técnicas de **desambiguação**.

Já Giabelli et al.(2020) apresentaram a extensão de um trabalho apresentado em 2018 na Conferência Internacional de Web Semântica e demonstraram o NEO, um sistema desenvolvido no âmbito das atividades de pesquisa em andamento na União Européia (UE) para disponibilização do primeiro sistema de monitorização em tempo real do mercado de trabalho e classificação de anúncios reais de todos os países da UE em 32 línguas. O sistema busca **enriquecer automaticamente a Taxonomia** europeia de ocupações e competências ESCO, por meio de

aprendizagem profunda. Foram propostos nesse sentido, uma nova métrica de similaridade semântica e um conjunto de medidas que estimam a adesão de novos termos a cada conceito. O NEO analisa características de conjuntos de palavras que ocorrem em contextos semelhantes e por isso tendem a ter um significado semelhante. O fluxo de processamento do sistema ocorre em três etapas principalmente: síntese de vetores de palavras, sugestão de novas entidades e votação e enriquecimento. A Figura 17 foi extraída e traduzida do estudo para demonstrar em cada passo a atuação da taxonomia e seus elementos constituintes. Além disso, foram empreendidos dois testes de hipóteses estatísticas que confirmaram a correlação entre as métricas GASC propostas do NEO e os julgamentos dos usuários. O NEO foi avaliado em mais de 2 milhões de anúncios de emprego no Reino Unido a partir de 2018, identificando 49 novas ocupações das quais 43 foram confirmadas por meio de um estudo de usuários com 10 participantes. A correlação entre os julgamentos dos usuários e as métricas GASC do NEO foi positiva e estatisticamente significativa.

Figura 17. Representação do fluxo de processamento dos principais módulos do sistema NEO, de Giabelli et al., 2020



Fonte: traduzido e adaptado de Giabelli et al., 2020, p. 16035

Huang et al.(2020) o artigo discutiu a construção e avaliação de uma taxonomia de tópicos denominada CoRel, que toma como entrada um *corpus* e uma **taxonomia semente** descrita por nomes de conceitos e constrói uma taxonomia mais completa baseada nos interesses dos usuários, onde cada nó é representado por um

agrupamento de termos coerentes. A CoRel completa a estrutura da taxonomia por um módulo de transferência de relação e enriquece a semântica dos nós de conceito por um módulo de aprendizagem. O módulo de transferência de relação aprende a relação de interesse do usuário preservada nos pares iniciais pai-filho e, em seguida, transfere-a ao longo de vários caminhos para expandir a taxonomia em largura e profundidade. O módulo de aprendizagem de conceitos encontra agrupamentos de tópicos discriminativos para cada conceito no processo de incorporação conjunta de conceitos e palavras. Experimentos extensos mostraram que ambos os módulos funcionam de forma eficaz na geração de uma taxonomia tópica de alta qualidade baseada em sementes fornecidas pelo usuário.

Manzoor et al. (2020) apresentaram um estudo experimental com o Arborist, uma abordagem de **expansão automática de taxonomias** que combina informações taxonômicas globais e locais para inferir os papéis dos nós sem supervisão (treinamento) explícita para classificação de funções. Por meio de avaliação extensiva em uma taxonomia com curadoria no Pinterest e em vários conjuntos de dados públicos, demonstram que o Arborist supera o estado da arte da expansão taxonômica, alcançando até 59% na classificação recíproca média e 83% de precisão entre as 15 principais previsões feitas pelo sistema.

Em diálogo, Shen et al.(2020) também propuseram um **modelo de expansão taxonômica**. No artigo os autores estudaram como expandir uma taxonomia existente adicionando um conjunto de novos conceitos e de modo a preservar as relações existentes nas taxonomias base. Foi proposta então uma nova estrutura autosupervisionada, chamada TaxoExpan, que gera automaticamente um conjunto de conceitos de consulta, formados por conceitos âncora e pares da taxonomia existente como dados de treinamento. Com base nesses dados a TaxoExpan aprende um modelo para prever se um conceito de consulta é o hipônimo direto de um conceito âncora. Uma rede neural gráfica incorpora informações posicionais hierárquicas. Cada conceito na taxonomia existente é visto como uma consulta, enquanto seus conceitos parentais são utilizados como uma âncora de verificação da confiabilidade do par. Foi avaliada a eficácia da estrutura TaxoExpan em três taxonomias reais de diferentes domínios: a taxonomia pública de campo de estudo da Microsoft Academic Graph; SemEval ask 14 Benchmark e WordNet 3.0. Os principais resultados do estudo mostraram que a TaxoExpan pode gerar taxonomias conceituais de alta qualidade em

domínios científicos e alcançar desempenho de última geração no desafio de expansão da taxonomia WordNet. Descobriu-se nos experimentos que ela também pode ser utilizada para **limpar uma taxonomia existente** com eficiência e confiabilidade.

Timmaraju et al. (2020) discutiram os desafios na construção de sistemas de classificação de conteúdo *web* e de aprendizagem multilíngue para conteúdo do Facebook (compreendidos como documentos *web*). O principal objetivo foi a classificação de documentos em uma taxonomia de classes que permitisse uma visão holística, já que os documentos *web* muitas vezes se caracterizam por conter não apenas um tipo de dado, mas associação de textos, imagens, vídeos e em alguns casos metadados categoriais, como *hashtags*, que auxiliam a classificação do conteúdo. A arquitetura de análise proposta se baseou no BERT, um modelo pré-treinado de processamento de linguagem natural, que é adaptado para uma configuração multimodal com vetores de pedaços de palavras, visuais e categoriais. Foram realizados experimentos em conjuntos de dados nos idiomas inglês e espanhol em dois tipos de documentos, texto e imagem. A **tradução e classificação multilíngue** foi realizada por meio de redes neurais; *hashtags* ou outros metadados são utilizados no ajuste fino e aprendizagem contínua do modelo demonstrando potencial de contribuição para a **classificação de conteúdo *web* em grande escala**. O estudo apontou que **as taxonomias fornecem uma estrutura semântica para classificar e compreender conteúdo *web*** e destaca os **desafios de rotular documentos dentro de uma grande taxonomia**.

O último estudo de 2020 nessa temática, Zahera e Sherif (2020) propõem o modelo ProBERT de classificação, treinado em conjuntos de dados de websites (títulos, descrições de produtos etc.) para **aprendizagem de representação semântica**. O objetivo era a **mineração de dados** de produtos *web* e **predição de categorias** que dão origem a uma taxonomia hierárquica. O modelo ProBERT é baseado no modelo neural BERT, que categoriza com base em metadados. A abordagem utiliza um modelo BERT simples que representa um único vetor de recursos do título e da descrição do produto e, em seguida, prevê suas categorias. O ProBERT aplica os recursos do BERT ao domínio dos dados do produto, aproveitando sua representação semântica para classificação.

Já em 2021, Caucheteux et al. (2021) propõem uma taxonomia para auxiliar na

fatoração das ativações de alta dimensão de modelos de linguagem construídos por meio de rede neural profunda GPT-2, um modelo de linguagem causal de alto desempenho, treinado para prever palavras a partir do contexto, que é conhecido por gerar representações semelhantes às do cérebro. O estudo examina como as ativações de **transformadores de linguagem** como o GPT-2 mapeiam linearmente a atividade cerebral durante a compreensão da fala. Quatro classes combinatórias dão suporte à análise: representações lexicais, onde são aplicados vetores de palavras para capturar seu significado básico; composicionais, que codificam como as palavras se combinam para formar frases ou sentenças; sintáticas, extraídas de camadas que codificam os aspectos estruturais da linguagem, como gramática e construção de frases; e semânticas, que diz respeito ao significado dos relacionamentos e que vão além da sintaxe. A **taxonomia ajuda a isolar as representações de construções linguísticas na atividade cerebral com vistas a fazer previsões simples, precisas e quantificáveis que auxiliem na investigação das representações linguísticas distribuídas no cérebro humano**. Os resultados dos experimentos empreendidos sugerem que as representações composicionais recrutam uma rede cortical mais difundida do que as lexicais e que a sintaxe e a semântica compartilham um substrato neural comum.

Martel e Zouaq (2021) retornam em 2021 para mostrar como o agrupamento hierárquico não supervisionado sobre vetores de incorporação pode ser usado para **extração de taxonomias**. O artigo explora o problema da necessidade de **incorporação de novos conceitos em taxonomias base de forma automatizada** e apresenta um método que utiliza agrupamento hierárquico não supervisionado sobre vetores de incorporação em grafos do conhecimento. Os vetores de incorporação de entidades são representações vetoriais densas das entidades de um gráfico: eles são baseados nas triplas em que cada entidade está envolvida, de modo que entidades semelhantes no gráfico tenham incorporações semelhantes no espaço vetorial. Ao aplicar agrupamento hierárquico a esses vetores obtém-se grupos de vetores geometricamente próximos, cujas entidades são similares semanticamente e gera-se uma árvore, que é então transformada em uma taxonomia por meio do mapeamento de tipos conhecidos em agrupamentos relevantes. Os resultados demonstram que, ao contrário dos métodos baseados em agrupamento supervisionado, esta abordagem também pode identificar novas classes a partir dos dados e rotulá-las com axiomas

lógicos, o que leva a um **aprendizado expressivo de taxonomias**. Para avaliação da abordagem a versão 2016-10 da DBpedia foi adotada. Para tanto, as instruções *rdfs:subClassOf* foram extraídas. Na visão dos autores os resultados demonstraram que a abordagem proposta foi capaz de criar uma visão rica e de alto nível do conteúdo de um grafo do conhecimento sem supervisão humana.

Prince et al.(2021) apresentaram o *Drug-Use Insights* (DUI), uma aplicação *web* pública e de código aberto desenvolvida por meio da coleta de dados de milhões de postagens de usuários de drogas nas **mídias sociais** ao abordarem o assunto do vício. A taxonomia hierárquica do DUI possui 108 categorias diferentes relacionadas com a dependência, consistindo em mais de 9.000 termos, onde cada categoria contém um conjunto de termos semanticamente relacionados e estabelecidos com base no documento do relatório da *Drug Enforcement Administration* (DEA), por **análise temática** e por vetores de palavras gerados automaticamente na análise das postagens. Para estudar a aplicação do DUI a dados de uma plataforma de mídia social que não foi utilizada na sua criação, foi analisado um conjunto de 318 *tweets* do Twitter e os resultados avaliados manualmente. O desempenho do DUI neste conjunto de dados foi comparável ao das postagens do Reddit. Concluiu-se então que o DUI foi eficaz na **identificação de termos relacionados ao uso de drogas** e pode contribuir para a análises semânticas no domínio. Uma **taxonomia de domínio** foi construída manualmente e **incrementada por meio de aprendizagem computacional**.

Em *Taxonomy Completion via Triplet Matching Network*, Zhang et al. (2021a) lançaram uma nova abordagem para **manutenção de taxonomias** pela inclusão automática de novos conceitos considerando não apenas hiperônimos, mas os pares hiperônimos-hipônimos para uma nova forma de pesquisa de conceitos dentro do modelo de dados, o que denomina **conclusão de taxonomia**. Assim, uma nova tarefa de aprendizagem é formulada e denominada *Triplet Matching Network* (TMN) e contrapõe os já sintetizados estudos de Shen et al. (2020) e Manzoor et al. (2020). As descobertas e as abordagens dos dois autores se basearam na busca por hiperônimos, enquanto a TMN se mostra capaz de encontrar relações mais complexas pela análise dos pares pai-filho. Experimentos em quatro aplicações do mundo real avaliaram a abordagem proposta e concluíram que o TMN apresentou a melhor performance em ambas as **tarefas de preenchimento de taxonomias** e tarefas

prévias de **expansão automatizada**. Os experimentos foram aplicados aos *datasets* Microsoft Academic Graph (MAG), que inclui MAG-Psychology e MAG-CS; e WordNet, com WordNet-verb e WordNet-Noun.

No ano de 2022 aparece nesta temática o estudo adicionado na etapa de leitura completa e identificado por meio do trabalho de Arous et al. (2023), sintetizado adiante. Jiang et al.(2022) apresentaram o TaxoEnrich, uma estrutura de **conclusão de taxonomia** que combina características semânticas e informações estruturais por meio do modelo de linguagem pré-treinado SciBERT, que supera outros modelos como o BERT. A proposta teve como enfoque a **conclusão da taxonomia** para **previsão de trigêmeos ⟨consulta, hiperônimo, hipônimo⟩** mais prováveis para um determinado conceito de consulta. Contemplou a **incorporação contextualizada de taxonomia**; a codificação de informações estruturais da taxonomia para aprender as representações das posições dos candidatos; codificação e agregação adaptativa com base na correspondência entre a consulta e posição; e um modelo de correspondência de posição de consulta que pode ser replicado em diferentes contextos. Os experimentos conduzidos demonstraram que o TaxoEnrich alcançou melhor desempenho entre todas as métricas de avaliação e métodos de última geração considerados, tais como: Arborist de Manzoor et al. (2020); TaxoExpan de Shen et al. (2020), considerado o estado da arte da expansão de taxonomias; STEAM de You et al. (2020); TMN de Zhang et al. (2021) e Gentaxo de Zeng et al. (2021). Foram adotados os mesmos conjuntos de dados de avaliação do TaxoExpan, MAG-CS, MAG-PSY, WirdNet-Noun e WordNet-Verb.

No mesmo ano, Li et al.(2022) propuseram a *Collaboratively Boosting Framework* (CBF), uma **estrutura de impulsionamento colaborativo para segmentação semântica** de imagens de *Remote sensing* (RS) ou sensoriamento remoto, que combina dois módulos baseados em dados: um que utiliza *Deep Semantic Segmentation Network* (DSSN), uma rede de segmentação semântica profunda, e outro voltado ao raciocínio ontológico guiado pelo conhecimento de maneira iterativa. Duas medidas principais formam a base do estudo, o canal de estimativa de sombras, que ajuda na identificação de sombras nas imagens de sensoriamento para corrigir erros de classificação devido à oclusão de luz, e o canal de estimativa de elevação, que distingue objetos com base em sua elevação para melhorar a capacidade discriminativa do classificador. Dessa forma, o módulo de

aprendizagem profunda integra a imagem original e os canais inferidos como entrada para aprendizagem. Já o módulo de raciocínio ontológico funciona por meio do conhecimento de domínio existente na taxonomia da ontologia, que segundo os autores é a chave para melhorar o desempenho da classificação. O raciocínio intra taxonomia corrige as classificações incorretas advindas da aprendizagem de máquina, enquanto o extra taxonomia realiza inferências e as integra de volta à rede como canais de entrada adicionais, que incluem estimativas de sombra e elevação. Esse processo iterativo de empregar o **raciocínio taxonômico** em canais leva ao aprimoramento mútuo nos dois módulos. É importante esclarecer, conforme apontado no estudo, que há limitações significativas que correspondem à baixa interpretabilidade e confiabilidade dos métodos de aprendizagem profunda atuais e o fato de ainda ser rara a adoção de raciocínio ontológico no domínio do sensoriamento remoto. Para validação da proposta foram conduzidos experimentos em dois conjuntos de dados abertos sobre sensoriamento remoto: UCM e ISPRS Potsdam. Os resultados mostraram que o CBF supera o *Semantic Referee*, usado como método comparativo, nos dois conjuntos de dados e alcança melhor desempenho de classificação e interpretabilidade. O método CBF melhora a segmentação semântica, especialmente para as categorias Vegetação, Solo, Pavimento e Água.

Já este ano, Arous et al.(2023) propõem o TaxoComplete, uma estrutura autosupervisionada (aprende com seus próprios dados) de preenchimento de taxonomia que aprende a representação de nós aproveitando sua posição na taxonomia. A estrutura constrói pares de nós candidatos e os usa para aprender um modelo de correspondência semântica aprimorado por posição. Ele mostra conceitos na taxonomia existente e os associa a um conjunto de âncoras, que inclui nós na vizinhança próxima e distante de um nó de consulta. Cada par de nó de consulta e nó âncora é rotulado com uma pontuação proporcional à distância do grafo que os separa na taxonomia. A estrutura utiliza um **mecanismo de correspondência semântica** de âncora de consulta e um módulo de propagação com reconhecimento de direção para combinar semanticamente os nós na taxonomia, o que otimiza a aprendizagem. O estudo também valida a estrutura proposta em quatro conjuntos de dados do mundo real, a saber: SemEval (baseado no WordNet 3.05) com dois conjuntos derivados, o SemEval-Noun de substantivos e o SemEval-Verb de verbos; e MAG-WIKI, derivado do conjunto de dados original do Microsoft Academic Graph (MAG), é um grafo

heterogêneo de campos de estudo e relações taxonômicas onde cada conceito foi correlacionado à sua definição da Wikipedia e derivados os conjuntos MAG-CS-WIKI e MAG-PSY-WIKI, das áreas de ciência da computação e psicologia, respectivamente. O Arborist proposto por Manzoor et al. (2020), o TaxoExpan de Shen et al. (2020), o TMN de Zhang et al.(2021c) e o TaxoEnrich de Jiang et al. (2022) serviram de linha de base para comparação. Os resultados demonstraram que o TaxoComplete supera os métodos de última geração ao prever com precisão a posição de um nó de consulta na taxonomia. Especificamente, descobriu-se que o TaxoComplete é 2 vezes mais eficaz em termos de HIT @k, que é uma métrica de avaliação comumente usada para tarefas de conclusão de taxonomia, isto é, adição de novos conceitos em qualquer nível dentro de uma taxonomia existente para manter sua relevância e precisão.

Bongiovanni et al. (2023) apresentou recentemente um método de classificação textual com base em uma taxonomia hierárquica personalizada e sem dados rotulados (anotados), que utiliza **modelagem profunda de linguagem**, do inglês *Deep Language Modeling* (DLM). Trata-se de uma modelagem projetada para compreender e gerar texto de maneira sofisticada, com uso de redes neurais profundas, para aprender padrões complexos e representações semânticas por meio de treinamento em grandes conjuntos de dados. É utilizada a abordagem *Zero-Shot Text Classification*, que se encaixa perfeitamente em tarefas que onde não há anotação de dados, uma vez que se refere à **capacidade de um sistema de reconhecer ou classificar durante o treinamento os dados textuais que nunca viu antes**. Os autores combinaram elementos para entender como cada documento relacionava-se semanticamente a cada rótulo da taxonomia. A abordagem foi validada testando cada etapa em três conjuntos de dados anotados disponíveis publicamente: Web Of Science (WoS), DBpedia Extract e Amazon Product Reviews. Os resultados demonstraram que a Classificação de Texto Semântico Zero-Shot foi capaz de produzir pontuações de relevância de última geração para rótulos por meio da taxonomia e apresentou melhora considerável após a implementação de uma técnica denominada Upwards Score Propagation (USP), que aproveita a estrutura hierárquica da taxonomia para reforçar a abordagem conduzida nos experimentos nos três conjuntos.

Por fim, Horta et al. (2023) tiveram como objetivo promover interpretabilidade do

conteúdo semântico e da estrutura conceitual que as *Convolutional Neural Networks* (CNN) - redes neurais convolucionais - adquirem a partir de seus dados de treinamento. Os autores apontam que os **avanços recentes em Explainable AI (XAI) permitiram às CNN aprender relações hierárquicas entre categorias semânticas na forma de classificações taxonômicas**. Assim, é proposto um método de extração de taxonomia para derivar uma taxonomia de domínio de uma CNN treinada, o que permite que seja avaliada a adequação semântica ou o nível de confiança da taxonomia extraída em relação à verdade fundamental, isto é, à precisão dos resultados da aprendizagem de máquina em relação ao mundo real. A abordagem é testada usando VGG-16, ResNet-18, ResNet-152 e treinada em CIFAR 100 e ImageNet. Resultados experimentais mostraram que as CNN podem aprender bem as relações de subclasses, especialmente quando são usadas ResNets com reconhecimento de hierarquia. A qualidade do conjunto de dados e da profundidade da arquitetura influenciam diretamente na adequação semântica. Os autores observam também que as técnicas existentes para injetar conhecimento externo aos modelos durante a fase de treinamento podem levar a melhores taxonomias.

4.2.6 Serviços Web e IoT

Apenas dois estudos compõem esta categoria, um de Zakharova (2020), que abordou hierarquia e categorização para compreensão e organização semântica de serviços *web*. Fala sobre correspondência de serviços *web* por meio de descrições contextuais baseadas em Descrição Lógica (DL) ontológica para combinar serviços *web* por contexto, com vistas à resolução de problemas de descoberta de serviços. Todos os serviços *web* possuem características próprias, resolvem problemas diferentes, utilizam métodos diferentes, mas todos possuem uma estrutura comum que permite classificá-los. Para tanto, o estudo determinada uma extensão da ontologia de DL de serviços e utiliza uma ontologia especial de descrição textual geral para cobrir todos os aspectos semânticos importantes das descrições contextuais; para determinar a correspondência entre um serviço e um pedido na *web*; e para categorizar semanticamente os serviços disponíveis. O principal objetivo da abordagem por ontologia é definir um sistema geral de conceitos maximamente completo que permita uma descrição semântica estruturada do contexto de qualquer serviço, independentemente do escopo de sua aplicação ou propósito de negócio. Mas somente quando o sistema de conceitos e a base de conhecimento apropriada

forem adequados e tiverem um alto grau de completude, tal ontologia poderá auxiliar na resolução eficaz de problemas de serviços da *web*.

E Huang et al.(2021), cujo estudo propôs uma **taxonomia para interações com eletrodomésticos e detecção de conflitos em casas inteligentes**, do domínio da *Internet of Things* (IoT) ou internet das coisas; um grafo do conhecimento para capturar as relações entre os serviços de IoT e propriedades do ambiente; e um algoritmo de detecção direta e indireta de conflitos baseado na taxonomia estruturada em quatro tipos de conflitos: um direto, *Function-function conflict*, e três indiretos, *Opposite-environment-impact conflict*, *Cumulative-environment-impact conflict* e *Transitive-environment-impact conflict*. Os resultados experimentais demonstraram o desempenho, a eficiência e a escalabilidade da estrutura proposta, além da detecção automática de conflitos, sem intervenção humana.

5 DISCUSSÃO E CONCLUSÃO

A presente pesquisa, uma revisão integrativa da produção científica disponível nos últimos cinco anos, buscou atualizar e sistematizar o estado da arte das taxonomias na ciência da informação e sua relação com a *web* semântica, de modo a captar pesquisas recentes e traçar convergências entre as temáticas.

Cabe retomar o histórico de desenvolvimento das taxonomias e seus usos no contexto da *web* semântica. No passado enormes esforços foram feitos no sentido da construção manual de taxonomias, que contava com a atuação de especialistas de domínio e, na maioria dos casos, consideráveis recursos. Em seguida, as formas semiautomáticas ganharam força, à medida que o conhecimento se expandia no ambiente virtual e as atividades de *e-commerce* se intensificavam no mundo. No entanto, o rápido crescimento do volume de dados da *web* logo demandou a evolução de técnicas que pudessem contribuir com estrutura e significado, o que favoreceu a demanda pela construção de taxonomias e expansão das existentes, iniciadas e em alguns casos mantidas por especialistas de domínio. A manutenção manual se tornava muito difícil conforme o volume crescia e os *softwares* auxiliares não apresentavam funcionalidades suficientes para lidar com o tamanho do esforço de garimpo. Inaugurava-se então uma era de uso da aprendizagem de máquina, do campo da inteligência artificial, que iniciou na esteira do desenvolvimento das tecnologias da *web* semântica e vem evoluindo rapidamente até os dias atuais.

As hipóteses da pesquisa se confirmaram a partir dos achados conjuntos das revisões narrativa e integrativa. Desse modo, é possível afirmar que a construção de uma taxonomia amplia efetivamente a capacidade semântica dos objetos digitais disponíveis na *web*, bem como sua interoperabilidade, tornando a rede mais eficiente; a evolução das tecnologias da *web* semântica vem promovendo múltiplos usos para as taxonomias em meio ao aumento do volume informacional; e o estabelecimento da *web* semântica gera desafios que exigem a intervenção humana no manejo de taxonomias para organização do conhecimento em diferentes áreas do saber.

Os resultados apontam importantes diferenças nas formas como as taxonomias foram adotadas ao longo do desenvolvimento do ambiente *web* e dialogam com Lambe (2007), ao considerar que as taxonomias podem contribuir para além da recuperação, desempenhando outros papéis como construção de conhecimento, identificação de novas relações, criação de sentido, gerenciamento de complexidade,

diagnóstico e tomada de decisão. Além das funções usuais, as taxonomias demonstraram versatilidade de aplicações no ambiente da *web* semântica, servindo à estruturação e ao enriquecimento semântico de grandes *corpus* multimodais, informação de contexto, gestão de riscos, detecção de conflitos e manejo de eletrodomésticos em casas inteligentes.

Nesse sentido, alguns estudos apontaram que os esforços de construção de taxonomias sem uma base mais sólida desenvolvida por especialistas de domínio não se mostraram eficazes. As estruturas construídas de forma intuitiva apresentaram vieses e os dados heterogêneos baseados na atuação de usuários no ambiente *web* não foram capazes de fornecer um ponto de partida confiável e autossustentável. Assim, as técnicas de aprendizagem passaram a se concentrar na expansão de confiáveis taxonomias e na indução de novos conceitos e relacionamentos com base na rápida evolução das diferentes áreas em todo o mundo e na crescente atividade da população mundial na *web*. No âmbito dos grafos do conhecimento se tornaram úteis as técnicas de extração taxonômica para facilitar a aprendizagem e a inferência de conceitos e relações.

Ao analisarmos as categorias e o conjunto de estudos selecionados pela revisão integrativa importantes aspectos se revelam, com destaque para estudos considerados de excelente e muito boa qualidade metodológica. Na categoria *classificação para contextualizar, compreender e recuperar informações*, Yang et al.(2019) abordaram a detecção automática de termos de domínios complexos e a importância da análise da camada taxonômica; Zhang et al.(2021c) o uso de taxonomia na categorização das técnicas de ameaça com alta precisão; Smith et al. (2022) utilizaram e atualizaram populares taxonomias de ocupação construídas por especialistas de domínio; enquanto Zhang e Sun (2022) tiveram como objetivo estimar a similaridade existente nas relações hiperônimo/hipônimo, elementos essenciais na estrutura taxonômica. Somente em 2023 aparece o conceito de *linked data* ou dados conectados com Jain (2023), apesar de não ser um conceito novo e representar o objetivo final da *web* semântica desde a sua concepção.

Na categoria *bases de conhecimento e vetores de palavras*, Malandri et al.(2020) propõem uma metodologia para gerar e avaliar vetores de palavras a partir de um *corpus* de texto, preservando as relações de similaridade semântica derivadas de uma taxonomia, considerada essencial na compreensão de máquina e processamento de linguagem natural.

Em Zhang et al.(2021b) um *grafo do conhecimento* com mais de cinco milhões de conceitos refinados extraídos de *logs* de pesquisa ruidosos foi construído tendo taxonomias como etapa de construção associada à mineração de conceitos de cauda longa.

Em *Linked (Open) Data,* Dilorenzo et al. (2020) apresentam uma engenharia de software orientada a reuso no contexto de uma abordagem de desenvolvimento ágil amplamente implementada para gerenciamento de projetos de diferentes áreas. Somente em 2023 surge nos estudos o conceito de LOD, acompanhado de princípios FAIR, com o trabalho de Golpayegani et al. (2023). O estudo se coloca no contexto da lei de inteligência artifical (IA) da União Européia, que tem como enfoque o impacto dos sistemas de IA considerados de “alto risco” para a saúde, a segurança e os direitos, discussão muito atual.

No âmbito da *Inteligência Artificial* se concentra a grande maioria dos estudos com enfoque na automatização de processos ligados às taxonomias. Giabelli et al. (2020) abordaram o enriquecimento automático de uma taxonomia; Shen et al. (2020) e Manzoor et al. (2020) abordaram e expansão automática de taxonomias; Zhang et al. (2021a) lançaram uma nova abordagem para manutenção de taxonomias, pela inclusão automática de novos conceitos considerando não apenas hiperônimos, mas os pares hiperônimos-hipônimos. Wang et al. (2019) empregaram grandes taxonomias para aprender como um termo se projeta para seus hiperônimos e não-hiperônimos em vetores de palavras, com abordagens monolíngue e multilíngue. Doan et al. (2020) tiveram como enfoque a extração taxonômica a partir de um *corpus* textual por meio de uma abordagem de dois estágios principais, extração de conceitos e extração de relações taxonômicas. Figueroa et al. (2019) apresentaram o resultado da fusão de taxonomias e sinalizaram que um tópico facilmente rotulado por um ser humano pode ser difícil de rotular por uma máquina e vice-versa.

Na esteira do desenvolvimento das IA, consideráveis avanços surgem a partir de 2022. Jiang et al. (2022) apresentaram o conceito de conclusão de taxonomia, que combina características semânticas e informações estruturais por meio do modelo de linguagem pré-treinado SciBERT, uma especialização do modelo de processamento de linguagem natural BERT. Enquanto Horta et al. (2023) afirmam que avanços recentes em IA explicável ou XAI permitiram às redes neurais convolucionais aprender relações hierárquicas entre categorias semânticas na forma de classificações taxonômicas. E no uso de IA para a compreensão de conteúdo multimídia destaca-se o estudo de Li et al. (2022), que propuseram a estrutura CBF de impulsionamento colaborativo para segmentação semântica de imagens. Essa evolução vai além do emprego de IA em *corpus* textuais e serve a uma diversidade de plataformas como Instagram e Youtube, cujo conteúdo se concentra nas formas de imagem e vídeo.

E na categoria *serviços web e IoT* destaca-se o estudo de Huang et al. (2021), que propuseram uma taxonomia para interações com eletrodomésticos e detecção de conflitos em casas inteligentes.

Como na visão de Hjørland (2016) sobre Processos de Organização do Conhecimento (POC) e Sistemas de Organização do Conhecimento (SOC), as taxonomias parecem ter assumido uma multiplicidade de versões com os recentes avanços na aprendizagem de máquina, onde atuam às vezes como um processo de construção semântica, às vezes como sistema estruturante. Tal compreensão pode contribuir para o desenvolvimento de abordagens mais efetivas no tocante à complexidade do conhecimento.

A obra *The accidental taxonomist*, de Heather Hedden, já possui três edições que acompanham desde 2009 o desenvolvimento do tema. A autora propõe na edição mais recente, publicada em 2022, a adoção de um significado mais amplo para taxonomias, de forma a englobar classificação hierárquica, taxonomias facetadas, listas de metadados, tesauros e entidades conectadas com ontologias. A multiplicidade de funções encontradas atualmente para as taxonomias coloca um ponto de reflexão sobre tal ideia no sentido da necessidade de uma nova conceituação e da elaboração de boas práticas embasadas pelos princípios classificatórios amplamente desenvolvidos no arcabouço da biblioteconomia e da CI. Também, não apenas no âmbito das taxonomias, mas talvez no estudo dos conceitos da área de organização do conhecimento frente à nova realidade tecnológica.

Uma das principais atualizações da última edição foi a mudança da base que as taxonomias tinham no modelo tradicional de tesauro para o modelo SKOS, padronizado pelo W3C. Trata-se de uma evolução que acompanha a necessidade global de recuperação de informações de diferentes formas e níveis de granularidade em ambientes heterogêneos de grande volume de dados, um contexto de mudança de paradigma dos termos enquanto *strings* para uma visão centrada no conceito e seus rótulos. Os mecanismos de busca e suas funcionalidades, que recuperam um conjunto considerável de itens e os organiza e apresenta de diversas formas já não basta, é preciso pesquisar de maneira semântica (Hedden, 2022).

A autora também sinaliza que atualmente há um comportamento mais íntimo entre as taxonomias e as tecnologias da *web* semântica, conforme demonstrado pelos achados, citando como exemplo as ontologias, que anteriormente eram tidas como substitutas das taxonomias, mas hoje se conectam a elas e as expandem, promovendo novos usos. Além disso, com maior capacidade de conexão entre computadores e servidores, se tornou comum a integração de *software* com outras aplicações por meio de API, o que modificou a maneira de utilizar taxonomias existentes, dispensando a exportação e importação e possibilitando a integração em ferramentas que oferecem também outras funcionalidades relacionadas, como a etiquetagem ou marcação automática (*auto tagging*). Afirma também, que com o aperfeiçoamento da inteligência artificial verifica-se hoje uma variedade maior de aplicações das taxonomias (inclusive as construídas manualmente), que estão mais valiosas especialmente para a aprendizagem de máquina, abrindo espaço para os profissionais que a autora denomina taxonomistas, isto é, especialistas em informação, bibliotecários, arquitetos ou engenheiros da informação, entre outros. Trata-se dos mesmos profissionais apontados nos estudos como fundamentais para validação e confiabilidade das taxonomias construídas, expandidas ou induzidas automaticamente.

No aspecto interdisciplinar, conforme os resultados da RI e a definição de Barité et al. (2015, p. 120) “[...] a organização do conhecimento é alimentada pelas contribuições recebidas da Ciência da Computação, Linguística, Terminologia e Ciências da Informação”. Atualmente, adicionam-se à interdisciplinaridade da CI os campos da ciência de dados, linguística cognitiva, linguística computacional e IA.

Estamos presenciando uma corrida tecnológica pelo desenvolvimento de redes neurais cada vez mais avançadas. O surgimento de estudos onde são propostas, avaliadas ou utilizadas como objeto de pesquisa variações dos modelos BERT e GPT não ocorre sem razão fundamentada, assim como a presença de marcas como OpenAI, Microsoft, IBM, Google, Facebook (hoje denominada Meta) e Pinterest. Todas as marcas citadas tem como foco do negócio o desenvolvimento de ferramentas *web* baseadas em conteúdo multimídia. Há consideráveis investimentos sendo empreendidos para compreensão das linguagens nessas diferentes modalidades de documento publicadas e compartilhadas em mídias sociais e outras plataformas de manipulação de conteúdo em texto, áudio (falada), vídeo (fala e expressões), entre outros. A evolução dos transformadores de linguagem e das técnicas de construção e manipulação de grafos do conhecimento conferem novos e importantes usos às taxonomias na organização e representação do conhecimento, aliadas às ontologias e embasadas pelas especificações do W3C, às quais já dialogam com as normas ISO. Nesse sentido, se torna fundamental a realização de pesquisas sobre o desenvolvimento de SOC, especialmente sobre taxonomias e seus novos usos, além de políticas públicas que favoreçam o acesso à capacidade computacional e de internet necessária ao avanço de pesquisas científicas ligadas à organização do conhecimento em tempos de IA.

A presente RI buscou preencher lacunas de conhecimento e mitigar a escassez de revisões sistemáticas encontradas na CI, especialmente quando relacionadas ao ambiente *web*, conforme estudo de Lopes, Aganette e Maculan (2020). Com a apresentação de um glossário e de uma taxonomia bilíngue pretende-se também contribuir para a produção científica em língua portuguesa, além do desenvolvimento do tema no Brasil, conforme apontaram Mendes e Bentes (2019) ao identificar maior emprego das taxonomias no âmbito internacional, destacando como provável motivo o desconhecimento das potencialidades que a taxonomia possa trazer para as organizações de um modo geral.

Também, limitações fazem parte de qualquer processo de pesquisa e precisam ser explicitadas. Estudos experimentais de diferentes campos, a maioria da Ciência da Computação, foram analisados e seus dados coletados no âmbito da presente revisão integrativa, de cunho qualitativo, não tendo sido realizadas, portanto, análises aprofundadas sobre todos os aspectos discutidos em cada estudo. A RI buscou

compreender o papel das taxonomias enquanto SOC no ambiente da *web* semântica, tendo coletado todas as evidências cabíveis restritas a este contexto. Além disso, as observações aqui colocadas apresentaram somente a visão e a compreensão da autora a partir dos estudos sintetizados e demais itens do *corpus* da pesquisa.

Compreende-se que a condução de uma revisão integrativa desta magnitude se beneficiaria de uma equipe para estabelecimento de discussões e consensos acerca das escolhas metodológicas e dos processos de análise. No entanto, a pesquisadora buscou principalmente mitigar ao máximo vieses de relato, por meio de consulta a fontes adicionais de diferentes campos, e garantir a aplicação de rigor metodológico em cada etapa de pesquisa pela adoção das melhores práticas recomendadas pelos guias de Toronto e Remington (2020) para condução de revisões integrativas e *Cochrane Library* de revisões sistemáticas, assim como o modelo PRISMA para o registro, representação e relatoria do processo de pesquisa, o que conferiu maior transparência e reprodutibilidade aos procedimentos.

Por fim, há diferentes formas de agregação de dados disponível nos apêndices e nos resultados, os quais também são disponibilizados em formato legível por máquina para promoção de reuso dos dados de pesquisa. Novas pesquisas com base nos estudos selecionados na revisão integrativa, ao serem conduzidos em áreas interdisciplinares à CI como ciência da computação, linguística, linguística cognitiva, linguística computacional, terminologia, ciência de dados e inteligência artificial, beneficiariam a própria CI com novos olhares e diferentes abordagens metodológicas.

REFERÊNCIAS

AGANETTE, E.; ALVARENGA, L.; SOUZA, R. R. Elementos constitutivos do conceito de taxonomia, *[s. l.]*. **Informação & Sociedade**, v. 20, n. 3, p. 77–93, 2010.

ALONSO, O. et al. Answering recreational *web* searches with relevant things to do results. **Information Processing and Management**, v. 57, n. 2, 2020.

ALVARENGA, L.; SILVA, D. L. da. Organização e representação do conhecimento na ciência da informação, *[s. l.]*. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 3, n. 1, 2010.

ALVARES, L.; ARAÚJO JÚNIOR, R. H. de. Marcos históricos da ciência da informação: breve cronologia dos pioneiros, das obras clássicas e dos eventos fundamentais, *[s. l.]*. **Transinformação**, v. 22, n. 3, p. 195–205, dez. 2010.

ALVARES, L. M. A. de R. **Teorias da Informação**. **Disciplina Cronografia da Ciência da Informação**. Programa de Pós-graduação em Ciência da Informação IBICT-UFRJ, 2022.

AQUINO, I. J.; CARLAN, E.; BRASCHER, M. B. Princípios Classificatórios para a Construção de Taxonomias. **Ponto de Acesso**, v. 3, n. 3, p. 196–2015, dez. 2009.

AROUS, I.; DOLAMIC, L.; CUDRÉ-MAUROUX, P. TaxoComplete: Self-Supervised Taxonomy Completion Leveraging Position-Enhanced Semantic Matching. *In*: THE WORLD WIDE WEB CONFERENCE, 2023, Austin. **Proceedings [...]**. New York: Association for Computing Machinery, 2023, p. 2509-2518. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159380259&doi=10.1145%2f3543507.3583342&partnerID=40&md5=b6200c059e86eecfbba92f8b6c5706ac>. Acesso em: 17 set. 2023.

ASIS&T. **History of ASIS&T**, 2023. Disponível em: <https://www.asist.org/about/history-of-asist/>. Acesso em: 17 maio. 2023.

BAPTISTA, D. M. O impacto dos metadados na representação descritiva. **Revista ACB: Biblioteconomia em Santa Catarina**, v. 12, n. 2, 2007.

BARITÉ, M. et al. **Diccionario de organización del conocimiento: clasificación, indización, terminología**. 6. ed. Montevideo: Ediciones Universitarias, 2015.

BERNERS-LEE, T. **Information Management: A Proposal**, 1989. Disponível em: <https://www.w3.org/History/1989/proposal.html>. Acesso em: 20 maio. 2023.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic *web*. **Scientific American**, v. 284, n. 5, p. 34–43, 2001.

BONGIOVANNI, L. et al. Zero-Shot Taxonomy Mapping for Document Classification. *In*: THE ACM SYMPOSIUM ON APPLIED COMPUTING, 38, 2023, Tallinn. **Proceedings [...]**. New York: Association for Computing Machinery, 2023, p. 911-918. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85162908762&doi=10.1145%2f3555776.3577653&partnerID=40&md5=ab7de8de362b392309668e5c1444a6b6>. Acesso em: 17 set. 2023.

BOOTH, A.; PAPAIOANNOU, D.; SUTTON, A. Systematic approaches to a successful literature review. *In*: BOOTH, A.; PAPAIOANNOU, D.; SUTTON, A. (Eds.). **The literature review: its role within research**. [s.l.] Thousand Oaks: Sage Publications Inc., 2012. v. 1, p. 17-35.

BRÄSCHER, M.; CAFÉ, L. M. A. Organização da informação ou organização do conhecimento? *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 9, 2008, São Paulo. **Anais[...]**. São Paulo: USP, 2008, p. 1-14. Disponível em: <https://brapci.inf.br/index.php/res/download/176536>. Acesso em: 29 abr. 2023

BRAUN, V.; CLARKE, V. Using thematic analysis in psychology. **Qualitative Research in Psychology**, v. 3, n. 2, p. 77–101, jan. 2006.

BUCKLAND, M. K. Information as thing. **Journal of the American Society for Information Science**, v. 42, n. 5, p. 351–360, jun. 1991.

CAFÉ, L.; BARROS, C. M.; FONSECA, M. S. da. Organização do conhecimento: uma análise conceitual nos anais do ENANCIB . **Em Questão**, v. 20, n. 1, p. 86–112, 2014.

CAMARGO JÚNIOR, R. N. C. et al. Revisão integrativa, sistemática e narrativa: aspectos importantes na elaboração de uma revisão de literatura. **Revista ACB: Biblioteconomia em Santa Catarina**, v. 28, n. 1, p. 1–15, 2023.

CAMPOS, M. L. de A.; GOMES, H. E. Taxonomia e Classificação: a categorização como princípio/Taxonomy and Classification: the principle of categorization. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 8, 2007, Salvador. **Anais [...]**. Salvador: ANCIB, 2007.

CARLAN, E.; BRÄSCHER, M. Sistemas de Organização do Conhecimento na visão da Ciência da Informação / Knowledge Organization Systems in the Information Science's approach. **Revista Ibero-Americana de Ciência da Informação**, v. 4, n. 2, p. 53–73, 2011.

CASARIN, S. T. et al. Tipos de revisão de literatura: considerações das editoras do Journal of Nursing and Health. **Journal of Nursing and Health**, v. 10, n. 5, 2020.

CATARINO, M. E.; CERVANTES, B. M. N.; ALMEIDA, I. A. de. A representação temática no contexto da *web* semântica. **Informação & Sociedade**, v. 25, n. 3, 2015.

CATARINO, M. E.; SOUZA, T. B. de. A representação descritiva no contexto da *web* semântica. **Transinformação**, v. 24, p. 77–90, 2012.

CAUCHETEUX, C.; GRAMFORT, A.; KING, J.-R. Disentangling Syntax and Semantics in the Brain with Deep Networks. *In*: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 38, 2021. **Proceedings[...]**. Maastricht: ML Research Press, v. 139, p. 1336 – 1348. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85161328616&partnerID=40&md5=21c3b2ce6b9cd645d950277b5fb9dc81>. Acesso em 17 set. 2023.

CERRAO, N. G.; CASTRO, F. F. de; JESUS, A. F. de. O método de revisão sistemática da literatura (RS) na área da Ciência da Informação no Brasil: análise de dados de pesquisa. **Informação & Tecnologia**, v. 5, n. 1, p. 105–116, 2018.

COPE, B.; KALANTZIS, M.; LIAM, M. **Towards a Semantic Web**. [s.l.] Sebastopol: Chandos Publishing, 2011.

CRITICAL APPRAISAL SKILLS PROGRAMME. **CASP Qualitative Checklist**, 2018. Disponível em: <https://casp-uk.net/casp-tools-checklists/>. Acesso em: 6 ago. 2023

CROTTI JUNIOR, A. et al. The ARK platform: Enabling risk management through semantic *web* technologies. *In*: INTERNATIONAL CONFERENCE ON BIOMEDICAL ONTOLOGIES, 11, 2020, Bolzano. **Proceedings [...].** Aachen: CEUR-WS, 2020. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101184475&partnerID=40&md5=7130b8fb5ed8b4f0a6b805acd78a91e7>. Acesso em: 17 set. 2023.

CURRÁS, E. **Ciencia de la informacion bajo postulados sistemicos y sistematicos**. [s.l.] Madrid:Edición personal, 2008.

DAHLBERG, I. **Fundamentos teórico-conceituais da classificação**. Revista de Biblioteconomia de Brasília, 6(1), 9–21, 1978. Disponível em: <https://periodicos.unb.br/index.php/rbbsb/article/view/29057/24860>.

DAHLBERG, I. Knowledge Organization: A New Science? **Knowledge Organization**, v. 33, n. 1, p. 11–19, 2006.

DANILEVITCH, O. Logical semantics approach for data modeling in XBRL taxonomies. *In*: INTERNATIONAL SEMANTIC WEB CONFERENCE, 18, 2019, Auckland. **Proceedings [...]**. Aachen: CEUR-WS, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85081553253&partnerID=40&md5=c38b20c7a75fb20aa55b1d43e08ba175>. Acesso em: 18 set. 2023.

DARAIO, E. et al. An explainable data-driven approach to *web* directory taxonomy mapping. *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE-BASED AND INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, 24, 2020. **Proceedings [...]**. Procedia Computer Science, v. 176, 2020, p. 1101-1110. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85093361917&doi=10.1016%2fj.procs.2020.09.106&partnerID=40&md5=c1d2d52e6d46224577b27cbe01216d52>. Acesso em: 17 set. 2023.

DESLANDES, S. F.; GOMES, R. **Pesquisa social: teoria, método e criatividade**. 25. ed. Petrópolis: Vozes, 2007.

DEXTRE CLARKE, STELLA G. The Information Retrieval Thesaurus. **Knowledge Organization** 46, no. 6: 439-59, 2019.

DILORENZO, E. et al. Enabling the Reuse of Software Development Assets through a Taxonomy for User Stories. **IEEE Access**, v. 8, p. 107285–107300, 2020.

DOAN, P. H.; ARCH-INT, N.; ARCH-INT, S. A semantic framework for extracting taxonomic relations from text corpus. **International Arab Journal of Information Technology**, v. 17, n. 3, p. 325–337, 2020.

FAUCEGLIA, N. R. et al. Automatic taxonomy induction and expansion. *In*: INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 9, 2019, Hong Kong. **Proceedings [...]**. Kerrville: Association for Computational Linguistics (ACL), 2019, p. 25-30. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85087435953&partnerID=40&md5=e6a757718ea22a1913f24ed788ad3a26>. Acesso em: 17 set. 2023.

FERREIRA, J. A.; SANTOS, P. L. V. A. da C. O modelo de dados Resource Description Framework (RDF) e o seu papel na descrição de recursos. **Informação & Sociedade: Estudos**, p. 13–23, 2013.

FIGO, A. O estouro da Bolha PontoCom que quebrou mais de 500 empresas e é uma assombração até hoje. **Infomoney**, São Paulo, ano 2020, 26 mai. 2023. Disponível em: < https://www.infomoney.com.br/mercados/o-estouro-da-bolha-pontocom-que-quebrou-mais-de-500-empresas-e-e-uma-assombracao-ate-hoje/>. Acesso em: 26 mai. 2023.

FIGUEROA, A.; GÓMEZ-PANTOJA, C.; NEUMANN, G. Integrating heterogeneous sources for predicting question temporal anchors across Yahoo! Answers. **Information Fusion**, v. 50, p. 112–125, 2019.

FOGL, J. Relations of the concepts ‘information’ and ‘knowledge’.**The Hague**, v. 4, n. 1, p. 21–24, 1979. International Fórum on Information and Documentation.

GIABELLI, A. et al. NEO: A Tool for Taxonomy Enrichment with New Emerging Occupations. **The Semantic Web -ISWC**, v. 2020, p. 568–584, 2020.

GOES, L. F. de; FERNANDEZ, C. Reflexões metodológicas sobre pesquisas do tipo estado da arte: investigando o conhecimento pedagógico do conteúdo. **Revista Electrónica de Enseñanza de las Ciencias**, v. 17, n. 1, p. 94–118, 2018.

GOLPAYEGANI, D.; PANDIT, H. J.; LEWIS, D. To Be High-Risk, or Not To Be - Semantic Specifications and Implications of the AI Act’s High-Risk AI Applications and Harmonised Standards. *In*: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 6, 2023, Chicago. **Proceedings [...]**. New York: Association for Computing Machinery, 12 jun. 2023, p. 905-915. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163616240&doi=10.1145%2f3593013.3594050&partnerID=40&md5=fca95758c584e64aa1826935e69642bd>. Acesso em: 17 set. 2023.

GOMES, H. E.; CAMPOS, M. L. de A. **A organização do conhecimento na Web: contribuições de Shiyali Ramamrita Ranganathan e Ingetraut Dahlberg**. Niterói: IACS/UFF, 2019. v. 1. Série Cadernos Acadêmicos.

HEDDEN, H. **The accidental taxonomist**. [s.l.] Medford: Information Today Inc., 2022.

HENDLER, J. Web 3.0 Emerging. **Computer**, v. 42, n. 1, p. 111–113, jan. 2009.

HIGGINS, J. P. T. et al. (eds.). **Cochrane handbook for systematic reviews of interventions**. Chichester: John Wiley & Sons, 2019.

HJØRLAND, B. What is Knowledge Organization (KO)? **Knowledge Organization**, v. 35, n. 2–3, p. 86–101, 2008.

HJØRLAND, B. Knowledge Organization (KO). **Knowledge Organization**, v. 43, p. 475–484, 2016.

HJØRLAND, B. Classification. **Knowledge Organization** 44, no. 2: 97-128, 2017.

HJØRLAND, B. The foundation of information science: one world or three? A discussion of Gnoli (2018). **Journal of Documentation**, v. 75, n. 1, p. 164–171, 14 jan. 2019.

HJØRLAND, B. Information Retrieval and Knowledge Organization: A Perspective from the Philosophy of Science. **Information**, v. 12, n. 3, p. 135, 20 mar. 2021.

HLAVA, M. M. K. The Taxobook: Principles and Practices of Building Taxonomies: Part 2 of a 3-Part Series. **Synthesis Lectures on Information Concepts, Retrieval, and Services**, v. 6, n. 4, p. 1–164, nov. 2014.

HORTA, V. A. C. et al. Semantic Interpretability of Convolutional Neural Networks by Taxonomy Extraction. *In*: INTERNATIONAL WORKSHOP ON NEURAL-SYMBOLIC LEARNING AND REASONING, 17, 2023, Siena. **Proceedings [...]**. Achen: CEUR – WS, v. 3432, 2023, p. 118-127. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85167458087&partnerID=40&md5=dd467fdac05fbab65b500b614ef49d82>. Acesso em: 17 set. 2023.

HUANG, B.; DONG, H.; BOUGUETTAYA, A. Conflict Detection in IoT-based Smart Homes. *In*: INTERNATIONAL CONFERENCE ON WEB SERVICES, 2021. **Proceedings [...]**. New York: Institute of Electrical and Electronics Engineers Inc., 2021, p. 303-313. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123201704&doi=10.1109%2fICWS53863.2021.00048&partnerID=40&md5=fe4b193f0a66fc5dbc670a44c4bdbcf6>. Acesso em: 17 set. 2023.

HUANG, J. et al. CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring. *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 26, 2020. **Proceedings [...]**. New York: Association for Computing Machinery, 2020, p. 1928-1936. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090404139&doi=10.1145%2f3394486.3403244&partnerID=40&md5=27e6124ca9d871f5627a7e8ed731de0c>. Acesso em: 17 set. 2023.

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

Disponível em: <https://www.iso.org/obp/ui/en/#iso:std:53658:en>. Acesso em: 18 ago. 2022

ISOTANI, S.; BITTENCOURT, I. Estruturação de Dados e Dados Abertos Conectados. *In*: DADOS ABERTOS CONECTADOS. [s.l.] São Paulo: **Novatec**, 2015.

JAIN, S. Product discovery utilizing the semantic data model. **Multimedia Tools and Applications**, v. 82, n. 6, p. 9173–9199, 2023.

JI, L. et al. Microsoft concept graph: Mining semantic concepts for short text understanding. **Data Intelligence**, v. 1, n. 3, p. 238–270, 2019.

JIANG, M. et al. TaxoEnrich: Self-Supervised Taxonomy Completion via Structure-Semantic Representations. *In*: THE WORLD WIDE WEB CONFERENCE, 2022, Lyon. **Proceedings [...]**. New York: Association for Computing Machinery, 25 abr. 2022, p. 925–934. Disponível em: < https://doi.org/10.1145/3485447.3511935>. Acesso em 18 set. 2023.

KASENCHAK, R. T. What is Semantic Search? and why is it important? **Information Services and Use**, v. 39, n. 3, p. 205–213, 2019.

LAMBE, P. **Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness**. Oxford: Chandos Publishing, 2007.

LE COADIC, Y.-F. **A ciência da informação**. 2. ed. Brasília: Briquet de Lemos, 2004.

LEINER, B. M. et al. A brief history of the internet. **ACM SIGCOMM Computer Communication Review**, v. 39, n. 5, p. 22–31, 7 out. 2009.

LI, Y.; OUYANG, S.; ZHANG, Y. Combining deep learning and ontology reasoning for remote sensing image semantic segmentation. **Knowledge-Based Systems**, v. 243, 2022.

LIMA, G. Â. B. de O. Da organização do conhecimento à recuperação da informação: teorias e técnicas. *In*: **Pesquisa Integrada em Organização do Conhecimento e Recuperação da Informação**. Belo Horizonte: [s.n.], 2016, p. 67-77.

LIMA, J. L. O.; ALVARES, L. Organização e representação da informação e do conhecimento. *In*: ALVARES, L. (Ed.). **Organização da informação e do conhecimento: conceitos, subsídios interdisciplinares e aplicações**. São Paulo: B4 Editores, 2012. v. 248, p. 21–48.

LOPES, M. V.; ALBUQUERQUE, A. C. de. Folksonomias na organização e representação do conhecimento: um estudo na literatura da área / Folksonomies in organization and knowledge representation: a study in the area's literature. **Revista Brasileira de Biblioteconomia e Documentação**, v. 18, p. 1–25, 2022.

LOPES, P. T. D.; AGANETTE, E. C.; MACULAN, B. C. M. dos S. Taxonomia corporativa e taxonomia facetada: usos e aplicações na ciência da informação no

Brasil. **Investigación Bibliotecológica: archivonomía, bibliotecología e información**, v. 34, n. 82, p. 159, 20 jan. 2020.

LOURDUSAMY, R.; JOHN, A. Metric based ontology quality evaluation. **International Journal of Engineering and Advanced Technology**, v. 8, n. 6, p. 3072–3077, 2019.

MACASKILL, W.; OWEN, D. Web 2.0 To Go. *In*: CONFERENCE OF THE LIBRARY AND INFORMATION ASSOCIATION OF NEW ZELANDA AOTEAROA, 2006, Wellington. **Proceedings [...].** Wellington, N.Z.: LIANZA, 2006. Disponível em: < https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=23740f46e5bd35fba28655e084a8828bbe575260>. Acesso em 25 abr. 2023.

MAIA, L. S.; de LIMA, G. Â.; MACULAN, B. C. M. dos S. Taxonomia dos tipos de relações semânticas para a organização e a representação do conhecimento: uma proposta a partir da literatura. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 10, n. 2, 2017.

MALANDRI, L. et al. MEET: A Method for Embeddings Evaluation for Taxonomic Data. *In*: INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS, 2020, Sorrento. **Proceedings [...]**. New York: Institute of Electrical and Electronics Engineers Inc., v. 2020-November, p. 31-38. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101360607&doi=10.1109%2fICDMW51313.2020.00014&partnerID=40&md5=e026395a07afdf9493484aa9978c4f2f>. Acesso em 17 set. 2023.

MALANDRI, L. et al. MEET-LM: A method for embeddings evaluation for taxonomic data in the labour market. **Computers in Industry**, v. 124, 2021.

MALDONADO, A.; KLUBIČKA, F.; KELLEHER, J. Size matters: The impact of training size in taxonomically-enriched word embeddings. **Open Computer Science**, v. 9, n. 1, p. 252–267, 2019.

MANZOOR, E. et al. Expanding Taxonomies with Implicit Edge Semantics. *In*: THE WORLD WIDE WEB CONFERENCE, 29, 2020, Taipei. **Proceedinsg [...]**. New York: Association for Computing Machinery, 20 abr. 2020. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85086581689&doi=10.1145%2f3366423.3380271&partnerID=40&md5=f339f67aa0662c313f9e8fe9ce2e9929>. Acesso em: 17 set. 2023.

MARTEL, F.; ZOUAQ, A. Taxonomy extraction using knowledge graph embeddings and hierarchical clustering. *In*: ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING, 2021. **Proceedings [...]**. New York: Association for Computing Machinery, 22 mar. 2021, p. 836-844. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104930460&doi=10.1145%2f3412841.3441959&partnerID=40&md5=fdbfc74db71d579444b65186762eb7d3>. Acesso em: 18 set. 2023.

MAZZOCCHI, F. Knowledge Organization System. **Knowledge Organization**, 45, no.1: 54-78, 2018.

MELO, M. A. F.; BRÄSCHER, M. Termo, conceito e relações conceituais: um estudo das propostas de Dahlberg e Hjorland. **Ciência da Informação**, v. 43, n. 1, 2014.

MENDES, I.; BENTES, V. Taxonomia nas áreas da Biblioteconomia e da Ciência da Informação: uma revisão sistemática. Porto: **Páginas a&b: Arquivos & Bibliotecas**, v. 12, p. 36–47, 2019.

MENDES, K. D. S.; SILVEIRA, R. C. de C. P.; GALVÃO, C. M. Revisão integrativa: método de pesquisa para a incorporação de evidências na saúde e na enfermagem. **Texto & Contexto - Enfermagem**, v. 17, n. 4, p. 758–764, dez. 2008.

MORAES, A. F. de; ARCELLO, E. N. O conhecimento e sua representação. **Informação & Sociedade: Estudos**, v. 10, n. 2, p. 105–121, 2000.

MOREIRA, W. **Sistemas de organização do conhecimento: aspectos teóricos, conceituais e metodológicos**. Marília: Universidade Estadual Paulista, 2018.

MOTA de SOUSA, L. M. et al. Revisões da literatura científica: tipos, métodos e aplicações em enfermagem. **Revista Portuguesa de Enfermagem de Reabilitação**, v. 1, n. 1, p. 45–55, 23 jun. 2018.

MUSGROVE, T.; WALSH, R. Remediating intentionally corrupted ontology-based data sets. *In*: JOINT ONTOLOGY WORKSHOPS EPISODE V: THE STYRIAN AUTUMN OF ONTOLOGY, 2019, Graz. **Proceedings [...]**. Aachen: CEUR-WS, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077700003&partnerID=40&md5=d3eb9ce2cb84db8fc634ebddaa96e3bc>. Acesso em: 18 set. 2023.

NATIONAL INFORMATION STANDARDS ORGANIZATION. **ANSI/NISO Z39.19-2005** (R2010) Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Baltimore: NISO, 13 mai. 2010. Disponível em: <https://groups.niso.org/higherlogic/ws/public/download/12591/z39-19-2005r2010.pdf>. Acesso em: 18 ago. 2022

NCSA MOSAIC. **NCSA Mosaic: Project Profile**, 2021. Disponível em: <http://www.ncsa.illinois.edu/research/project-highlights/ncsa-mosaic/>. Acesso em: 25 maio. 2023.

OHLY, P. Ingetraut Dahlberg (1927-2017). **Knowledge Organization**,47, no. 2: 173-182, 2020.

O'REILLY, T. **What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software**, 2005. Disponível em: <https://www.oreilly.com/pub/a/web2/archive/what-is-*web*-20.html>. Acesso em: 25 maio. 2023.

PATEL, A.; JAIN, S.; SHANDILYA, S. K. Data of semantic *web* as unit of knowledge. **Journal of Web Engineering**, v. 17, n. 8, p. 647–674, 2019.

PINHEIRO, L. V. R.; FERREZ, H. D. Tesauro Brasileiro de Ciência da Informação. Rio de Janeiro: **Instituto Brasileiro de Informação em Ciência e Tecnologia**, 2014.

POIRIER, L. Classification as catachresis: Double binds of representing difference with semiotic infrastructure. **Canadian Journal of Communication**, v. 44, n. 3, p. 361–371, 2019.

PRINCE, Z.; JHA, D.; SINGH, R. DUI: The drug use insights *web* server. Oxford: **Bioinformatics**, v. 37, n. 24, p. 4895–4897, 2021.

RAZA, M. A. et al. A Taxonomy and Survey of Semantic Approaches for Query Expansion. New York: **IEEE Access**, v. 7, p. 17823–17833, 2019.

ROCHA, L. de L.; SALES, L. F.; SAYÃO, L. F. Descrever para preservar: metadados como ferramenta para gestão de dados de pesquisa. Belém: **ISKO Brasil**, v. 5, p. 194–201, 2017.

RODRIGUES, A. S. P.; SACHINSKI, G. P.; MARTINS, P. L. O. Contribuições da revisão integrativa para a pesquisa qualitativa em Educação. Brasília: **Linhas Críticas**, v. 28, 24 mar. 2022.

ROSA, L. S. da; MACKEDANZ, L. F. A análise temática como metodologia na pesquisa qualitativa em educação em ciências. Blumenau: **Atos de Pesquisa em Educação**, v. 16, p. 8574, 27 abr. 2021.

ROTHER, E. T. Revisão sistemática X revisão narrativa. São Paulo: **Acta Paulista de Enfermagem**, v. 20, n. 2, p. v–vi, jun. 2007.

SALATINO, A. A. et al. The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. **Data Intelligence**, v. 2, n. 3, p. 379–416, 2020.

SANT'ANNA RAMOS VOSGERAU, D.; PAULIN ROMANOWSKI, J. Estudos de revisão: implicações conceituais e metodológicas. Curitiba: **Revista Diálogo Educacional**, v. 14, n. 41, p. 165, 12 jul. 2014.

SAUMURE, K.; SHIRI, A. Knowledge organization trends in library and information studies: a preliminary comparison of the pre- and post-*web* eras. London: **Journal of Information Science**, v. 34, n. 5, p. 651–666, 3 out. 2008.

SAUNDERS, M. N. K.; LEWIS, P.; THORNHILL, ADRIAN. **Research methods for business students**. 8. ed. Harlow: Mark N.K. Saunders, Philip Lewis and Adrian Thornhill, 2019.

SAYÃO, L. F.; SALES, L. F. Afinal, o que é dado de pesquisa? **BIBLOS**, v. 34, n. 2, 31 dez. 2020.

SCHIESSL, M.; SHINTAKU, M. Sistemas de Organização do Conhecimento. *In*: ALVARES, L. (Ed.). **Organização da informação e do conhecimento: conceitos, subsídios interdisciplinares e aplicações**. São Paulo: B4 Editores, 2012. v. 248p. 49–118.

SCHREIBER, G.; RAIMOND, Y. **RDF 1.1 Primer**. [s.l: s.n.]. Disponível em: <https://www.w3.org/TR/rdf11-primer/>. Acesso em: 21 set. 2022.

SHA, A. S.; NUNES, B. P.; HALLER, A. Link Topics from Q&A Platforms using Wikidata: A Tool for Cross-platform Hierarchical Classification. *In*: ACM INTERNATIONAL CONFERENCE PROCEEDING SERIES, 15, 2023, Austin. **Proceedings [...]**. New York: Association for Computing Machinery, 30 abr. 2023, p. 357 - 362. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159172287&doi=10.1145%2f3578503.3583625&partnerID=40&md5=bef89e13e98713d2b28139ce3104a851>. Acesso em: 18 set. 2023.

SHANG, J. et al. NetTaxo: Automated Topic Taxonomy Construction from Text-Rich Network. *In*: THE WORLD WIDE WEB CONFERENCE, 29, 2020, Taipei. **Proceedings [...]**. New York: Association for Computing Machinery, Inc., 20 abr. 2020, p. 1908-1919. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85086584453&doi=10.1145%2f3366423.3380259&partnerID=40&md5=9e42f41128a192fdc8b6ae681b15f193>. Acesso em: 18 set. 2023.

SHARIFPOUR, R.; WU, M.; ZHANG, X. Large-scale analysis of query logs to profile users for dataset search. **Journal of Documentation**, v. 79, n. 1, p. 66–85, 2023.

SHEN, J. et al. TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network. *In*: THE WORLD WIDE WEB CONFERENCE, 29, 2020, Taipei. **Proceedings [...]**. New York: Association for Computing Machinery, Inc., 20 abr. 2020, p. 486 - 497. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85086080593&doi=10.1145%2f3366423.3380132&partnerID=40&md5=b4e082aa696224aaa8850b51e29a5c82>. Acesso em: 18 set. 2023.

SILVA, A. R. da; BAPTISTA, D. M. Organização do conhecimento enquanto área de pesquisa na Ciência da Informação brasileira. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16, 2015, João Pessoa. **Anais [...]**. ANCIB, 2015. Disponível em: <https://repositorio.unb.br/handle/10482/31613>. Acesso em: 1 maio. 2023.

SMITH, A. **Simple Knowledge Organization System (SKOS)**, 2022. (Nota técnica).

SMITH, S.; DOOLEY, D.; HE, Y. Foundational Development of an Occupation Ontology. *In*: Joint Ontology Workshops, Episode VIII: The Svear Sommar of Ontology, 3, 2022, Jonkoping. **Proceedings[...].** New York: Association for Computing Machinery, Inc., v. 3249, 2022. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141399571&partnerID=40&md5=8d5291634f6238ce04f6850fa783461f>. Acesso em: 18 set. 2023.

SOARES, E. et al. Enabling contextualized ontology modeling with a collaborative multi-view system. *In*: INTERNATIONAL SEMANTIC WEB CONFERENCE ON DEMOS AND INDUSTRY TRACKS: FROM NOVEL IDEAS TO INDUSTRIAL PRACTICE, 19, 2020. **Proceedings [...]**. New York: Association for Computing Machinery, Inc, v. 2721, p. 163 – 167, 2020. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096236099&partnerID=40&md5=db06bcf015569c352f75c6b6a921857f>. Acesso em 18 set. 2023.

SOUZA, M. T. de; SILVA, M. D. da; CARVALHO, R. de. Integrative review: what is it? How to do it? São Paulo: **Einstein**, v. 8, n. 1, p. 102–106, mar. 2010.

SOUZA, R. F. de. Organização do conhecimento. *In*: Para entender a ciência da informação. Salvador: **EDUFBA**, 2007.

SOUZA, R. R.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, v. 33, p. 132–141, 2004.

TEIXEIRA, R. da S.; SOUSA, B. P. de. O uso de figuras de linguagens do domínio da lesbiandade no acervo fotográfico do Lesbian Herstory Archives: uma proposta de taxonomia. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 25, p. 01–21, 2 set. 2020.

TERRA, J. C. et al. Taxonomia: elemento fundamental para a gestão do conhecimento. **Biblioteca Terra Fórum Consultores**, 2005.

TIMMARAJU, A. S.; LIU, A.; TRIPATHI, P. Addressing Challenges in Building Web-Scale Content Classification Systems. *In*: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 2020, Barcelona. **Proceedings [...]**. New York: Institute of Electrical and Electronics Engineers Inc., v. 2020-May, p. 8134 – 8138. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089215587&doi=10.1109%2fICASSP40776.2020.9054158&partnerID=40&md5=fbd519ca8fa5102c9fce05d5930e38e6>. Acesso em: 17 set. 2023.

TOLEDANO, C. A. Web 2.0: the origin of the word that has changed the way we understand public relations. *In*: International PPR 2013 Conference, 3, 2013, Barcelona. **Proceedings [...].** Barcelona: jul. 2013.

TORONTO, C. E.; REMINGTON, R. **A Step-by-Step Guide to Conducting an Integrative Review**. Cham: Springer International Publishing, 2020.

TOUTAIN, L. M. B. B. **Para entender a ciência da informação**. Salvador: EDUFBA, 2007.

VIGNOLI, R. G.; SOUTO, D. V. B.; CERVANTES, B. M. N. Sistemas de organização do conhecimento com foco em ontologias e taxonomias. **Informação & Sociedade: Estudos**, v. 23, n. 2, 2013.

W3C. **Semantic Web Activity**. Disponível em: <https://www.w3.org/2001/sw/>. Acesso em: 1 nov. 2022.

W3C. **Semantic Web**. Disponível em: <https://www.w3.org/standards/semanticweb>. Acesso em: 21 set. 2022.

W3C. **Long Description of W3C10 Timeline Graphic**, 2005. Disponível em: <https://www.w3.org/2005/01/timelines/description>. Acesso em: 21 maio. 2023.

WANG, C. et al. A family of fuzzy orthogonal projection models for monolingual and cross-lingual hypernymy prediction. *In*: The World Wide Web Conference, 2019, San Francisco. **Proceedings [...].** New York: Association for Computing Machinery, Inc.,

13 mai. 2019, p. 1965-1976, 2019a. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85066917337&doi=10.1145%2f3308558.3313439&partnerID=40&md5=077d17fd78628437b33d01e2a93e54e7>. Acesso em: 18 set. 2023.

WANG, C.; HE, X.; ZHOU, A. Improving Hypernymy Prediction via Taxonomy Enhanced Adversarial Learning. *In*: 33rd AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, 31st INNOVATIVE APPLICATIONS OF ARTIFICIAL INTELLIGENCE CONFERENCE AND THE 9th AAAI SYMPOSIUM ON EDUCATIONAL ADVANCES IN ARTIFICIAL INTELLIGENCE, 2019, Honolulu. **Proceedings [...]**. Cambridge: AAAI Press, p. 7128 – 7135, 2019. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85094104062&partnerID=40&md5=d4731a808e266f8401db4471010870e3>. Acesso em: 18 set. 2023.

WANG, Q. et al. Set-based Noise Elimination for Is-a Relations in a Large-Scale Lexical Taxonomy. *In*: IEEE INTERNATIONAL CONFERENCE ON POWER DATA SCIENCE, 2019, Taizhou. **Proceedings [...].** New York: Institute of Electrical and Electronics Engineers Inc., p. 101 – 104, 2019b. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85082500574&doi=10.1109%2fICPDS47662.2019.9017169&partnerID=40&md5=fc88e0115a434b7c9d6ea6c329300a3d>. Acesso em 17 set. 2023.

WEN, S.-F.; YAMIN, M. M.; KATT, B. Ontology-Based Scenario Modeling for Cyber Security Exercise. *In*: IEEE EUROPEAN SYMPOSIUM ON SECURITY AND PRIVACY WORKSHOPS, 6, 2021, Vienna. **Proceedings [...]**. New York: Institute of Electrical and Electronics Engineers Inc., p. 249-258. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119058951&doi=10.1109%2fEuroSPW54576.2021.00032&partnerID=40&md5=8f3f07e80eea1d181ac160ef8d6512c7>. Acesso em: 17 set. 2023.

WESER, M. et al. An Ontology-based Metamodel for Capability Descriptions. *In*: IEEE INTERNATIONAL CONFERENCE ON EMERGING TECHNOLOGIES AND FACTORY AUTOMATION, 25, 2020, Vienna. **Proceedings [...]**. New York: Institute of Electrical and Electronics Engineers Inc., v. 2020-September, p. 1679 – 1686. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85093357753&doi=10.1109%2fETFA46521.2020.9212104&partnerID=40&md5=c8b9c3cdac32b5ff6d816130f5527b56>. Acesso em: 18 set. 2023.

YAMAN, B.; THOMPSON, K.; BRENNAN, R. A SKOS taxonomy of the UN global Geospatial information management data themes. *In*: INTERNATIONAL WORKSHOP ON GEOSPATIAL LINKED DATA, 4, 2021. **Proceedings [...].** Aachen: CEUR-WS, v. 2977, p. 89 – 96. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117796821&partnerID=40&md5=e754167f490d2508ab36921564c81089>. Acesso em 17 set. 2023.

YANG, L.; CORMICAN, K.; YU, M. Learning systems engineering domain ontologies from text documents. *In*: IEEE INTERNATIONAL SYMPOSIUM ON SYSTEMS ENGINEERING, 5, 2019, Edinburgh. **Proceedings [...]**. New York: Institute of Electrical and Electronics Engineers Inc., 3 out. 2019. Disponível em:

<https://www.scopus.com/inward/record.uri?eid=2-s2.0-85081091143&doi=10.1109%2fISSE46696.2019.8984550&partnerID=40&md5=c7106fc9f330c10ba92f950bebe70743>. Acesso em 17 set. 2023.

ZAHERA, H. M.; SHERIF, M. A. ProBERT: Product data classification with fine-tuning BERT model. *In*: SEMANTIC WEB CHALLENGE ON MINING THE WEB OF HTML-EMBEDDED PRODUCT DATA, 1, 2020, Athens. **Proceedings [...].** Aachen: CEUR-WS, 2020. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85095976845&partnerID=40&md5=afe921f9679113de8fe4996153968fc7>. Acesso em: 18 set. 2023.

ZAKHAROVA, O. Context *web* services matching in the discovery task resolving. Ontological approaches. *In:* INTERNATIONAL SCIENTIFIC AND PRACTICAL CONFERENCE OF PROGRAMMING, 12, 2020, Kyiv. **Proceedings [...]**. Aachen: CEUR-WS, v. 2866, 2020, p. 39-47. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85106751467&partnerID=40&md5=c62bd21d210f1af25db5c58d2e3d9da0>. Acesso em: 17 set. 2023.

ZENG, M. L. Knowledge Organization Systems (KOS). **Knowledge Organization**, v. 35, n. 2–3, p. 160–182, 2008.

ZHANG, J. et al. Taxonomy Completion via Triplet Matching Network. *In*: AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, 35, 2021. **Proceedings [...].** Cambridge: AAAI Press, v. 5B, 2021, p. 4662-4670, 2021a. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85106115115&partnerID=40&md5=9e57ec43c33748538bbb5be215507463>. Acesso em: 18 set. 2023.

ZHANG, N. et al. AliCG: Fine-grained and Evolvable Conceptual Graph Construction for Semantic Search at Alibaba. *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 35, 2021. **Proceedings [...].** Cambridge: AAAI Press, v. 5B, p. 4662-4670, 2021b. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108387990&doi=10.1145%2f3447548.3467057&partnerID=40&md5=9ad4d6bccce0f774f8c9bb57f2299353>. Acesso em: 18 set. 2023.

ZHANG, P. et al. CrawlPhish: Large-scale analysis of client-side cloaking techniques in phishing. *In*: IEEE SYMPOSIUM ON SECURITY AND PRIVACY, 42, 2021, San Franscisco. **Proceedings [...]**. New York: Institute of Electrical and Electronics Engineers Inc., v. 2021-May, p. 1109 – 1124, 2021c. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115071617&doi=10.1109%2fSP40001.2021.00021&partnerID=40&md5=de0d6454db851906bcb2c61c053c4e51>. Acesso em: 18 set. 2023.

ZHANG, X.; SUN, L. Research on the Similarity between Nodes with Hypernymy/Hyponymy Relations based on IC and Taxonomical Structure. **International Arab Journal of Information Technology**, v. 19, n. 3, p. 388–395, 2022.

ZINS, C. Conceptual approaches for defining data, information, and knowledge. **Journal of the American Society for Information Science and Technology**, v. 58, n. 4, p. 479–493, 15 fev. 2007.

ZOUAQ, A.; MARTEL, F. What is the schema of your knowledge graph?: Leveraging knowledge graph embeddings and clustering for expressive taxonomy learning. *In*: INTERNATIONAL WORKSHOP ON ON SEMANTIC BIG DATA, 2020, Portland. **Proceedings [...]**. New York: Association for Computing Machinery, Inc., 2020. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85087110106&doi=10.1145%2f3391274.3393637&partnerID=40&md5=d49a8504fdd58ac0652e501c4af650da>. Acesso em: 18 set. 2023.

GLOSSÁRIO

API: *Application Programming Interface* (API) ou Interface de Programação de Aplicação. API são mecanismos que permitem que dois componentes de software se comuniquem usando um conjunto de definições e protocolos. Todos os serviços da Web são API.

Aplicações *downstream*: aplicações que utilizam dados de usuários como base para suas operações;

Aprendizagem de taxonomias: do inglês *taxonomy learning*, é repetidamente agrupar hierarquicamente vetores de entidades e criar uma estrutura sobre grupos de entidades semanticamente próximas. Os agrupamentos criados são rotulados com axiomas que explicam uma divisão de agrupamento.

Axioma: premissa ou declaração aceita como verdadeira sem a necessidade de prova dentro de um sistema lógico ou teoria formal.

***Backlog* do produto**: lista que indica requisitos, funcionalidades, melhorias e correções que devem ser desenvolvidos e priorizados no âmbito do desenvolvimento de um produto, geralmente no contexto de aplicação de gerenciamento ágil de projetosBases de conhecimento / Knowledge bases

Conclusão de taxonomia: do inglês *taxonomy completion*, vai além da expansão ao encontrar pares de hiperônimos e hipônimos para novos conceitos, aumentando a precisão no contexto do preenchimento automático de taxonomias.

***Corpus* naturais**: conjuntos de dados textuais em linguagem natural compreensível por máquina.

Corpus: conjuntos de dados, geralmente textuais, também chamados corporas.

Crowdsourcing: técnica empregada com o propósito de reunir diferentes pessoas em torno da realização de uma tarefa ou da solução de um problema.

Embeddings: ou *word embeddings*, são vetores de palavras ou representações numéricas de palavras em um espaço vetorial contínuo. Esses vetores são criados por algoritmos de aprendizagem de máquina, como Word2Vec, glove ou fasttext, que capturam as relações semânticas e contextuais entre palavras com base em grandes quantidades de dados textuais. Variações do termo também podem ser encontradas.

Engenharia de ontologias: do inglês *ontology engeneering;,*engloba todos os processos concernentes à construção, manutenção e expansão de ontologias; aparece muitas vezes ligado à expansão da taxonomia da ontologia.

Expansão de taxonomias: do inglês taxonomy expansion, refere-se a uma taxonomia existente a partir da qual um conjunto de novos conceitos emerge.

***Explainable* AI (XAI)**: novo conceito de Inteligência Artificial Explicável, que propõe um conjunto de medidas que visa combater a falta de transparência dos modelos populares de aprendizagem de máquina.

Extração taxonômica: do inglês t*axonomic extraction*, presume a existência de uma estrutura hierárquica e se refere à extração de esquema ou estrutura de grafo do conhecimento de um conjunto de dados.

Grafo do conhecimento: do inglês Knowledge Graphs (KG), são estruturas de dados que representam o conhecimento por meio de grafos ou triplos na forma “entidade principal-propriedade-entidade final”. Geralmente dependem de um esquema de representação de classes como ponto de partida.

HTML: Hipertext Markup Language. É uma linguagem de marcação de hipertexto que codifica formatações, links e outros recursos em páginas da *web*.

Indução de taxonomias: do inglês *taxonomy induction,* se refere ao processo de inferência de uma estrutura hierárquica ou taxonomia a partir de um conjunto de dados.

Internet das coisas: do inglês *Internet of Things (IoT)*, refere-se a uma rede de dispositivos geralmente utilizados no ambiente doméstico incorporados com internet e tecnologias que possibilitem a agregação de funções relacionadas à conectividade.

Log: é um registro eletrônico que fornece uma auditoria de atividades e modificações num sistema computacional. Pode ser de eventos, de acesso, de processamento, entre outros registros de atividade humana ou de máquina num sistema.

Marcação: do inglês *tagging,* também é chamada etiquetagem e se refere à marcação livre de palavras-chaves a determinado conteúdo; está geralmente associada às mídias sociais e aos *blogs*.

Motifs: padrões recorrentes ou estruturas semelhantes encontradas nos dados que estão sendo analisados para criar uma taxonomia.

Rotulagem: do inglês *labeling*, se refere ao produto da ação de conferir rótulos a conjuntos de dados ou conteúdo com base numa estrutura de classificação, seguindo critérios específicos; são também denominados termos alternativos ou rótulos alternativos no contexto das taxonomias.

Serviços *web*: do inglês *Web Services,* estruturas de comunicação e interoperabilidade entre aplicativos e sistemas distribuídos pela Internet.

String: Sequência de caracteres alfanuméricos (letras, números e/ou símbolos) amplamente utilizadas em programação e no ambiente *web*.

Synset: conjunto de sinônimos.

Taxonomia de tópicos: taxonomia em árvore (hierarquicamente estruturada), onde os nós contém um conjunto de termos semanticamente similares.

Token: dispositivo ou protocolo de identificação segura.

XML: extensible Markup Language, uma linguagem de marcação que permite aos indivíduos definir e utilizar tags próprias com o objetivo facilitar o compartilhamento de informações.

.

APÊNDICE A - Relatório *Checklist* PRISMA

Seção/Tópico	Item #	Item do *Checklist*	Local onde o item é relatado
TÍTULO			-
Título	1	Identificar o relato como uma revisão integrativa.	Título do trabalho e Capítulo 3. Metodologia
RESUMO/ABSTRACT			-
Abstract	2	Consulte a lista de verificação PRISMA 2020 para resumos.	RESUMO
INTRODUÇÃO			-
Justificativa	3	Descreva a justificativa para a revisão no contexto do conhecimento existente.	Item 1.1
Objetivos	4	Forneça uma declaração explícita do(s) objetivo(s) ou questão(ões) abordada(s) pela revisão.	Item 1.2
MÉTODOS			
Critérios de elegibilidade	5	Especifique os critérios de inclusão e exclusão para a revisão e como os estudos foram agrupados para as sínteses.	Item 3.3.1
Fontes de informação	6	Especifique todas as bases de dados, registros, sites, organizações, listas de referência e outras fontes pesquisadas ou consultadas para identificar estudos. Especifique a data em que cada fonte foi pesquisada ou consultada pela última vez.	Item 3.3.2
Estratégia de busca	7	Apresentar as estratégias de busca completas para todos os bancos de dados, registros e sites, incluindo quaisquer filtros e limites usados.	Item 3.3.3
Processo de seleção de estudos	8	Especifique os métodos usados para decidir se um estudo atendeu aos critérios de inclusão da revisão, incluindo quantos revisores examinaram cada registro e cada relatório recuperado, se trabalharam de forma independente e, se aplicável, detalhes das ferramentas de automação usadas no processo.	Item 3.3.4.1
Processo de coleta de dados	9	Especifique os métodos usados para coletar dados dos relatórios, incluindo quantos revisores coletaram dados de cada relatório, se eles trabalharam de forma independente, quaisquer processos para obter ou confirmar dados dos investigadores do estudo e, se aplicável, detalhes das ferramentas de automação usadas no processo.	Item 3.3.4.2
Itens de dados	10a	Liste e defina todos os resultados para os quais os dados foram buscados. Especifique se todos os resultados compatíveis com cada domínio de resultado em cada estudo foram buscados (por exemplo, para todas as medidas, pontos de tempo, análises) e, se não, os métodos usados para decidir quais resultados coletar.	Itens 1.2; 3.1 e 3.3
	10b	Liste e defina todas as outras variáveis para as quais os dados foram solicitados (por exemplo, características dos participantes e da intervenção, fontes de financiamento). Descreva quaisquer suposições feitas sobre qualquer informação ausente ou pouco clara.	Não se aplica
Avaliação do risco de viés do estudo	11	Especifique os métodos usados para avaliar o risco de viés nos estudos incluídos, incluindo detalhes da(s) ferramenta(s) usada(s), quantos revisores avaliaram cada estudo e se eles trabalharam de forma independente e, se aplicável, detalhes das ferramentas de automação usado no processo.	Contemplado no Item 3.3.4.2

Medidas de efeito	12	Especifique para cada resultado a(s) medida(s) de efeito (por exemplo, taxa de risco, diferença média) usada na síntese ou apresentação dos resultados.	Não se aplica, pesquisa qualitativa.
Métodos de síntese	13a	Descreva os processos usados para decidir quais estudos eram elegíveis para cada síntese (por exemplo, tabulação das características da intervenção do estudo e comparação com os grupos planejados para cada síntese (item nº 5)).	3.3.4
	13b	Descreva quaisquer métodos necessários para preparar os dados para apresentação ou síntese, como tratamento de estatísticas resumidas ausentes ou conversões de dados.	3.3.4
	13c	Descreva quaisquer métodos usados para tabular ou exibir visualmente os resultados de estudos e sínteses individuais.	3.3.4
	13d	Descreva quaisquer métodos usados para sintetizar resultados e justifique a(s) escolha(s). Se a meta-análise foi realizada, descreva o(s) modelo(s), método(s) para identificar a presença e a extensão da heterogeneidade estatística e o(s) pacote(s) de software usado(s).	Não se aplica, pesquisa qualitativa.
	13e	Descreva quaisquer métodos usados para explorar possíveis causas de heterogeneidade entre os resultados do estudo (por exemplo, análise de subgrupo, meta-regressão).	
	13f	Descreva quaisquer análises de sensibilidade realizadas para avaliar a robustez dos resultados sintetizados.	
Avaliação de viés de relato	14	Descreva quaisquer métodos usados para avaliar o risco de viés devido à falta de resultados em uma síntese (decorrente de vieses de relato).	3.3.4
Avaliação de confiança	15	Descreva quaisquer métodos usados para avaliar a certeza (ou confiança) no conjunto de evidências de um resultado.	3.3.4
RESULTADOS			
Seleção de estudos	16a	Descreva os resultados do processo de busca e seleção, desde o número de registros identificados na busca até o número de estudos incluídos na revisão, de preferência usando um diagrama de fluxo.	3.3
	16b	Cite estudos que parecem atender aos critérios de inclusão, mas que foram excluídos, e explique por que foram excluídos.	APÊNDICE B
Características do estudo	17	Cite cada estudo incluído e apresente suas características.	APÊNDICE B
Risco de viés nos estudos	18	Apresentar avaliações de risco de viés para cada estudo incluído.	Não se aplica na RI, abordado em 3.3.4 para o conjunto.
Resultados de estudos individuais	19	Para todos os resultados, apresente, para cada estudo: (a) estatísticas resumidas para cada grupo (quando apropriado) e (b) uma estimativa de efeito e sua precisão (por exemplo, intervalo de confiança/crível), idealmente usando tabelas estruturadas ou parcelas.	APÊNDICE D
Resultados das sínteses	20a	Para cada síntese, resuma brevemente as características e o risco de viés entre os estudos contribuintes.	Não se aplica na RI, abordado em 3.3.4 para o conjunto.
	20b	Apresentar os resultados de todas as sínteses estatísticas realizadas. Se a metanálise foi feita, apresente para cada estimativa resumida e sua precisão (por exemplo, intervalo de confiança/crível) e medidas de heterogeneidade estatística. Se comparar grupos, descreva a direção do efeito.	
	20c	Apresentar resultados de todas as investigações de possíveis causas de heterogeneidade entre os resultados do estudo.	

	20d	Apresentar os resultados de todas as análises de sensibilidade realizadas para avaliar a robustez dos resultados sintetizados.	
Vieses de relato	21	Apresentar avaliações de risco de viés devido a resultados ausentes (decorrentes de vieses de relato) para cada síntese avaliada.	Não se aplica na RI, contemplado no Item 3.3.4.2
Certeza da evidência	22	Apresentar avaliações de certeza (ou confiança) no conjunto de evidências para cada resultado avaliado.	Não se aplica, pesquisa qualitativa.
DISCUSSÃO			-
Discussão	23a	Forneça uma interpretação geral dos resultados no contexto de outras evidências.	Item 5
	23b	Discuta quaisquer limitações das evidências incluídas na revisão.	
	23c	Discuta quaisquer limitações dos processos de revisão usados.	
	23d	Discuta as implicações dos resultados para a prática, política e pesquisas futuras.	
OUTRAS INFORMAÇÕES			
Registro e protocolo	24a	Fornecer informações cadastrais da resenha, incluindo nome do registro e número do registro, ou informar que a resenha não foi registrada.	Não se aplica na RI. Item 3
	24b	Indique onde o protocolo de revisão pode ser acessado ou declare que um protocolo não foi preparado.	
	24c	Descrever e explicar quaisquer alterações nas informações fornecidas no registro ou no protocolo.	
Apoio	25	Descreva as fontes de apoio financeiro ou não financeiro para a revisão e o papel dos financiadores ou patrocinadores na revisão.	Não se aplica
Conflito de interesses	26	Declare quaisquer interesses conflitantes dos autores de resenhas.	Não se aplica
Gestão de Dados	27	Relate quais dos itens a seguir estão disponíveis ao público e onde podem ser encontrados: modelos de formulários de coleta de dados; dados extraídos dos estudos incluídos; dados usados para todas as análises; código analítico; quaisquer outros materiais usados na revisão.	Biblioteca Digital Brasileira de Teses e Dissertações (BDTD)

Fonte: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

APÊNDICE B - Estudos avaliados para elegibilidade e qualidade metodológica

Elegibilidade	Estudo	Título	Tipo de pesquisa	Avaliação da qualidade	Justificativa
Sim	Alonso et al., 2020	*Answering recreational web searches with relevant things to do results*	Experimental	Razoável	-
Excluído	Alrumaih et al., 2020	*Domain Ontology for Requirements Classification in Requirements Engineering Context*	Experimental		Aborda somente ontologia, sem que aspectos das taxonomias sejam discutidos.
Sim	Arous et al., 2023	*TaxoComplete: Self-Supervised Taxonomy Completion Leveraging Position-Enhanced Semantic Matching*	Experimental	Muito boa	-
Excluído	Bhutani e Saha, 2019	*Towards an evolved information food chain of world wide web and taxonomy of semantic web mining*	Estudo de revisão		Estudo de revisão
Excluído	Bhutani et al., 2021	*Semantic framework for facilitating product discovery*	Experimental		Propõe uma nova estrutura de mecanismo de busca semântica e discute a implementação de sua própria ontologia de classificação de produtos sem que seja abordada a estrutura hierárquica ou citado o termo taxonomia.
Sim	Bongiovanni et al., 2023	*Zero-Shot Taxonomy Mapping for Document Classification*	Experimental	Boa	-
Excluído	Bonte et al., 2019	*C-Sprite: Efficient hierarchical reasoning for rapid RDF stream processing*	Experimental		Estudo excluído por não abordar taxonomias ou hierarquias de maneira mais direta.
Sim	Caucheteux et al., 2021	*Disentangling Syntax and Semantics in the Brain with Deep Networks*	Experimental	Boa	Este foi um estudo particularmente difícil de sintetizar e a decisão de mantê-lo passou pela importância da evidência de uso de taxonomias como auxiliar

Elegibilidade	Estudo	Título	Tipo de pesquisa	Avaliação da qualidade	Justificativa
					em estudos de redes neurais profundas.
Sim	Crotti Junior et al., 2020	*The ARK platform: Enabling risk management through semantic web technologies*	Experimental	Razoável	-
Sim	Danilevitch, 2019	*Logical semantics approach for data modeling in XBRL taxonomies*	Experimental	Boa	-
Sim	Daraio et al., 2020	*An explainable data-driven approach to web directory taxonomy mapping*	Experimental	Boa	-
Sim	Dilorenzo et al., 2020	*Enabling the Reuse of Software Development Assets through a Taxonomy for User Stories*	Experimental	Excelente	-
Sim	Doan et al., 2020	*A semantic framework for extracting taxonomic relations from text corpus*	Experimental	Muito boa	-
Sim	Fauceglia et al., 2019	*Automatic taxonomy induction and expansion*	Experimental	Razoável	-
Sim	Figueroa et al., 2019	*Integrating heterogeneous sources for predicting question temporal anchors across Yahoo! Answers*	Experimental	Muito boa	-
Sim	Giabelli et al., 2020	*NEO: A Tool for Taxonomy Enrichment with New Emerging Occupations*	Experimental	Muito boa	-
Sim	Golpayegani et al., 2023	*To Be High-Risk, or Not To Be - Semantic Specifications and Implications of the AI Act's High-Risk AI Applications and Harmonised Standards*	Experimental	Excelente	-
Excluído	Guizzardi et al., 2021	*Types and taxonomic structures in conceptual modeling: A novel ontological theory and engineering support*	Estudo de revisão		Estudo teórico; de base bibliográfica

Elegibilidade	Estudo	Título	Tipo de pesquisa	Avaliação da qualidade	Justificativa
Sim	Horta et al., 2023	*Semantic Interpretability of Convolutional Neural Networks by Taxonomy Extraction*	Experimental	Muito boa	-
Excluído	Hossain et al., 2021	*A survey on automatically constructed universal knowledge bases*	Estudo de revisão		-
Sim	Huang et al., 2020	*CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring*	Experimental	Muito boa	-
Sim	Huang et al., 2021	*Conflict Detection in IoT-based Smart Homes*	Experimental	Razoável	Estudo incluído visto que a Internet das Coisas pode se beneficiar da Web Semântica ao utilizar seus padrões e tecnologias para facilitar a interoperabilidade, a integração e a interpretação dos dados gerados pelos dispositivos IoT.
Sim	Jain, 2023	*Product discovery utilizing the semantic data model*	Experimental	Excelente	-
Sim	Ji et al., 2019	*Microsoft concept graph: Mining semantic concepts for short text understanding*	Experimental	Boa	-
Incluído	Jiang et al., 2022	*TaxoEnrich: Self-Supervised Taxonomy Completion via Structure-Semantic Representations*	Experimental	Muito Boa	Identificado nas referências de Arous et al., 2023
Excluído	Jindal e Kaur, 2020	*Information Retrieval from Software Bug Ontology Exploiting Formal Concept Analysis*	Experimental		Não aborda taxonomias, somente análise formal de conceitos e ontologias
Sim	Kasenchak, 2019	*What is Semantic Search? and why is it important?*	Experimental	Razoável	-

Elegibilidade	Estudo	Título	Tipo de pesquisa	Avaliação da qualidade	Justificativa
Excluído	Li et al., 2021	*Semiautomatic Generation of Code Ontology Using ifcOWL in Compliance Checking*	Experimental		Apesar de a pesquisa abordar os relacionamentos hierárquicos, não aborda taxonomias, cita timidamente em dois momentos, mas se concentra em relacionamentos mais profundos no âmbito ontológico como forma de reduzir a necessidade de intervenção humana.
Sim	Li et al., 2022	*Combining deep learning and ontology reasoning for remote sensing image semantic segmentation*	Experimental	Excelente	-
Sim	Lourdusamy e John, 2019	*Metric based ontology quality evaluation*	Experimental	Razoável	Avalia especificamente a camada taxonômica da ontologia para entender as relações e a organização dos conceitos
Sim	Malandri et al., 2020	*MEET: A Method for Embeddings Evaluation for Taxonomic Data*	Experimental	Muito boa	-
Sim	Malandri et al., 2021	*MEET-LM: A method for embeddings evaluation for taxonomic data in the labour market*	Experimental	Muito boa	-
Sim	Maldonado et al., 2019	*Size matters: The impact of training size in taxonomically-enriched word embeddings*	Experimental	Boa	-
Sim	Manzoor et al., 2020	*Expanding Taxonomies with Implicit Edge Semantics*	Experimental	Muito boa	-
Sim	Martel e Zouaq, 2021	*Taxonomy extraction using knowledge graph embeddings and hierarchical clustering*	Experimental	Boa	-
Sim	Musgrove e Walsh, 2019	*Remediating intentionally corrupted ontology-based data sets*	Experimental	Razoável	-
Sim	Patel et al., 2019	*Data of semantic web as unit of knowledge*	Teórica	Razoável	-

Elegibilidade	Estudo	Título	Tipo de pesquisa	Avaliação da qualidade	Justificativa
Sim	Poirier, 2019	*Classification as catachresis: Double binds of representing difference with semiotic infrastructure*	Teórica	Boa	-
Sim	Prince et al., 2021	*DUI: The drug use insights web server*	Experimental	Muito boa	-
Sim	Raza et al., 2019	*A Taxonomy and Survey of Semantic Approaches for Query Expansion*	Experimental	Razoável	-
Excluído	Rovetto et al., 2020	*Orbital debris ontology, terminology, and knowledge modeling*	Estudo de revisão		Estudo de revisão
Sim	Salatino et al., 2020	*The computer science ontology: A comprehensive automatically-generated taxonomy of research areas*	Experimental	Boa	-
Sim	Sha et al., 2023	*Link Topics from Q&A Platforms using Wikidata: A Tool for Cross-platform Hierarchical Classification*	Experimental	Boa	O artigo foi incluído por abordar fortemente relações taxonômicas e valorizar a estrutura do domínio própria das taxonomias, mesmo que no contexto das ontologias.
Sim	Shang et al., 2020	*NetTaxo: Automated Topic Taxonomy Construction from Text-Rich Network*	Experimental	Boa	-
Sim	Sharifpour et al., 2023	*Large-scale analysis of query logs to profile users for dataset search*	Experimental	Boa	-
Sim	Shen et al., 2020	*TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network*	Experimental	Excelente	-

Elegibilidade	Estudo	Título	Tipo de pesquisa	Avaliação da qualidade	Justificativa
Excluído	Shukla et al., 2023	*Disaster management ontology- an ontological approach to disaster management automation*	Experimental		O estudo não aborda taxonomias, há a ocorrência de algo relacionado a taxonomia apenas 3 vezes.
Excluído	Silva, 2020	A organização da informação na interface de *website* a partir de arranjos facetados	Estudo de revisão		Estudo de revisão
Excluído	Silva e Miranda, 2020	*Panoramic investigation on the adoption of faceted structures within the view of information architecture*	Estudo de revisão		Estudo de revisão
Sim	Smith et al., 2022	*Foundational Development of an Occupation Ontology*	Experimental	Excelente	-
Sim	Soares et al., 2020	*Enabling contextualized ontology modeling with a collaborative multi-view system*	Experimental	Razoável	-
Excluído	Stephen et al., 2023	*A Pattern for Representing Scientific Taxonomies*	Experimental		Descreve o desenvolvimento de uma ontologia a partir do ponto em que já existe uma taxonomia bem estruturada. Foco na modelagem ontológica.
Sim	Teixeira e Sousa et al., 2020	*The use of figures of speech in the lesbianity domain in lesbian herstory archives's photographic collection: A taxonomy proposal*	Experimental	Boa	-
Sim	Timmaraju et al., 2020	*Addressing Challenges in Building Web-Scale Content Classification Systems*	Experimental	Razoável	-
Sim	Wang et al., 2019b	*Set-based Noise Elimination for Is-a Relations in a Large-Scale Lexical Taxonomy*	Experimental	Boa	-

Elegibilidade	Estudo	Título	Tipo de pesquisa	Avaliação da qualidade	Justificativa
Sim	Wang et al., 2019	*Improving Hypernymy Prediction via Taxonomy Enhanced Adversarial Learning*	Experimental	Muito boa	-
Sim	Wang et al., 2019a	*A family of fuzzy orthogonal projection models for monolingual and cross-lingual hypernymy prediction*	Experimental	Muito boa	-
Sim	Wen et al., 2021	*Ontology-Based Scenario Modeling for Cyber Security Exercise*	Experimental	Boa	-
Sim	Weser et al., 2020	*An Ontology-based Metamodel for Capability Descriptions*	Experimental	Boa	-
Sim	Yaman et al., 2021	*A SKOS taxonomy of the UN global Geospatial information management data themes*	Experimental	Razoável	-
Sim	Yang et al., 2019	*Learning systems engineering domain ontologies from text documents*	Experimental	Muito boa	Apesar de não abordar especificamente o tema das taxonomias, considerou-se importante sinalizar que as taxonomias são tidas como igualmente importantes na análise de um domínio, mesmo para profissionais da ontologia.
Sim	Zahera e Sherif, 2020	*ProBERT: Product data classification with fine-tuning BERT model*	Experimental	Regular	-
Sim	Zakharova, 2020	*Context web services matching in the discovery task resolving. Ontological approaches*	Experimental	Regular	-
Sim	Zhang e Sun, 2022	*Research on the Similarity between Nodes with Hypernymy/Hyponymy Relations based on IC and Taxonomical Structure*	Experimental	Muito boa	-

Elegibilidade	Estudo	Título	Tipo de pesquisa	Avaliação da qualidade	Justificativa
Sim	Zhang et al., 2021b	*AliCG: Fine-grained and Evolvable Conceptual Graph Construction for Semantic Search at Alibaba*	Experimental	Muito boa	-
Sim	Zhang et al., 2021c	*CrawlPhish: Large-scale analysis of client-side cloaking techniques in phishing*	Experimental	Muito boa	-
Sim	Zhang et al., 2021a	*Taxonomy Completion via Triplet Matching Network*	Experimental	Excelente	-
Sim	Zouaq e Martel, 2020	*What is the schema of your knowledge graph?: Leveraging knowledge graph embeddings and clustering for expressive taxonomy learning*	Experimental	Boa	-

Fonte: elaborado pelo autor (2023)

APÊNDICE C - Contextualização conceitual dos estudos incluídos na revisão integrativa

Estudo	Área	Palavras-chave	Conceito na taxonomia da RI	Conceitos adicionais citados	Teorias citadas	Modelos analisados	Modelos propostos
Alonso et al., 2020	CC	*budget control; complex networks; digital storage; search engines; semantics; websites*	Mecanismo de busca		-	*Location-based social networks (LBSN)*	-
Arous et al., 2023	Engenharia	*taxonomy completion; self-supervision; semantic matching;*	Conclusão de taxonomia	*Conclusão de taxonomia*	-	Arborist; TaxoExpan; TMN; TaxoEnrich	TaxoComplete
Bongiovanni et al., 2023	CC	*hierarchical text classification; natural language processing; zero-shot text classification*	PLN	*hierarchical text classification; Deep Language Modeling (DLM)*	-	Web Of Science (WoS), DBpedia Extract e Amazon Product Reviews; Upwards Score Propagation (USP)	Zero-Shot Text Classification
Caucheteux et al., 2021	Linguística	*brain functional; neuro imaging; magnetic resonance imaging; neurophysiology; semantic web; semantics; syntactics*	Aprendizagem profunda	-	-	-	-
Crotti Junior et al., 2020	CC	*ark platform; organizational change; risk management*	LD	-	Teoria dos Sistemas Socio-técnicos	-	Acess Risk Knowledge (ARK)

Estudo	Área	Palavras-chave	Conceito na taxonomia da RI	Conceitos adicionais citados	Teorias citadas	Modelos analisados	Modelos propostos
Danilevitch, 2019	Economia e CC	*logical semantics; natural language processing; formal language; structural semantics; semantic data model; XBRL taxonomy; analytic and synthetic languages*	PLN	*logical semantics*	-	*Teconologia XBRL; Taxonomia XBRL*	-
Daraio et al., 2020	CC	*automated taxonomy mapping; web directories; text categorization; vector representations of text*	Engenharia de ontologias	*IA Explicável ou EXplainable Artificial Intelligence (XAI)*	-	-	-
Dilorenzo et al., 2020	Engenharia Elétrica e Informática	*agile software development; software reuse; user stories; information retrieval; technology acceptance model*	LD	*Desenvolvimento Ágil de Software (ASD); Engenharia de Software Orientada à Reutilização (RDSE); backlog do produto; Scrum; framework ágil; Web Information System (WIS); Technology Acceptance Model (TAM)*	-	-	-
Doan et al., 2020	CC	*taxonomic relation; ontology construction; word sense disambiguation; knowledge aquisition*	Extração taxonômica	*Word Sense Disambiguation (WSD)*	-	WSD e abordagens *web*	-

Estudo	Área	Palavras-chave	Conceito na taxonomia da RI	Conceitos adicionais citados	Teorias citadas	Modelos analisados	Modelos propostos
Fauceglia et al., 2019	Linguística computacio-nal	-	KG	*similaridade semântica (KB e vetores de palavras); supervised learning al-gorithm*	Teoria da Ordem	Knowledge Graph Induction Service (KGIS); KG induction job; Word2vec; Wikidata	Smart Spreadsheet (SSS); Strict Partial Order Networks (SPON)
Figueroa et al., 2019	Engenharia	*multi-view learning; transfer learning; question classification; natural language processing; intelligent information retrieval; web mining*	PLN	*question temporal answers; crowdsourcing*	-	Sequential Forward Floating Search (SFFS)	-
Giabelli et al., 2020	Estatística; Informática; CC	-	Aprendizagem profunda	*Hierarchical Semantic Relatedness (HSR)*	-	Word2Vec, GloVe, FastText	NEO
Golpayegani et al., 2023	CC	*AI act; harmonised standards; high-risk AI; semantic web; taxonomy*	LOD	*FAIR; SKOS; FAIRness*	-	-	-
Horta et al., 2023	Ciência de Dados	*explainable AI; taxonomy extraction; convolutional neural networks*	Aprendizagem profunda	*Rede Neural Convolucional (ConvNet / Convolutional Neural Network / CNN); XAI; verdade fundamental*	-	-	-
Huang et al., 2020	CC	*relation extraction; semantic computing; taxonomy construction; topic discovery*	Expansão de taxonomias	*Clustering and classifica; Computing methodologies; Information extraction; Ontology engineering*	-	-	CoRel

Estudo	Área	Palavras-chave	Conceito na taxonomia da RI	Conceitos adicionais citados	Teorias citadas	Modelos analisados	Modelos propostos
Huang et al., 2021	CC	*conflict detection; internet of things; IoT service; knowledge graph; smart home*	IoT	*IoT*	-	-	-
Jain, 2023	CC	*knowledge graph; ontology; engineering equipment; multimedia data; product categorization; product matching; recommendation*	ontologia	-	-	-	-
Ji et al., 2019	CC	*knowledge extraction; conceptualization; text understanding*	KG	*relações is-a, denominadas no estudo como fatos isA;*	Basic-level Categorization (BLC)	WordNet; DBPedia; YAGO; Freebase; ConceptNet; NELL; WikiTaxonomy; KnowItAll	Microsoft Concept Graph e Probase
Jiang et al., 2022	CC	*taxonomy completion; self-supervised learning; knowledge representation*	Conclusão de taxonomia	*Conclusão de taxonomia*	-	Arborist; TaxoExpan; STEAM; TMN; Gentaxo	TaxoEnrich
Kasenchak, 2019	CI	*semantic search, natural language processing, knowledge graphs, taxonomy, text analytics*	Mecanismo de busca	*semantic search; variantes lexicais; fuzzy matching; pesquisa contextual*	-	Google Knowledge Graph; Analyzer	-
Li et al., 2022	Engenharia da Informação	*collaboratively boosting framework (CBF); deep learning; deep semantic segmentation network (DSSN); ontology reasoning; remote sensing (RS) imagery*	Aprendizagem profunda		-	*RS;DSSN*	CBF

Estudo	Área	Palavras-chave	Conceito na taxonomia da RI	Conceitos adicionais citados	Teorias citadas	Modelos analisados	Modelos propostos
Lourdusamy e John, 2019	CC	*ontology evaluation; ontology quality; semantic measures*	ontologia	-	-	-	Metric based Ontology Quality Analyzer (MOQA)
Malandri et al., 2020	Estatística; Informática; CC	*embeddings evaluation; taxonomies; semantic similarity*	Vetores	*Hierarchical Semantic Similarity (HSS)*	-	WordNet; Wikipedia	Method for Embeddings Evaluation for Taxonomic Data (MEET) e HSS
Malandri et al., 2021	Estatística; Informática; CC	*embeddings evaluation; taxonomies; semantic hierarchies; labour market; ICT*	Vetores	-	-	-	MEET-LM
Maldonado et al., 2019	CC	*word embeddings; taxonomic embeddings; wordnet; semantic similarity; taxonomic enrichment; retrofitting*	Vetores	similaridade taxonômica*; random walk; retrofitting.*	-	Skip-Gram; SemEval-17	-
Manzoor et al., 2020	Content *web*; Social media	*learning systems; semantics; world wide web*	Aprendizagem de máquina	semântica de borda	-	-	Arborist
Martel e Zouaq, 2021	CC	*knowledge graph embeddings; expressive taxonomies; ontology*	Aprendizagem de taxonomias	*Clustering; Computing methodologies; Ontology engineering.*	-	-	-

Estudo	Área	Palavras-chave	Conceito na taxonomia da RI	Conceitos adicionais citados	Teorias citadas	Modelos analisados	Modelos propostos
Musgrove e Walsh, 2019	Comunicação	*bad actors; book publishing; cosine similarity; data corruption; knowledge discovery; middle-level ontology; ontological distance; ontology mapping; ontology revision; semantic similarity*	ontologia	-	-	-	-
Patel et al., 2019	Engenharia	*knowledge representation scheme; ontology; semantic web; unit of knowledge*	ontologia		-	OWL, RDF/XML	-
Poirier, 2019	Comunicação	*ethnography; metadata; science and techonology studies; knowledge representation; critical data studies; semantic web*	Classificação	infraestruturas semióticas; catacrese	filosofia analítica	-	-
Prince et al., 2021	Bioinformática	*computer; drug dependence; human; semantics; social media; software*	Aprendizagem de máquina	-	-	Skip-Gram; CBOW	Drug-Use Insights (DUI)
Raza et al., 2019	CC	*information retrieval; morphological expansion; ontology; semantic query expansion*	Mecanismo de busca	*Query expansion (QE);*	-	Abordagem semântica de QE dos tipos linguística, baseada em ontologia, híbrida e tradicional.	-

Estudo	Área	Palavras-chave	Conceito na taxonomia da RI	Conceitos adicionais citados	Teorias citadas	Modelos analisados	Modelos propostos
Salatino et al., 2020	CC	*scholarly data; ontology learning; bibliographic data; scholarly ontologies; semantic web*	Taxonomia da ontologia	-	-	-	CSO e Klink-2
Sha et al., 2023	CC	*topic classification; wikidata ontology; entity linking; q&a platforms*	LD	*Enterprise ontologies; ruled-based topic classification; entity linking*	-	Zhihu, Quora, Wikidata	-
Shang et al., 2020	CC	*embeddings; semantics; world wide web*	Taxonomia hierárquica	*Motifs;* taxonomia de tópicos; *downstream applications; embedding space; neural network classifier;*	-	-	NetTaxo plataforma top-down recursiva
Sharifpour et al., 2023	CC	*dataset search; log analysis; search behaviour; clustering; semantic text similarity; word embedding*	Similaridade semântica	-	-	-	-
Shen et al., 2020	CC	*taxonomy expansion; self-supervised learning*	Expansão de taxonomias	-	-	Field-of-Study (FoS); Microsoft Academic Graph (MAG)	TaxoExpan

Estudo	Área	Palavras-chave	Conceito na taxonomia da RI	Conceitos adicionais citados	Teorias citadas	Modelos analisados	Modelos propostos
Smith et al., 2022	Medicina e Estatística	*occupation;* ESCO; ISCO; US BLS O*Net/SOC; UK National Statistics; ontology; OBO Foundry; CEUR-WS; ENVO; ISIC; ILO	ontologia	-	-	ESCO; ISCO; US BLS O*Net/SOC; UK National Statistics; ontology; OBO Foundry; CEUR-WS; ENVO; ISIC; ILO	Occupational Classification compliant ontology (OccO)
Soares et al., 2020	CC	*hyperknowledge; hybrid knoeledge representation; knowledge management; knowledge visualization; knoeledge curation*	ontologia	*conteúdo simbólico; conteúdo não simbólico*	-	-	Knowledge Explorer System (KES); Hyperknowledge (HK)
Teixeira e Sousa et al., 2020	Biblioteconomia e CI	*figures of speech; lesbian herstory archives; lesbianity; taxonomy; thematic representation*	Taxonomia facetada	*representação temática*	-	Lesbian Herstory Archives (LHA)	-
Timmaraju et al., 2020	Comunicação	*self-attention; weakly supervised pre-training; multi-lingual classification; cross-document transfer; multi-modality*	Aprendizagem profunda	*Transfer Learning*	-	-	BERT
Wang et al., 2019b	Ciência de Dados	*knowledge engineering; knowledge representations; emantic web*	KB	*similaridade semântica;* taxonomias lexicais	-	SimRank; Probase	-

Estudo	Área	Palavras-chave	Conceito na taxonomia da RI	Conceitos adicionais citados	Teorias citadas	Modelos analisados	Modelos propostos
Wang et al., 2019	CC	*artificial intelligence; forecasting; natural language processing systems; semantics*	Aprendizagem de máquina	*NLP tasks; Web scale taxonomies; hypernymy prediction; downstream NLP applications; text entailment; adversarial learning*	-	Microsoft Concept Graph	TEAL, U-TEAL, S-TEAL e AS-TEAL
Wang et al., 2019a	CC	*hypernymy prediction; multi-wahba projection; cross-lingual transfer learning;*	Aprendizagem de máquina	*Fuzzy Orthogonal Projection Models*	Wahba	-	Projeção Multi-Wahba (MWP); Projeção Multi-Wahba de Transferência (TMWP); Iterative Transfer Multi-Wahba Projection (ITMWP)
Wen et al., 2021	CC	*cyber security exercise; scenario modeling; ontology*	Taxonomia da ontologia	-	-	-	-
Weser et al., 2020	Automação industrial; Engenharia elétrica e eletrônica	*capabilities; skills; ontology; semantic web; manufacturing*	ontologia	*ontology-based capability matching for production resources*	-	-	Capability for industry (C41)
Yaman et al., 2021	Ciência de Dados	*data handling; information management; linked data; metadata*	LD	*data catalog*	-	DCAT, RDF e SKOS; *ordnance Survey Ireland (OSi)*	Taxonomia UN-CGIM
Yang et al., 2019	Engenharia; Informática	*ontology learning; natural language processing; systems engineering*	ontologia	*Engenharia de domínio*	-	-	-

Estudo	Área	Palavras-chave	Conceito na taxonomia da RI	Conceitos adicionais citados	Teorias citadas	Modelos analisados	Modelos propostos
Zahera e Sherif, 2020	Ciência de Dados	*classification (of information); HTML; websites*	Mineração de dados	-	-	-	ProBERT
Zakharova, 2020	CC	*web service; discovery task; matching process; matching types; matching degree defining methods; semantical matching; contextual matching; ontological approaches; top level service ontology; web services categorization and classification; a taxonomy of general textual description of web services; matching of web services by context.*	Serviços *web*	*description Logic (DL)*	-	ontologia DL	-
Zhang e Sun, 2022	Engenharia e Informática	*computing similarity; information content; ontology and knowledge graph; wordnet; hypernymy/hyponymy*	Similaridade semântica	*similaridade semântica*	-	WordNet	-
Zhang et al., 2021b	Engenharia do Conhecimento	*information systems; query representation; information extraction*	KG	*Conceptual graphs*	-	modelo BERT	AliCG
Zhang et al., 2021c	CC	*cloaking; evasion; javascript; phishing; web-security*	Taxonomias manuais	-	-	-	CrawPish
Zhang et al., 2021a	CC; Engenharia	*artificial intelligence; large dataset; semantics*	Manutenção de taxonomias	*completion taxonomy*	-	-	Triplet Matching Network (TMN)

Estudo	Área	Palavras-chave	Conceito na taxonomia da RI	Conceitos adicionais citados	Teorias citadas	Modelos analisados	Modelos propostos
Zouaq e Martel, 2020	CC	*knowledge graph embeddings; expressive taxonomies; ontology learning; hierarchical clustering*	KG	*expressive taxonomies; ontology learning; hierarchical clustering; taxonomy induction; taxonomy extraction*	-	-	-

Fonte: elaborado pelo autor (2023)

APÊNDICE D - Resultados e aplicabilidade dos estudos incluídos na revisão integrativa

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
Alonso et al., 2020	Experimental	Razoável	Desenvolveu uma estrutura combinando registros de consultas de pesquisa com dados do LBSN para abordar consultas recreativas. Demonstrou uma melhor classificação dos POIs em relação a um simples classificador baseado em check-in. Criou uma taxonomia a partir de conteúdo gerado pelo usuário e registros comportamentais para modelar a experiência do usuário. Realizou uma avaliação offline com *crowdsourcing* interno, validando a abordagem.	Aplicações *web* em geral que possuam alguma funcionalidade de recomendação por localização.
Arous et al., 2023	Experimental	Muito boa	Os resultados demonstraram que o TaxoComplete supera os métodos de última geração ao prever com precisão a posição de um nó de consulta na taxonomia. Especificamente, descobriu-se que o TaxoComplete é 2 vezes mais eficaz em termos de HIT @k, que é uma métrica de avaliação comumente usada para tarefas de conclusão de taxonomia, isto é, adição de novos conceitos em qualquer nível dentro de uma taxonomia existente para manter sua relevância e precisão.	*E-commerce* e plataformas baseadas em conhecimento
Bongiovanni et al., 2023	Experimental	Boa	Os resultados demonstraram que a Classificação de Texto Semântico Zero-Shot foi capaz de produzir pontuações de relevância de última geração para rótulos por meio da taxonomia e apresentou melhora considerável após a implementação de uma técnica denominada Upwards Score Propagation (USP), que aproveita a estrutura hierárquica da taxonomia para reforçar a abordagem conduzida nos experimentos nos três conjuntos.	Datasets; Mídias sociais e folksonomias
Caucheteux et al., 2021	Experimental	Boa	Os resultados dos experimentos empreendidos sugerem que as representações composicionais recrutam uma rede cortical mais difundida do que as lexicais e que a sintaxe e a semântica compartilham um substrato neural comum.	Linguística e Linguística computacional
Crotti Junior et al., 2020	Experimental	Razoável	A plataforma ARK foi considerada útil e de fácil aprendizado e fornece uma visão abrante dos casos clínicos. Funcionalidades como interligação de fontes de dados como evidências, definições da taxonomia disponíveis na interface e possibilidade de ocultar funções se destacaram.	Psicologia organizacional; gerenciamento de riscos para sistemas em

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
				diferentes domínios, tais como saúde, aviação e finanças.
Danilevitch, 2019	Experimental	Boa	Linguagens naturais sintéticas são mais desafiadoras para o design de taxonomia. Para domínios de conhecimento complexos diretamente ligados a pronunciamentos, regras ou leis regulatórias, a classificação deve ser projetada por pessoas que entendam a natureza técnica dessas regras de negócios. A modelagem de dados em taxonomias XBRL representa um exemplo de abordagem baseada em humanos para classificação de informações legíveis por humanos em um formato legível por máquina. Acreditamos que a pesquisa proposta avançaria no campo da Web Semântica em relação aos meios de transformação de dados não estruturados em relatórios estruturados.	Linguística computacional
Daraio et al., 2020	Experimental	Boa	A representação de texto baseada em frequência mostrou-se particularmente adequada para conduzir o procedimento automatizado de mapeamento de taxonomia, enquanto o espaço de incorporação tem sido usado de forma lucrativa para fornecer explicações locais das atribuições de categoria.	e-commerce; Internet of Things
Dilorenzo et al., 2020	Experimental	Excelente	A taxonomia permite a reutilização de histórias de usuários e seus ativos relacionados. No estudo a taxonomia desempenhou com 100% de cobertura e granularidade e 89,8% de heterogeneidade. Os experimentos demonstraram 90,17% de cobertura, alta granularidade e 80,11% de heterogeneidade. A taxonomia foi prontamente adotada pelas equipes de desenvolvimento de software, que classificaram como útil e de fácil manuseio.	Gerenciamento de projetos; suporte à decisão; dataset com histórias de usuários previamente classificadas;
Doan et al., 2020	Experimental	Muito boa	Os reusltados demonstram que o modelo proposto é capaz de extrair eficientemente as relações taxonômicas de corporas de texto, mesmo quando os dados são insuficientes ou carecem de treinamento. Na avaliação do modelo foi demonstrada alta precisão e recordação na extração de termos-chave dos conjuntos de dados, superando um método de	Linguística computacional

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
			última geração em conjuntos de dados com altos resultados nas técnicas de desambiguação.	
Fauceglia et al., 2019	Experimental	Razoável	O KGIS induz taxonomias não-hierárquicas com base em 3 abordagens: uma baseada na identificação de padrões, para extrair relacionamentos tipo-instância (is-a); outra baseada na similaridade semântica reconhecida pelo cruzamento com entidades da base de conhecimento Wikidata; e outra por meio de uma arquitetura de rede neural desenvolvida com base na teoria da ordem para o estabelecimento de relações is-a estritas de ordem parcial, isto é, relações transitivas, assimétricas e irreflexivas, mantendo a consistência das taxonomias geradas e evitando dependências circulares nas relações hierárquicas entre os conceitos na taxonomia.	Linguística computacional
Figueroa et al., 2019	Experimental	Muito boa	Usando o Sequential Forward Floating Search (SFFS) como linha de base para a aprendizagem multivisualização, observou-se que a informação linguística para a identificação de âncoras temporais e a pesquisa *web* para identificar fragmentos de texto relevantes são substanciais, o que justifica a realização da pesquisa. O estudo demonstrou que humanos e máquinas mostram diferentes graus de dificuldade ao rotular questões de diversos tópicos. Um tópico que é fácil de rotular por um ser humano pode ser difícil de rotular por uma máquina e vice-versa.	Linguística computacional
Giabelli et al., 2020	Experimental	Muito boa	O NEO foi avaliado em mais de 2 milhões de anúncios de emprego no Reino Unido a partir de 2018, identificando 49 novas ocupações das quais 43 foram confirmadas por meio de um estudo de usuários com 10 participantes. A correlação entre os julgamentos dos usuários e as métricas GASC do NEO foi positiva e estatisticamente significativa.	Planejamento de Políticas Públicas; Identificação de tendências em sistemas *downstream*.
Golpayegani et al., 2023	Experimental	Excelente	VAIR, uma taxonomia formal e FAIR para representar hierarquias de IA e conceitos de risco	IA; Mercado de seguros; Saúde; Governo; setor financeiro; Avaliação de riscos em geral.

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
Horta et al., 2023	Experimental	Muito boa	Resultados experimentais mostraram que as CNN podem aprender bem as relações de subclasses, especialmente quando são usadas ResNets com reconhecimento de hierarquia. A qualidade do conjunto de dados e da profundidade da arquitetura influenciam diretamente na adequação semântica. Os autores observam também que as técnicas existentes para injetar conhecimento externo aos modelos durante a fase de treinamento podem levar a melhores taxonomias.	IA e automatização de taxonomias em diferentes domínios
Hossain et al., 2021	Estudo de revisão		Experimentos extensos mostraram que ambos os módulos funcionam de forma eficaz na geração de uma taxonomia tópica de alta qualidade baseada em sementes fornecidas pelo usuário.	Organização do conhecimento em geral
Huang et al., 2020	Experimental	Muito boa	Os resultados experimentais demonstraram o desempenho, a eficiência e a escalabilidade da estrutura proposta, além da detecção automática de conflitos, sem intervenção humana.	Internet das Coisas
Huang et al., 2021	Experimental	Razoável	O trabalho fornece uma plataforma comum para representar e armazenar descrições de produtos complexas e heterogêneas na *web* semântica. Utiliza grafos do conhecimento na compreensão de atributos dos produtos e cálculos de similaridade com as especificações de entrada. Resultados de diversos experimentos comparativos com portais de e-commerce semelhantes demonstraram principalmente que a aplicação de tecnologias da *web* semântica e de inteligência artificial aprimoram a integração e a interoperabilidade de dados para a descoberta de produtos.	Taxonomia para interações com eletrodomésticos em casas inteligentes
Jain, 2023	Experimental	Excelente	A Probase é a taxonomia de maior cobertura dentre as concorrentes; apresenta uma precisão avaliada em 92,8% na média de todos os conceitos existentes (cerca de 5,4 milhões). O processo de conceituação do Microsoft Concept Graph é considerado efetivo e robusto.	Base de conhecimento para pesquisa científica, organização do conhecimento em geral
Ji et al., 2019	Experimental	Boa	Os experimentos conduzidos demonstraram que o TaxoEnrich alcançou melhor desempenho entre todas as métricas de avaliação e métodos de última geração considerados.	Manutenção automática de taxonomias em

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
				diferentes aplicações
Jiang et al., 2022	Experimental	Muito Boa	O estudo aponta que as abordagens tradicionais de pesquisa falham principalmente por manipular termos ao invés de conceitos, os quais carregam significado fundamental para a correta compreensão do que está sendo pesquisado. Editores de grandes corporas ou de conteúdo especializado apontam como principais problemas: a falta de variantes lexicais, PLN e outras combinações conceituais; a existência de ambiguidade na linguagem; e a dificuldade de manter vocabulários controlados devido às rápidas mudanças ocorridas ao longo do tempo.	Mecanismos de busca
Kasenchak, 2019	Experimental	Razoável	Os resultados mostraram que o CBF supera o Semantic Referee, usado como método comparativo, nos dois conjuntos de dados e alcança melhor desempenho de classificação e interpretabilidade. O método CBF melhora a segmentação semântica, especialmente para as categorias Vegetação, Solo, Pavimento e Água.	Sensoriamento remoto para diferentes finalidades
Li et al., 2022	Experimental	Excelente	Desenvolveu uma estrutura para avaliação de ontologia baseada em métricas que verifica a sintaxe, a estrutura e a semântica das ontologias. Criou o MOQA, uma ferramenta baseada na *web* que avalia ontologias enviadas no formato.owl em aspectos sintáticos, estruturais e semânticos. Demonstrou o uso de questões de competência dentro da estrutura para avaliar a capacidade da ontologia de responder a consultas específicas de um domínio. Testou com sucesso a ferramenta MOQA quanto à funcionalidade e eficácia na avaliação de ontologias.	Avaliação de ontologias
Malandri et al., 2020	Experimental	Muito boa	Os resultados experimentais mostram que o HSS supera as medidas de última geração para medir a similaridade semântica na taxonomia.O artigo demonstra que a incorporação selecionada por meio do MEET tem um desempenho melhor do que as incorporações selecionadas pela literatura sobre tarefas de benchmark.	Bases de conhecimento e organização do conhecimento em geral; varejo

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
			A incorporação de palavras provou ser eficaz na captura de semelhanças léxicas e semânticas para enriquecer taxonomias a partir de dados de texto.	online; pesquisa científica.
Malandri et al., 2021	Experimental	Muito boa	Os resultados experimentais mostraram que o MEET-LM é capaz de captar, incorporar e representar de forma eficaz as semelhanças entre esses empregos dentro da taxonomia padrão da ESCO à medida que emergem da procura real do mercado de trabalho. O MEET-LM alcança uma precisão de 99,4%.	Aprimoramento de dados e pesquisas semânticas em palataforma virtuais
Maldonado et al., 2019	Experimental	Boa	Vetores treinados em pseudocorpos de caminhada randomizada são capazes de codificar informações taxonômicas, conforme demonstrado pelo bom desempenho em uma tarefa de similaridade sinônimica. O estudo demonstrou que o enriquecimento taxonômico de vetores de *corpus* naturais por meio de combinação de vetores nem sempre aumenta o desempenho dos vetores de palavras resultantes, o que aponta cenários mais propícios para tal combinação. O enriquecimento através da combinação de vetores pode ajudar quando vetores de *corpus* naturais são treinados em corpora de pequeno a médio porte. Se forem treinados em corpora muito grandes, o enriquecimento taxonômico não oferece nenhum benefício e pode, na verdade, prejudicar o desempenho. A similaridade taxonômica é impulsionada principalmente por vetores de caminhada randomizada. Somente tamanhos médios de dados de caminhada randomizada são necessários: há poucos benefícios em treinar vetores em caminhadas randomizadas muito extensas. A combinação vetorial é útil quando o *corpus* natural é grande. Os experimentos mostraram que a força dos vetores taxonômicos via caminhada randomizada pelo WordNet veio da repetição de conceitos WordNet bem conectados nos pseudocorpora gerados, o que efetivamente reforçou as relações topológicas dentro da taxonomia (relações de proximidade, vizinhança, dependência, associação).	Normalização de bases de conhecimento; aprendizagem de máquina; criação, uso e gestão de corporas; modelos preditivos em diversas áreas.
Manzoor et al., 2020	Experimental	Muito boa	Através de uma avaliação extensiva de uma taxonomia em larga escala no Pinterest e de vários conjuntos de dados públicos, demonstramos que o	Manutenção e criação de

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
			Arborist supera o estado da arte. Os papéis taxonômicos inferidos induzem ainda uma categorização útil dos nós taxonômicos que merece uma exploração mais aprofundada.	taxonomias em escala *web*
Martel e Zouaq, 2021	Experimental	Boa	Os resultados demonstram que, ao contrário dos métodos baseados em agrupamento supervisionado, esta abordagem também pode identificar novas classes a partir dos dados e rotulá-las com axiomas lógicos, o que leva a um aprendizado expressivo de taxonomias.	Manutenção automática de taxonomias em diferentes aplicações
Musgrove e Walsh, 2019	Experimental	Razoável	O método extensional não é infalível por si só e se baseia na frequência de co-ocorrências de tags em uma folksonomia para validar uma ontologia formal; alta precisão na classificação de casos duvidosos é alcançada quando aplicadas as duas abordagens em conjunto, extensional e intencional.	Conformação de folksonomias
Patel et al., 2019	Teórica	Razoável	Poucos claros, mas basicamente mostram a eficácia das taxonomias na organização e representação do conhecimento, especialmente em bases de conhecimento. Aplica tecnologias da *web* semântica para mapeamento, correspondência, fusão, particionamento e raciocínio de ontologias para aplicação no ambiente *web* em geral, em bases de conhecimento e em organizações.	Estruturação de dados na *web*
Poirier, 2019	Teórica	Boa	"Os projetistas de infraestrutura semiótica não têm escolha a não ser considerar compensações quando se trata de decidir sobre a complexidade/restritividade, flexibilidade/dureza, desleixo/limpeza de seus padrões e estruturas. Para tomar decisões éticas e justas ao avaliar essas opções, os projetistas de infraestrutura semiótica precisam de habilidade, não apenas para encontrar os limites dos termos ou modelar logicamente o significado, mas também para discernir quando e por que é importante representar a complexidade e quando e por que é importante promover uma linguagem partilhada ou restritiva" (p.369, tradução nossa). O estudo aponta que o grupo denominado como projetistas de infraestrutura semiótica é contido por teóricos e praticantes, os quais se autodenominando especialistas em representação do conhecimento; ontologistas, cientistas de dados ou engenheiros de dados. Frequentemente possuem formação em filosofia analítica, ciência da computação ou ciência da informação, com	Linguística computacional; Ciência de dados; Modelos preditivos; Modelagem conceitual.

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
			conhecimentos em linguística teórica. Trata-se de um grupo que compartilha de uma meta comum, que é a construção de modelos de dados para preservação de significado conforme se movem por diferentes domínios. "Os projetistas de infraestrutura semiótica encontram constantemente exceções às regras que suas linguagens codificam. Mesmo quando pretendem explicitamente codificar uma diversidade de conceitos e ideias, são constantemente confrontados com mais conceitos e ideias que não se enquadram nas categorias identificadas. Eles estão descobrindo que a aplicação das abordagens formais para codificar a igualdade e a diferença teorizadas na filosofia analítica não pode ajudá-los a escapar deste vínculo" (p. 367-368, tradução nossa). "Para aqueles que abordam o projeto de infraestrutura semiótica com o objetivo explícito de representar diferença de maneiras justas e equitativas, descobrir quanta complexidade representar versus quanta complexidade restringir não é simples. Frequentemente, nomear uma diferença pode causar tanta violência quanto restringir uma linguagem para representar apenas certas diferenças. O desafio de conceber infraestruturas apenas semióticas exige considerar quando é que faz diferença política, ética e prática codificar a diferença. Na prossecução deste objetivo, os projetistas de infraestruturas semióticas em certos domínios começaram a adotar uma lente cognitiva diferente da lente analítica, que orientou historicamente o trabalho de representação do conhecimento; estão aprendendo a se sintonizar com as formas como os seus sistemas de representação do conhecimento estão envolvidos numa política de representação e, ao fazê-lo, trazem novas mentalidades sobre identidade e diferença para o seu trabalho"(p.368, tradução nossa)	
Prince et al., 2021	Experimental	Muito boa	O desempenho do DUI neste conjunto de dados foi comparável ao das postagens do Reddit. Conclui-se então que o DUI é eficaz na identificação de termos relacionados ao uso de drogas e pode contribuir para a análises semânticas no domínio.	Análise de dados epidemiológicos; Saúde

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
Raza et al., 2019	Experimental	Razoável	Apresenta a estrutura do conhecimento como potencial fonte de termos semanticamente relacionados, são exemplos os vocabulários controlados de domínios específicos ou gerais, as relações taxonômicas e as relações não-taxonômicas. O objetivo de usar essas relações taxonômicas simples na expansão da consulta (QE) é permitir a recuperação de conceitos mais gerais ou específicos relacionados aos termos da consulta de pesquisa. Ao percorrer a taxonomia para cima ou para baixo, o processo de QE pode expandir a consulta com conceitos estreitamente relacionados. No entanto, um desafio em usar essas relações taxonômicas simples é determinar a distância hierárquica apropriada para obter conceitos de expansão candidatos. Em outras palavras, é crucial decidir quantos níveis de distância do conceito original devem ser considerados ao selecionar conceitos de expansão da estrutura de conhecimento; a recomendação é de uma distância conceitual de menos de cinco níveis. A abordagem semântica de QE do tipo linguística é mais efetiva para desambiguação; o tipo baseado em ontologia nem sempre melhora a eficiência da recuperação da informação por conter termos ambíguos e muito gerais; e a abordagem híbrida, apesar de complexa, explora o que há de mlehor das abordagens linguística e baseada em ontologia.	Sistemas de recuperação da informação de domínios específicos
Salatino et al., 2020	Experimental	Boa	Os resultados apontam que o novo CSO supera as abordagens atuais de evolução de ontologia, demonstra maior precisão e recordação na previsão de novas áreas de pesquisa em comparação com métodos baseados em palavras-chave e alcançou a maior média de concordância em comparações com classificações feitas por anotadores especialistas de domínio.	No âmbito científico no campo da Ciência da Computação e áreas afins.
Sha et al., 2023	Experimental	Boa	A abordagem proposta demonstrou que contribui para a ligação de informações não estruturadas com dados estruturados; demonstra que taxonomias como a Wikidata podem otimizar a estrutura do tópico a partir de outras plataformas; pode auxiliar na classificação e análise de conteúdo Web multilíngue e multimodal; é passível de aplicação a outras redes sociais; pode servir como referência para avaliar o desempenho de tarefas de classificação hierárquica semelhantes e fornecer funcionalidade de filtragem para ajustar	Ecossistemas de informação

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
			classificadores existentes com base na profundidade semântica personalizada.	
Shang et al., 2020	Experimental	Boa	Foi criada uma nova estrutura de incorporação e agrupamento de termos hierárquicos NetTaxo, que consome uma rede rica em texto como entrada, para a construção automática de taxonomia de tópicos. Experimentos extensos em dois conjuntos de dados demonstram a superioridade da estrutura em comparação com as linhas de base. Experimentos de ablação confirmaram a necessidade e eficácia da seleção de motifs proposta em nível de instância.	aplicações downstream, pesquisa *web* e recomendação de conteúdo
Sharifpour et al., 2023	Experimental	Boa	A semelhança semântica identificou seis perfis ou classes: Expert Research, Expert Search, Expert Explore, Novice Research, Novice Search e Novice Explore. Os autores propuseram uma taxonomia de perfis de usuários de pesquisa de nível de especialização que compartilha semelhanças com taxonomias previamente identificadas na área de pesquisa na *web* e no comércio eletrônico. A taxonomia proposta contempla tanto a experiência dos usuários no domínio, quanto o objetivo ou a motivação da busca no portal de dados de pesquisa, contribuindo para identificar uma classificação mais significativa dos utilizadores, tendo em conta não só o seu comportamento de pesquisa, mas também as motivações subjacentes às suas características de pesquisa.	Área acadêmica; Bases de conhecimento; Mídias sociais; Datasets
Shen et al., 2020	Experimental	Excelente	Os principais resultados do estudo mostraram que a TaxoExpan pode gerar taxonomias conceituais de alta qualidade em domínios científicos e alcançar desempenho de última geração no desafio de expansão da taxonomia WordNet. Descobriu-se nos experimentos que ela também pode ser utilizada para limpar uma taxonomia existente com eficiência e confiabilidade.	Plataformas digitais em geral, especialmente e-commerce e bibliotecas
Smith et al., 2022	Experimental	Excelente	Recomenda-se o uso dos princípios da OBO Foundry. A restrição do escopo às áreas abrangidas pelas taxonomias de ocupação popular, com foco nas vocações, permite que a riqueza de informações patrimoniais já existentes possa ser aproveitada. O emprego da estrutura ontológica de nível superior do BFO ilustraria uma nova implementação da primeira ontologia padronizada internacionalmente. Nosso exemplo de taxonomia selecionado, o SOC dos	economia, educação, investigação, infraestruturas organizacionais, bem como para a

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
			EUA, permitiu que os principais agrupamentos fossem alinhados com o campo de trabalho. No entanto, com o campo de trabalho tratado como uma propriedade de objeto, pode ser preferível o uso dos dez grupos principais relacionados ao nível de habilidade da ISCO. A implementação da ISCO na ESCO da UE tem melhorias significativas e é sugerida como uma melhor base para começar, seguida pela incorporação de elementos adicionais de outras taxonomias padrão. A seleção cuidadosa e o uso das propriedades do objeto melhorariam significativamente o OccO, sugerindo que o ISIC fosse usado para capturar o campo de trabalho, e a existente BFO Environment Ontology ENVO fosse usada para a propriedade do ambiente. Competências, habilidades e áreas de conhecimento estão representadas entre as taxonomias padrão, mas a harmonização entre estas referências é necessária e seria uma realização benéfica. O Wikidata ou outros recursos da Web Semântica poderiam empregar os resultados do OccO para melhor categorizar as ocupações; esses recursos são importantes demais para não terem uma estrutura ontológica sólida. Uma vez implementada a estrutura da ontologia ocupacional, ela poderá ser o modelo para esferas adicionais da atividade humana: as ocupações, os papéis familiares e os papéis sociais poderiam todos ser implementados dentro desta estrutura. Ter uma tal representação integrada da actividade humana pode trazer benefícios generalizados na coordenação das actividades mundiais em medicina, política social, educação, emprego e serviços sociais.	análise de saúde e segurança
Soares et al., 2020	Experimental	Razoável	A visão de modelagem de ontologia KES oferece suporte à modelagem eficiente de ontologias contextualizadas, enquanto a visão estrutural apóia a exploração, curadoria e integração com conteúdo não simbólico.	Projetos de IA
Teixeira e Sousa et al., 2020	Experimental	Boa	O processo de análise dos metadados para a construção da taxonomia revelou inconsistências terminológicas ou incongruências na estrutura de representação temática, potenciais ruídos para a recuperação da informação.	Representação da informação e do conhecimento em domínios ligados a minorias

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
Timmaraju et al., 2020	Experimental	Razoável	A análise de hashtags pode contribuir para a classificação de conteúdo *web* quando analisadas por meio de modelos pré-treinados e redes neurais. "Compreender o significado semântico do conteúdo na *web* através das lentes de uma taxonomia tem muitas vantagens práticas" (p. 8134, tradução nossa).	Gestão de conteúdo *web*; Gestão e análise de mídias sociais
Wang et al., 2019b	Experimental	Boa	A precisão do algoritmo e a taxa de detecção de ruídos são altas. O algoritmo elimina mais ruídos de forma eficaz do que o Agony2, um algoritmo anterior utilizado como comparativo.	Inteligência Artificial (chatbots, assistentes virtuais); Aprendizagem de Máquina; Linguística computacional.
Wang et al., 2019	Experimental	Muito boa	O modelo proposto é comparável aos principais concorrentes em termos de acurácia. O estudo mostra que o TEAL pode inferir novas relações de hiperonímia por meio do enriquecimento automático de taxonomias com 87% de precisão. Também é capaz de prever hiperonímia para outras línguas e elos perdidos em taxonomias. O AS-TEAL pode aproveitar tanto vetores derivados de grandes *corpus* quanto as taxonomias existentes, benéficas para predição em domínios específicos. "a performance da predição de hiperonímia pode ser melhorada pela integração de taxonomias" (p.1733. tradução nossa) Aponta que taxonomias probabilísticas como Probase, utilizada por meio do Microsoft Concept Graph, contém geralmente muitos erros que precisam ser filtradas e avaliadas, quando ocorrem menos de cinco vezes são eliminadas por serem dadas como não confiáveis.	Linguística computacional
Wang et al., 2019a	Experimental	Muito boa	Houve progresso na previsão de hiperonímia em inglês, mas existem desafios para outros idiomas devido à dependência do idioma e à falta de dados de treinamento. O Multi-Wahba Projection (MWP) supera todas as abordagens básicas	Aprendizagem de máquina, Modelos preditivos

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
			anteriores. O MWP tem maior precisão do que HyperVec, BLESS e ENTAILMENT.	
Wen et al., 2021	Experimental	Boa	Curadoria de taxonomias do domínio da segurança cibernética.	Segurança cibernética
Weser et al., 2020	Experimental	Boa	A representação formal de ontologias é uma tecnologia adequada para codificar conhecimento que anteriormente era descrito de forma não estruturada ou nem sequer descrito. Os modelos de ontologia podem ser mais facilmente mantidos e reutilizados quando separados da linguagem de programação. A especificação de capacidades com base em um vocabulário padronizado permite a substituição e interação contínua de recursos de diferentes fabricantes.	Automação industrial
Yaman et al., 2021	Experimental	Razoável	demonstrou a contribuição da tecnologia da Web Semântica para a governança de dados espaciais, bem como a sua importância para a publicação de dados.	Dados FAIR para organizações governamentais e privadas.
Yang et al., 2019	Experimental	Muito boa	Um *corpus* de domínio de engenharia de sistemas é criado com tags de classes gramaticais. Os resultados da geração semiautomática, quando comparados com os construídos manualmente, denotam que a ontologia gerada automaticamente possui excelente precisão e recuperação na extração de conceitos, o que abre um novo capítulo no desenvolvimento de ontologias de engenharia de sistemas.	Engenharia de domínio; análise de domínio
Zahera e Sherif, 2020	Experimental	Regular	Os experimentos sugerem que o ProBERT é uma boa base para avaliar a tarefa de classificação de produtos.	*E-commerce*
Zakharova, 2020	Experimental	Regular	A conveniência do uso de DL é justificada em todas as etapas da solução do problema de descoberta. É a ferramenta que fornece a definição formal do serviço e dos raciocinadores que possibilitam as heranças lógicas. Cada serviço é representado pela ontologia DL baseada na ontologia de nível superior. A ontologia do serviço é baseada na base de conhecimento da ontologia DL integrada do domínio.	Arquitetura de sistemas e serviços *web*

Estudo	Tipo de pesquisa	Avaliação da qualidade	Resultados	Aplicabilidade
			As ontologias por si só não descrevem as possibilidades de serviços, mas fornecem conceitos adicionais que permitem compreender corretamente a especificação das possibilidades de serviços. A criação e utilização da ontologia especial dos tipos de conceitos mais comuns permite aumentar a eficiência das soluções de descoberta e permite categorizar/classificar os serviços: por área de utilização, funções realizadas, etc.	
Zhang e Sun, 2022	Experimental	Muito boa	Os resultados experimentais mostram que o método proposto obteve melhor desempenho no conjunto de dados padrão M&C.	Estatística
Zhang et al., 2021b	Experimental	Muito boa	O modelo extrai com acurácia os conceitos e melhora a performance das classificações realizadas e a nomeação de entidades. A abordagem foi considerada eficaz. O AliCG está atualmente implantado no Alibaba para oferecer suporte a uma variedade de cenários de negócios, incluindo o produto Alibaba UC Browser.	E-commerce
Zhang et al., 2021c	Experimental	Muito boa	Além de ser capaz de detectar e categorizar a evasão do lado do cliente com alta precisão, nossa abordagem também pode rastrear a origem de diferentes implementações.	Prevenção de crimes cibernéticos
Zhang et al., 2021a	Experimental	Excelente	Experimentos em quatro aplicações do mundo real avaliam a abordagem proposta e concluem que o TMN apresenta a melhor performance em ambas as tarefas de preenchimento de taxonomias e tarefas prévias de expansão automatizada.	Linguística computacional; modelagem de dados
Zouaq e Martel, 2020	Experimental	Boa	Avaliar a qualidade de taxonomias expressivas é difícil pela inexistência de um padrão ouro para tanto. É possível extrair uma taxonomia de um grafo do conhecimento de modo que a taxonomia reflita o conteúdo do grafo ou suas instâncias de dados.	Construção de taxonomias e manutenção de ontologias.

Fonte: elaborado pelo autor (2023)

APÊNDICE E - Abordagens utilizadas para criação, gestão e/ou uso de taxonomias na *web* semântica nos últimos cinco anos

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
Alonso et al., 2020	Experimental	Razoável	É proposto o uso de uma taxonomia para filtragem e classificação de pontos de interesse geográficos relevantes para consultas recreativas.	Taxonomia na modelagem do conhecimento	Aplicações *web* em geral que possuam alguma funcionalidade de recomendação por localização.
Arous et al., 2023	Experimental	Muito boa	Conclusão de taxonomias / manutenção automática de taxonomias, enquanto instrumentos fundamentais de guarda de conceitos e suas relações.	Manutenção e curadoria automatizada de taxonomias	*E-commerce* e plataformas baseadas em conhecimento
Bongiovanni et al., 2023	Experimental	Boa	O estudo busca propor um método de classificação textual com base em uma taxonomia hierárquica personalizada e sem dados rotulados (anotados), que utiliza modelagem profunda de linguagem, do inglês Deep Language Modeling (DLM). A estrutura hierárquica da taxonomia se mostrou fundamental para a aprendizagem e inferência dos rótulos.	Taxonomia como classificação hierárquica em modelos de aprendizagem profunda e PLN	Datasets; Mídias sociais e folksonomias
Caucheteux et al., 2021	Experimental	Boa	Uso de taxonomia como base na modelagem conceitual utilizada para aprendizagem com uso de redes neurais profundas.	Uso de taxonomia na modelagem conceitual	Linguística e Linguística computacional
Crotti Junior et al., 2020	Experimental	Razoável	Uma taxonomia de gerenciamento de riscos clínicos foi manualmente desenvolvida e posteriormente expressa em SKOS para análise automatizada de dados qualitativos. Importância da taxonomia para semântica no LD	Taxonomia como vocabulário controlado e esquema de classificação	Psicologia organizacional; gerenciamento de riscos para sistemas em diferentes domínios, tais como saúde,

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
					aviação e finanças.
Danilevitch, 2019	Experimental	Boa	Uso de taxonomias na modelagem de dados e na integração semântica de domínios. "A estrutura construída sobre taxonomias e ontologias altamente expressivas permite gerar dados automaticamente a partir de sua fonte com o mínimo de aumento e preparação adicional. Esses dados estruturados podem ser analisados posteriormente por mecanismos de IA legíveis por máquina e torná-los comparáveis e nas configurações contextuais mais complexas. Para conceber adequadamente taxonomias altamente expressivas, os modeladores de dados devem levar em consideração os aspectos linguísticos das línguas naturais, as regras técnicas do domínio profissional específico e os meios aplicáveis das tecnologias informáticas. Acreditamos que a pesquisa de todos esses aspectos geraria resultados úteis que avançariam ainda mais no campo da Web Semântica em geral"(p. 4, tradução nossa). Para domínios de conhecimento complexos diretamente ligados a pronunciamentos, regras ou leis regulatórias, a classificação deve ser projetada por pessoas que entendam a natureza técnica dessas regras de negócios. A modelagem de dados em taxonomias XBRL representa um exemplo de abordagem baseada em humanos para classificação de informações legíveis por humanos em um formato legível por máquina. Acreditamos que a pesquisa proposta avançaria no campo da Web Semântica em	Construção de taxonomia de domínio	Linguística computacional

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
			relação aos meios de transformação de dados não estruturados em relatórios estruturados.		
Daraio et al., 2020	Experimental	Boa	Mapeamento, categorização ou construção de taxonomias por meio de IA e manutenção de taxonomias por meio de modelos preditivos. "O presente artigo aborda a tarefa de mapeamento automatizado de taxonomia, que é um subdo problema mais geral de mapeamento de ontologias".	Construção automática de taxonomias	e-commerce; Internet of Things
Dilorenzo et al., 2020	Experimental	Excelente	É sabido que as taxonomias facetadas são adequadas para organizar ativos reutilizáveis de desenvolvimento de software dada a sua flexibilidade para expansão e manutenção, além da facilidade de implementação como um banco de dados relacional. Abre um novo campo de pesquisa que explora o reuso de artefatos de software em projetos ágeis, mesmo que entre projetos diferentes, bem como a possibilidade de adicionar algoritmos inteligentes para auxiliar desenvolvedores de software. Taxonomias com potencial de contribuição para o reuso no ambiente *web*	Taxonomia como esquema de classificação; taxonomia facetada para classificação e recuperação de informação	Gerenciamento de projetos; suporte à decisão; *dataset* com histórias de usuários previamente classificadas;
Doan et al., 2020	Experimental	Muito boa	Apresenta abordagens de extração taxonômica e aponta o desafio de coletar conhecimento relevante da *web* para a construção de taxonomias consistentes.	Extração de relações taxonômicas como base para o aprimoramento semântico	Linguística computacional
Fauceglia et al., 2019	Experimental	Razoável	O estudo demonstra a importância das taxonomias para o processamento de linguagem natural e para a aprendizagem de máquina. Enquanto estruturas as taxonomias adicionam semântica ao conjunto de dados, facilitando a aprendizagem e guiando o processamento de consultas. Os relacionamentos taxonômicos estabelecidos contribuem para a ligação	Taxonomias conferindo semântica a conjuntos de dados	Linguística computacional

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
			entre termos e conceitos num conjunto de dados complexo. No estudo, as taxonomias são a estrutura e o produto principal gerado pelo KGIS e armazenado como SSS para manipulação. A consistência dos relacionamentos gerados promove melhor compreensão dos dados para a geração de informação a partir deles e permite que a semântica seja expandida e aprofundada por outros conjuntos de dados advindos de bases de conhecimento também estruturadas como taxonomias ou por usuários. Desse modo, podemos concluir que ao mesmo tempo em que as taxonomias estruturam, também dão flexibilidade para o crescimento sustentável do conjunto e possíveis integrações.		
Figueroa et al., 2019	Experimental	Muito boa	Aplicação de taxonomias como base para o processamento de linguagem natural e especificamente para aplicação de análises linguísticas por meio de máquinas.	Uso de taxonomias como anotação de corpus; uso de facetas temporais na resposta a perguntas comunitárias.	Linguística computacional
Giabelli et al., 2020	Experimental	Muito boa	Taxonomias como instrumento de primeira escolha quando se precisa conhecer um domínio e, principalmente revelar conhecimento emergente. Atuam como um SOC que permite compreender com rapidez e facilidade as dinâmicas de evolução e mudança de determinado campo ou temática ao longo do tempo, seja por especialistas ou não especialistas.	Enriquecimento de taxonomia existente	Planejamento de Políticas Públicas; Identificação de tendências em sistemas *downstream*.
Golpayegani et al., 2023	Experimental	Excelente	Aplicação de taxonomias na classificação de riscos como instrumento que confere confiabilidade ao sistema	Uso de taxonomias existentes; taxonomia de domínio	IA; Mercado de seguros; Saúde; Governo; setor financeiro;

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
					Avaliação de riscos em geral.
Horta et al., 2023	Experimental	Muito boa	"Avanços recentes em IA explicável mostraram que as CNN são capazes de aprender relações hierárquicas entre categorias semânticas na forma de classificações taxonômicas" (n.p.; tradução nossa)	Extração de relações taxonômicas por redes neurais convolucionais	IA e automatização de taxonomias em diferentes domínios
Hossain et al., 2021	Estudo de revisão		"A taxonomia não é apenas uma forma fundamental de representação do conhecimento, mas também crucial para vastas aplicações ricas em conhecimento, como resposta a perguntas e pesquisa na *web*. A maioria dos métodos de construção de taxonomia existentes extraem pares de entidades hiperônimo-hipônimo para organizar uma taxonomia 'universal'" (n.p., tradução nossa).	Taxonomia como esquema de classificação e ferramenta semântica para conjuntos de dados	Organização do conhecimento em geral
Huang et al., 2020	Experimental	Muito boa	Taxonomia para interações com eletrodomésticos em casas inteligentes, numa abordagem orientada a objetos para modelagem de dispositivos.	Taxonomia de domínio para compreensão semântica em aprendizagem de máquina	Internet das Coisas
Huang et al., 2021	Experimental	Razoável	"Metadados, taxonomias e ontologias juntos estruturam os dados não estruturados e fornecem informações às máquinas em um formato legível por máquina, para que possam ser feitas inferências significativas." (p. 9182, tradução nossa) ontologias existentes, como eCLSS OWL e ontologia Good Relations, usam taxonomias para representação textual de metadados de produtos. A ontologia de Equipamentos de Engenharia (EEO) serve como um modelo geral de dados e estrutura de taxonomia para descrições de produtos.	Taxonomia como uma das tecnologias da *web* semântica fundamentais para interoperabilidade.	Taxonomia para interações com eletrodomésticos em casas inteligentes

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
Jain, 2023	Experimental	Excelente	Extraída Figura 1 e adaptada	data mining e aprendizagem de máquina para construção de taxonomias	Base de conhecimento para pesquisa científica, organização do conhecimento em geral
Ji et al., 2019	Experimental	Boa	Manutenção e curadoria automatizada de taxonomias	Conclusão e expansão de taxonomia; taxonomia de domínio	Manutenção automática de taxonomias em diferentes aplicações
Jiang et al., 2022	Experimental	Muito Boa	O estudo traz a importância das taxonomias para o aprimoramento da experiência de busca e recuperação da informação sob dois aspectos, (a) como a principal ferramenta no alcance do objetivo de uma pesquisa semântica, que é acessar conceitos escondidos em termos em documentos e bases de dados; e (b) aproveitamento de metadados semânticos advindos das taxonomias, além da marcação de conteúdo útil para grandes e complexos corporas, já que permite uma pesquisa mais granular e precisa ao associar o conteúdo a conceitos ou tópicos específicos.	Taxonomias como fonte de metadados	Mecanismos de busca
Kasenchak, 2019	Experimental	Razoável	Uso de taxonomia em mecanismos de inferência dentro da ontologia e aprendizagem profunda	Taxonomia da ontologia/taxonomia de domínio e raciocínio taxonômico	Sensoriamento remoto para diferentes finalidades
Li et al., 2022	Experimental	Excelente	Uma avaliação sintática da taxonomia da ontologia é empreendida para verificação de inconsistências, erros e redundâncias. A avaliação feita por similaridade semântica considera	Taxonomias na identificação inconsistências, erros e redundâncias, no	Avaliação de ontologias

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
			o conhecimento definido na taxonomia, considerado o mais confiável.	contexto da avaliação de ontologias.	
Malandri et al., 2020	Experimental	Muito boa	As taxonomias são a base de uma ampla gama de aplicações na *web* semântica e fornecem suporte formal para trocas sintáticas e semânticas. Conseqüentemente, eles são essenciais para a compreensão da máquina no processamento de linguagem natural e em tarefas orientadas para negócios. No entanto, manter essas hierarquias atualizadas e capazes de representar o domínio do qual foram extraídas ainda é uma atividade demorada e propensa a erros, ponto onde a aprendizagem de máquina atua. No âmbito da similaridade semântica: tarefas downstream dependem da representação de conhecimento estruturado subjacente, como recomendações personalizadas para varejistas on-line, compreensão de consultas para mecanismos de pesquisa e compreensão de texto.	Refinamento, Enriquecimento e Aprendizagem de taxonomias por meio de vetores Manutenção de taxonomias por meio de aprendizagem de máquina	Bases de conhecimento e organização do conhecimento em geral; varejo online; pesquisa científica.
Malandri et al., 2021	Experimental	Muito boa	Uma taxonomia existente é utilizada e enriquecida com dados	Taxonomia de domínio	Aprimoramento de dados e pesquisas semânticas em palataforma virtuais
Maldonado et al., 2019	Experimental	Boa	O estudo contribui significativamente para a compreensão das facetas que envolvem as relações hierárquicas numa taxonomia e seu papel para a linguística computacional.	uso de bases de conhecimento com dados taxonômicos para melhor desempenho na aprendizagem de	Normalização de bases de conhecimento; aprendizagem de máquina; criação, uso e gestão de

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
				máquina; enriquecimento taxonômico para melhor compreensão das informações e do conhecimento existente nas bases.	corporas; modelos preditivos em diversas áreas.
Manzoor et al., 2020	Experimental	Muito boa	"Taxonomias curadas melhoram o desempenho de sistemas de aprendizagem de máquina por meio de conhecimento estruturado de alta qualidade. No entanto, a curadoria manual de uma taxonomia grande e em rápida evolução é inviável" (p. 2044, tradução nossa).	A contribuição da curadoria de taxonomias para a qualidade do conhecimento utilizado pelos sistemas de aprendizagem de máquina. Expansão e indução de taxonomia Predição de hiperônimos e aprendizado de taxonomia	Manutenção e criação de taxonomias em escala *web*
Martel e Zouaq, 2021	Experimental	Boa	ontologias e taxonomias são cruciais para a Web Semântica, mas são caras de construir manualmente, o que estimula pesquisas sobre automação. *Expressive taxonomy*: taxonomias expressivas são taxonomias aprimoradas que incluem novas classes identificadas a partir de dados e descritas com axiomas lógicos. Talvez seja um exemplo prático do ponto exato onde uma taxonomia e uma ontologia independentes se tocam e se complementam.	Taxonomia no contexto de ontologias e criação de KG para expansão de taxonomias. Taxonomia hierárquica taxonomia em árvore	Manutenção automática de taxonomias em diferentes aplicações

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
Musgrove e Walsh, 2019	Experimental	Razoável	Aborda o relacionamento entre ontologias, taxonomias e folksonomias	Complementaridade na aplicação de ontologias, taxonomias e folksonomias	Conformação de folksonomias
Patel et al., 2019	Teórica	Razoável	"A representação taxonômica é uma forma poderosa de organizar o conhecimento. A ideia básica da taxonomia é que um item é subordinado a outro item. As classes são inseridas na taxonomia e a classe mais específica herda automaticamente todas as propriedades da classe mais geral, à qual está diretamente subordinada na taxonomia. A classe mais geral é chamada de superclasse e as classes especializadas são chamadas de subclasses. A propriedade de herança da representação do conhecimento taxonômico permite herdar propriedades de múltiplas superclasses. Especificidade e generalização são dois operadores usados para derivar heranças numa hierarquia" (p. 658, tradução nossa). Utiliza o termo ontologia ao discutir a visão de dados da *web* semântica como unidades de conhecimento, porém exalta aspectos de atribuição de conceitos, estruturação hierárquica e mecanismos de herança, característicos das taxonomias, que podem compor ontologias no ambiente *web*. "A alma da *web* semântica é a ontologia que mantém a hierarquia das entidades e provê mecanismos de herança"(p. 654, tradução nossa). Os autores apontam dois principais benefícios do conhecimento taxonômico ou das taxonomias: redução de redundâncias e de complexidade nos mecanismos	exploração teórica da representação do conhecimento na *web* semântica representação taxonômica	Estruturação de dados na *web*

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
			de busca. Citam de forma não detalhada a importância do conhecimento taxonômico para a precisão dos resultados em mecanismos de busca e recuperação. Os autores abordam a *web* semântica como uma *web* inteligente, onde predomina o raciocínio de máquina. Assim, lançam luz à aprendizagem de máquina e apontam que todos os componentes de conhecimento da inteligência artificial estão armazenados numa estrutura hierárquica e suas unidades de conhecimento são capazes de receber novos conhecimentos sem que antigos sejam afetados, aspecto que agrega valor ao raciocínio que busca assemelhar-se ao comportamento humano (memorização, pensamento e imaginação). Dialoga com Poirier (2019) ao afirmar que a base de conhecimento da *web* semântica deve ser o mais generalizada possível para comportar a natureza dinâmica das informações globais. Para Poirier, os desafios impostos pelas especificidades de diferentes naturezas no mundo devem ser encarados por meio do aprofundamento da capacidade de conceituação das diferenças com maior riqueza de detalhes e sensibilidade humana.		
Poirier, 2019	Teórica	Boa	O estudo é fundamental para o aprofundamento da discussão sobre as diversas facetas da estrutura hierárquica das taxonomias na modelagem conceitual de dados e no refinamento de bases de conhecimento e corpora. Também coloca as taxonomias e as ontologias, enquanto SOC, no núcleo dos desafios de representação do conhecimento em meio à diversidade e às diferenças socioculturais.	Aborda aspectos do papel semântico dos SOC no ambiente *web*.	Linguística computacional; Ciência de dados; Modelos preditivos; Modelagem conceitual.

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
			Traz o aspecto da informação taxonômica como informação de contexto fundamental na representação de dados. "nunca foi tão importante garantir que os conceitos representados nos dados sejam contextualizados de forma robusta (p. 369, tradução nossa)		
Prince et al., 2021	Experimental	Muito boa	Taxonomia de domínio construída manualmente e incremento por meios automatizados de aprendizagem computacional.	Taxonomia de domínio e expansão automática de taxonomias	Análise de dados epidemiológicos; Saúde
Raza et al., 2019	Experimental	Razoável	Importância dos relacionamentos taxonômicos para a recuperação da informação	Taxonomia hierárquica	Sistemas de recuperação da informação de domínios específicos
Salatino et al., 2020	Experimental	Boa	Contribui para a compreensão do papel das taxonomias como estruturantes de um domínio e dentro do âmbito da ontologia, além do auxílio à recuperação de informações e descoberta de novos conhecimentos. O objetivo final dos autores é criar um conjunto abrangente de taxonomias de tópicos em grande escala que descrevam os diferentes ramos da ciência.	Taxonomia como esquema de classificação e ferramenta semântica associada a ontologias.	No âmbito científico no campo da Ciência da Computação e áreas afins.
Sha et al., 2023	Experimental	Boa	Importância da taxonomia de domínio, mesmo no contexto de bases de conhecimento e aprendizagem de máquina	Taxonomia de domínio; taxonomia da ontologia	Ecossistemas de informação
Shang et al., 2020	Experimental	Boa	"uma das maiores vantagens da construção automática de taxonomias é a habilidade de captura de informações de *corpus* específicos e adotar para diferentes cenários" (n.p.) "organizar artigos científicos dentro de uma estrutura de taxonomia dá aos pesquisadores uma visão de	Taxonomia em árvore	aplicações *downstream*, pesquisa *web* e recomendação de conteúdo

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
			águia sobre o campo, e então eles podem rapidamente identificar seus interesses, e facilmente adquirir a informação que desejam" (n.p.)		
Sharifpour et al., 2023	Experimental	Boa	É construída a primeira taxonomia baseada em *logs* de usuários em um *dataset* de dados de pesquisa	Construção de taxonomia	Área acadêmica; Bases de conhecimento; Mídias sociais; *Datasets*
Shen et al., 2020	Experimental	Excelente	Emprega uma taxonomia para organizar abordagens semânticas de expansão de consulta	Expansão e indução automática de taxonomias	Plataformas digitais em geral, especialmente e-commerce e bibliotecas
Smith et al., 2022	Experimental	Excelente	São listadas diversas taxonomias existentes e informações sobre elas, além de comparações valiosas para o desenvolvimento, manutenção e expansão de outras taxonomias, além do desenvolvimento de pesquisas.	Uso de taxonomias existentes como forma de padronização e criação de consenso em um domínio	economia, educação, investigação, infraestruturas organizacionais, bem como para a análise de saúde e segurança
Soares et al., 2020	Experimental	Razoável	Aponta que os conjuntos de dados RDF têm interpretação ambígua e não têm relações explícitas com os subgráficos. Assim como as abordagens atuais de modelagem de ontologias não podem abordar o conhecimento contextual.	Mecanismo de visualização de taxonomias e sua importância na organização do conhecimento, especialmente o contextual (instâncias).	Projetos de IA

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
Teixeira e Sousa et al., 2020	Experimental	Boa	Taxonomias como esquema de representação temática da informação, tornando-se um vocabulário comum para o tema, qualificado e ponto de partida para o desenvolvimento de estudos sobre lesbianidade e temáticas afins.	Taxonomias como esquema de representação temática da informação	Representação da informação e do conhecimento em domínios ligados a minorias
Timmaraju et al., 2020	Experimental	Razoável	"Compreender o significado semântico do conteúdo na *web* através das lentes de uma taxonomia tem muitas vantagens práticas" (p. 8134) Aponta que as taxonomias fornecem uma estrutura semântica para classificar e compreender conteúdo *web*. Destaca os desafios de rotular documentos dentro de uma grande taxonomia. As taxonomias fornecem uma lente estruturada para interpretar uma grande variedade de conteúdo gerado pelos usuários nas plataformas *web*. As taxonomias facilitam a classificação de conteúdo em grande escala, hoje propiciada por tecnologias de processamento de linguagem natural e inteligências artificiais.	Taxonomia como esquema de classificação; taxonomias no contexto da aprendizagem profunda e uso de inteligência artificial.	Gestão de conteúdo *web*; Gestão e análise de mídias sociais
Wang et al., 2019b	Experimental	Boa	Taxonomias contribuem para a precisão e a confiabilidade de uma base de conhecimento, uma vez que dão suporte para identificação de relacionamentos ruidosos que prejudicam a correta compreensão das informações ali existentes.	identificação de ruídos nos termos e nos relacionamentos tipo-instância numa taxonomia lexical construída automaticamente	Inteligência Artificial (chatbots, assistentes virtuais); Aprendizagem de Máquina; linguística computacional.
Wang et al., 2019	Experimental	Muito boa	Muitos modelos e aplicações são taxonomias denominadas de outra forma, as mais comuns são Knowledge graph e base de conhecimento (não todas, mas aquelas que focam nas relações tipo-instância).	usa o Microsoft Concept Graph como taxonomia; "a performance da	Liguística computacional

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
			Extraída Figura 1 página 7130	predição de hiperonímia pode ser melhorada pela integração de taxonomias" (p.1733. tradução nossa)	
Wang et al., 2019a	Experimental	Muito boa	"A semântica das relações de hiperonímia entre pares conceito-conceito (subclasse de) e pares conceito-instância (instância de) são diferentes. Tais pares, no entanto, são frequentemente misturados em cenários do mundo real"(p. 1965).	Discute hiperônimos, um dos componentes das taxonomias	Aprendizagem de máquina, Modelos preditivos
Wen et al., 2021	Experimental	Boa	Taxonomias como vocabulário controlado estruturante para organização do domínio da cibersegurança Pesquisamos e revisamos os padrões e ontologias de segurança cibernética existentes para incorporar as taxonomias mais comumente usadas na ontologia e para fornecer um entendimento comum do domínio da segurança cibernética. Taxonomias existentes: Common Attack Pattern Enumeration and Classification (CAPEC) European Union Agency for Cybersecurity (ENISA) Software Engineering Institution (SEI) Common Weakness Enumeration (CWE)	Modelagem de ontologia com base na construção de uma taxonomia e no uso de taxonomias externas para complementaridade de informações semânticas.	Segurança cibernética
Weser et al., 2020	Experimental	Boa	Ao confiar em uma taxonomia conhecida publicamente, ela garante que haja um entendimento e um acordo comuns sobre a classificação e categorização das capacidades. Isso permite que descrições e inferências consistentes e padronizadas sejam feitas.	Como base para raciocinar e fazer inferências sobre a capacidade produtiva e vocabulário controlado.	Automação industrial

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
Yaman et al., 2021	Experimental	Razoável	Este artigo apresenta uma taxonomia SKOS de licença aberta documentada para os temas de dados UN-GGIM que segue as melhores práticas de Linked Data. Boas Práticas de Gestão de Dados Abertos Conectados	Taxonomia como parte fundamental na construção de um ambiente de dados conectados. Foi implementada como um vocabulário controlado de dados conectados para padronizar definições.	Dados FAIR para organizações governamentais e privadas.
Yang et al., 2019	Experimental	Muito boa	Extraída figura de camadas ontológicas. Demonstra a importância da camada taxonômica para o aprofundamento do estudo ontológico de domínio. "O processo de aprendizagem de ontologia proposto é um ponto de partida inicial para gerar ontologias de domínio de engenharia de sistemas. Vamos continuar o estudo em profundidade para focar na camada taxonômica no futuro".(n.p.)	Taxonomia da ontologia	Engenharia de domínio; análise de domínio
Zahera e Sherif, 2020	Experimental	Regular	Construção automática de taxonomias, previsão de categorias e taxonomias na estruturação de conteúdo de websites (indexação).	Construção automática de taxonomias	*E-commerce*
Zakharova, 2020	Experimental	Regular	Hierarquia e categorização para compreensão e organização semântica de serviços *web*	Estrutura hierárquica, categorização	Arquitetura de sistemas e serviços *web*
Zhang e Sun, 2022	Experimental	Muito boa	Taxonomia como elemento fundamental em métodos de estimativa de similaridade semântica, neste estudo denominada similaridade computacional.	Hierarquia da taxonomia como abordagem para medição da similaridade semântica em conjuntos de dados	Estatística

Estudo	Tipo de pesquisa	Avaliação da qualidade	Contribuição	Abordagem	Aplicabilidade
Zhang et al., 2021b	Experimental	Muito boa	Importância da construção de taxonomias como etapa fundamental na construção de KG	Taxonomia como etapa de construção de um grafo do conhecimento	E-commerce
Zhang et al., 2021c	Experimental	Muito boa	A taxonomia ajuda a avaliar a eficácia das técnicas de camuflagem para evitar a detecção de phishing baseada em navegador. A taxonomia possibilita ao CrawPish categorizar a evasão do lado do cliente com alta precisão.	Taxonomia construída manualmente e empregada na organização e compreensão do conhecimento de um domínio.	Prevenção de crimes cibernéticos
Zhang et al., 2021a	Experimental	Excelente	Introduziu o conceito de conclusão de taxonomia, que vai além da expansão ao encontrar pares de hiperônimos e hipônimos para novos conceitos, aumentando a precisão da conclusão da taxonomia.	Manutenção e curadoria automatizada de taxonomias	Linguística computacional; modelagem de dados
Zouaq e Martel, 2020	Experimental	Boa	No estudo os conceitos de *ontology learning*, *taxonomy induction* e *taxonomy extraction* são associados como quase sinônimos ou variações uns dos outros. Futuros trabalhos dos autores desejam demonstrar como construir uma taxonomia para identificação de erros, inconsistências ou incompletudes de dados.	Construção automática de taxonomias	Construção de taxonomias e manutenção de ontologias.

Fonte: elaborado pelo autor (2023)

ANEXO 1 - CASP Qualitative Checklist (2018)

CASP
Critical Appraisal Skills Programme

Paper for appraisal and reference: .....

Section A: Are the results valid?

1. Was there a clear statement of the aims of the research?

Yes ☐ Can't Tell ☐ No ☐

HINT: Consider
- what was the goal of the research
- why it was thought important
- its relevance

Comments:

Na matriz de análise: coluna OBJETIVO e item da coluna METODOLOGIA (a) objetivos

2. Is a qualitative methodology appropriate?

Yes ☐ Can't Tell ☐ No ☐

HINT: Consider
- If the research seeks to interpret or illuminate the actions and/or subjective experiences of research participants
- Is qualitative research the right methodology for addressing the research goal

Comments:

Na matriz de análise: coluna METODOLOGIA, item (b) métodos

Is it worth continuing?

3. Was the research design appropriate to address the aims of the research?

Yes ☐ Can't Tell ☐ No ☐

HINT: Consider
- if the researcher has justified the research design (e.g. have they discussed how they decided which method to use)

Comments:

Na matriz de análise: coluna METODOLOGIA, item (b) métodos

CASP
Critical Appraisal Skills Programme

4. Was the recruitment strategy appropriate to the aims of the research?

Yes ☐
Can't Tell ☐
No ☐

HINT: Consider
- If the researcher has explained how the participants were selected
- If they explained why the participants they selected were the most appropriate to provide access to the type of knowledge sought by the study
- If there are any discussions around recruitment (e.g. why some people chose not to take part)

Comments:

Não se aplica aos estudos encontrados

5. Was the data collected in a way that addressed the research issue?

Yes ☐
Can't Tell ☐
No ☐

HINT: Consider
- If the setting for the data collection was justified
- If it is clear how data were collected (e.g. focus group, semi-structured interview etc.)
- If the researcher has justified the methods chosen
- If the researcher has made the methods explicit (e.g. for interview method, is there an indication of how interviews are conducted, or did they use a topic guide)
- If methods were modified during the study. If so, has the researcher explained how and why
- If the form of data is clear (e.g. tape recordings, video material, notes etc.)
- If the researcher has discussed saturation of data

Comments:

Na matriz de análise: coluna METODOLOGIA, itens (c) transparência nos relatos/ética e (d) análise de dados/medidas/instrumentos

CASP
Critical Appraisal Skills Programme

6. Has the relationship between researcher and participants been adequately considered?

Yes ☐
Can't Tell ☐
No ☐

HINT: Consider
- If the researcher critically examined their own role, potential bias and influence during (a) formulation of the research questions (b) data collection, including sample recruitment and choice of location
- How the researcher responded to events during the study and whether they considered the implications of any changes in the research design

Comments:

Não se aplica aos estudos encontrados

Section B: What are the results?

7. Have ethical issues been taken into consideration?

Yes ☐
Can't Tell ☐
No ☐

HINT: Consider
- If there are sufficient details of how the research was explained to participants for the reader to assess whether ethical standards were maintained
- If the researcher has discussed issues raised by the study (e.g. issues around informed consent or confidentiality or how they have handled the effects of the study on the participants during and after the study)
- If approval has been sought from the ethics committee

Comments:

Na matriz de análise: coluna METODOLOGIA, itens (c) transparência nos relatos/ética e (e) vieses

CASP
Critical Appraisal Skills Programme

8. Was the data analysis sufficiently rigorous?

Yes ☐
Can't Tell ☐
No ☐

HINT: Consider
- If there is an in-depth description of the analysis process
- If thematic analysis is used. If so, is it clear how the categories/themes were derived from the data
- Whether the researcher explains how the data presented were selected from the original sample to demonstrate the analysis process
- If sufficient data are presented to support the findings
- To what extent contradictory data are taken into account
- Whether the researcher critically examined their own role, potential bias and influence during analysis and selection of data for presentation

Comments:

Na matriz de análise: coluna METODOLOGIA, item (d) análise de dados/medidas/instrumentos

9. Is there a clear statement of findings?

Yes ☐
Can't Tell ☐
No ☐

HINT: Consider whether
- If the findings are explicit
- If there is adequate discussion of the evidence both for and against the researcher's arguments
- If the researcher has discussed the credibility of their findings (e.g. triangulation, respondent validation, more than one analyst)
- If the findings are discussed in relation to the original research question

Comments:

Na matriz de análise: coluna METODOLOGIA, item (f) achados

CASP
Critical Appraisal Skills Programme

Section C: Will the results help locally?

10. How valuable is the research?	HINT: Consider • If the researcher discusses the contribution the study makes to existing knowledge or understanding (e.g. do they consider the findings in relation to current practice or policy, or relevant research-based literature • If they identify new areas where research is necessary • If the researchers have discussed whether or how the findings can be transferred to other populations or considered other ways the research may be used

Comments:

Na matriz de análise: coluna METODOLOGIA, item (g) referências; coluna APLICABILIDADE; coluna CONTRIBUIÇÃO; e AVALIAÇÃO DE QUALIDADE DO ESTUDO