UNIVERSIDADE DE BRASÍLIA
FACULDADE DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

JEAN CARLOS BORGES BRITO

Organização da Informação: uma proposta de *framework* genérico para geração automática de assuntos, indexação e busca facetada em repositórios digitais

Brasília/DF
2023

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

JEAN CARLOS BORGES BRITO

Organização da Informação: uma proposta de *framework* genérico para geração automática de assuntos, indexação e busca facetada em repositórios digitais

Tese submetida ao Programa de Pós-Graduação em Ciência da Informação da Universidade de Brasília, como requisito parcial para obtenção do título de doutor em Ciência da Informação.

Área de Concentração: Organização da Informação

Orientador: Prof. Dr. Dalton Lopes Martins

Brasília/DF
2023

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

BB862o

Brito, Jean Carlos Borges
Organização da Informação: uma proposta de framework genérico para geração automática de assuntos, indexação e busca facetada em repositórios digitais / Jean Carlos Borges Brito; orientador Dalton Lopes Martins. -- Brasília, 2023.
138 p.

Tese(Doutorado em Ciência da Informação) -- Universidade de Brasília, 2023.

1. Repositórios Digitais. 2. Geração Automática e Semiautomática. 3. Metadados. 4. Tesauros. 5. Indexação. I. Martins, Dalton Lopes, orient. II. Título.

UNIVERSIDADE DE BRASÍLIA

PROGRAMA DE PÓS GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Ata Nº: 30

Aos vinte e um dias do mês de novembro do ano de dois mil e vinte e três, instalou-se a banca examinadora de Tese de Doutorado do aluno Jean Carlos Borges Brito,matrícula 20/0001736. A banca examinadora foi composta pelos professores Dr. Marcio de Carvalho Victorino, membro titular interno, PPGCINF/UnB, Dra. Daniela Lucas da Silva Lemos, membro externo, Universidade Federal do Espírito Santo (UFES), Dr. Henrique Monteiro Cristovão, membro externo, Universidade Federal do Espírito Santo (UFES) e Dr. Dalton Lopes Martins, PPGCINF/UnB, orientador/presidente. O discente apresentou o trabalho intitulado "Organização da Informação: uma proposta de framework genérico para geração automática de assuntos, indexação e busca facetada em repositórios digitais".

Concluída a exposição, procedeu-se a arguição do(a) candidato(a), e após as considerações dos examinadores o resultado da avaliação do trabalho foi:

( ) Pela aprovação do trabalho;

(X) Pela aprovação do trabalho, com revisão de forma, indicando o prazo de até 30 dias para apresentação definitiva do trabalho revisado;

( ) Pela reformulação do trabalho, indicando o prazo de (Nº DE MESES) para nova versão;

( ) Pela reprovação do trabalho, conforme as normas vigentes na Universidade de Brasília.

Conforme os Artigos 34, 39 e 40 da Resolução 0080/2021 - CEPE, o(a) candidato(a) não terá o título se não cumprir as exigências acima.

Dr. Dalton Lopes Martins (PPGCINF/UnB)

(PRESIDENTE)

Dr. Marcio de Carvalho Victorino (PPGCINF/UnB)

(MEMBRO TITULAR INTERNO)

Dra. Daniela Lucas da Silva Lemos (UFES)

(MEMBRO EXTERNO)

Dr. Henrique Monteiro Cristovão (UFES)

(MEMBRO EXTERNO)

Dr. João de Melo Maricato (PPGCINF/UnB)

(SUPLENTE)

Jean Carlos Borges Brito

(Doutorando)

Documento assinado eletronicamente por **Dalton Lopes Martins**, **Professor(a) de Magistério Superior da Faculdade de Ciência da Informação**, em 12/12/2023, às 11:18, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Marcio de Carvalho Victorino**, **Professor(a) de Magistério Superior da Faculdade de Ciência da Informação**, em 21/12/2023, às 21:04, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Daniela Lucas da Silva Lemos**, **Usuário Externo**, em 23/12/2023, às 11:00, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Jean Carlos Borges Brito**, **Usuário Externo**, em 26/12/2023, às 19:01, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Henrique Monteiro Cristovão**, **Usuário Externo**, em 03/01/2024, às 18:13, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **10445979** e o código CRC **F267C156**.

Referência: Processo nº 23106.121801/2023-20 SEI nº 10445979

Universidade de Brasília

Centro de custo: Colegiado da Pós-Graduação

Para: DPG/DIRPG

Prezados,

Informamos que o discente Jean Carlos Borges Brito apresentou a revisão de forma e o trabalho foi aprovado.

Favor considerar o termo de autorização - Teses e dissertações 10753610 em correção ao termo de autorização - Teses e dissertações (SEI nº 10445981).

Atenciosamente,

Em 21/08/2023.

Documento assinado eletronicamente por **Clovis Carvalho Britto**, **Coordenador(a) da Pós-Graduação da Faculdade de Ciência da Informação**, em 29/12/2023, às 20:55, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Dalton Lopes Martins**, **Professor(a) de Magistério Superior da Faculdade de Ciência da Informação**, em 03/01/2024, às 09:32, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **10756847** e o código CRC **464DD40C**.

Referência: Processo nº 23106.121801/2023-20 SEI nº 10756847

À minha mãe Avane, minha esposa Kelly Ferrer, meus filhos Daniel Felipe e Eduardo Richard, à minha irmã Viviane Isene, sobrinho Christian Isene e minha tia Eulália. Esses seres humanos foram os meus laços de amor e me deram inspiração dia após dia para execução dessa pesquisa.

AGRADECIMENTOS

Agradeço

...a todos os meus mestres desde o jardim de infância, do 1º e 2º graus, da Graduação, Especializações, Mestrado e Doutorado e demais professores de cursos realizados, pela transmissão do conhecimento e incentivo para aprender cada vez mais.

...às instituições que tive o privilégio de estudar: Centro de Ensino Fundamental 19 (Taguatinga); Centro de Educação Infantil 05 (Taguatinga); Escola Classe 03 (Ceilândia Norte); Centro de Ensino Fundamental 16 (Ceilândia Norte); Centro de Ensino Médio 02 (Ceilândia Norte); Centro Social Cantinho do Girassol (Ceilândia); SENAC-Ceilândia; Faculdade Cenecista de Brasília – Ceilândia; Faculdade Santa Terezinha – Taguatinga; Universidade Católica de Brasília – Asa Norte; Instituto de Gestão, Economia e Políticas Públicas (IGEPP); Escola Superior de Defesa – ESD; Universidade de Brasília – UnB; Teatro Nacional Cláudio Santoro; Escola Nacional de Administração Pública – ENAP; dentre diversos outros estabelecimentos que agregaram conhecimento e experiência.

...aos empreendimentos e instituições que tive a oportunidade de trabalhar: Supermercado Mais Barato, Hot Dog JV, Sacaria Borges, Força Aérea Brasileira (SERENS-6, BINFA e COMGEP), Ministério Público Militar, Transoft Informática, Operador Nacional do Sistema Elétrico, Serviço Nacional de Aprendizagem do Cooperativismo, Dataprev, Ibama, Centro Universitário UniProjeção e Agência Espacial Brasileira.

...ao Prof. Dr. Dalton Lopes Martins (UnB) que orienta essa pesquisa e coordena projetos sobre os seguintes temas: objetos e repositórios digitais, acervos digitais e estratégias de interoperabilidade de sistemas de informação, dados abertos ligados, ciência de dados e aprendizagem de máquina com ênfase na análise de objetos digitais.

...à Biblioteca Nacional da Finlândia por disponibilizar a ferramenta Annif em formato aberto e possibilitar sua utilização com diversas técnicas de aprendizagem para indexação automática de metadados. Especial agradecimento ao Sr. Osma Suominen e sua equipe.

...à população do Brasil, pois no período de dedicação para este doutorado passamos por uma pandemia de COVID-19, ocasionando aulas remotas, *lockdown*, dificuldades de interação e o mais dramático: "a perda de 708.021 mil vidas de brasileiros", conforme consulta ao site <https://covid.saude.gov.br/> em 08/12/2023. Esta pesquisa é dedicada a essas pessoas que travaram suas batalhas pessoais para continuar vivendo, mas que infelizmente perderam suas vidas, deixando familiares e amigos de luto.

...à todas as pessoas que conversei e interagi durante a minha trajetória de vida. Todas essas pessoas influenciaram o ser humano que sou hoje.

...**à Deus**, pois tudo o que faço é exclusivamente para a **honra e glória dele**.

O conhecimento e a informação são os recursos estratégicos para o desenvolvimento de qualquer país. Os portadores desses recursos são as pessoas.

Peter Drucker

"*Knowledge is power. Information is liberating. Education is the premise of progress, in every society, in every family*."

Kofi Annan

BRITO, Jean Carlos Borges. Percurso **Organização da Informação: uma proposta de *framework* genérico para geração automática de assuntos, indexação e busca facetada em repositórios digitais**. Brasília, 2023. Tese (Doutorado em Ciência da Informação) – Programa de Pós-graduação em Ciência da Informação, Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, 2023, 138f.

RESUMO

Introdução. Com o aumento exponencial de dados, devido a sua característica digital, amplia-se também os problemas de ausência ou a discrepância de metadados cadastrados, tornando um trabalho oneroso e árduo para um ser humano corrigi-los manualmente. Surge a necessidade de investigações para melhorar a organização e facilitar recuperação da informação. Nesse contexto, a aplicação de inteligência artificial através da aprendizagem de máquina, utilizando ferramentas automáticas e semiautomáticas de forma complementar para coleta de metadados e geração de assuntos, propicia a melhoria de sua indexação e buscas nos repositórios digitais.
Objetivo. Propor um *framework* genérico com um conjunto de atividades e técnicas para executar a geração e indexação automática de assuntos em um repositório digital, visando a organização e a recuperação da informação.
Métodos. Revisão Sistemática da Literatura, estudando sobre o tema de geração automática e semiautomática de metadados, suas ferramentas, técnicas, características e funções. Realizou-se pesquisa exploratória em bases de dados científicas da Ciência da Informação, selecionando periódicos específicos para a avaliação de acordo com as principais listas de classificação. Utilizou-se método misto na análise dos dados, com abordagens quantitativas e qualitativas, sendo definido um protocolo rigoroso de revisão. Identificou-se ferramentas para auxiliar a pesquisa aplicada, através da sua customização e uso em conjunto de vários algoritmos de aprendizagem de máquina que auxiliassem no processo de geração automática de assuntos. Ao final, realizou-se um estudo de caso aplicado para o "modelo de pesquisa".
Resultados. Conclui-se que as técnicas de geração automática de metadados auxiliam na sugestão de assuntos para documentos robustos como uma tese e dissertação, ampliando o quantitativo de descritores, de modo a facilitar a configuração de taxonomias, filtros e facetas. Esse trabalho propôs o *framework* genérico validado pelo modelo de pesquisa, através do estudo de caso aplicado. Esse *framework* pode ser adequado e aplicado em qualquer área do conhecimento, com intuito de melhorar e facilitar a busca e a recuperação da informação nos repositórios digitais pelos usuários e gestores desses acervos.

Descritores: Repositórios Digitais. Geração Automática e Semiautomática. Metadados. Tesauros. Corpus de conhecimento. Indexação.

BRITO, Jean Carlos Borges. **Information Organization: a proposal for a generic framework for automatic generation of subjects, indexing and faceted search in digital repositories**. Brasília, 2023. Thesis (Doctorate in Information Science) – Postgraduate Program in Information Science, Faculty of Information Science, University of Brasília, Brasília, 2023, 138f.

ABSTRACT

Introduction. With the exponential increase in data, due to its digital nature, problems with the absence or discrepancy of registered metadata also increase, making it an expensive and arduous job for a human being to correct them manually. There is a need for investigations to improve the organization and facilitate information retrieval. In this context, the application of artificial intelligence through machine learning, using automatic and semi-automatic tools in a complementary way to collect metadata and generate subjects, improves indexing and searches in digital repositories.
Goal. Propose a generic framework with a set of activities and techniques to automatically generate and index subjects in a digital repository, aiming at organizing and retrieving information.
Methods. Systematic Literature Review, studying the subject of automatic and semi-automatic generation of metadata, its tools, techniques, characteristics and functions. Exploratory research was carried out in scientific databases of Information Science, selecting specific journals for evaluation according to the main classification lists. A mixed method was used for data analysis, with quantitative and qualitative approaches, with a strict review protocol being defined. Tools were identified to help applied research, through their customization and joint use of several machine learning algorithms that would help in the process of automatic subject generation. At the end, a case study applied to the "research model" was carried out.
Results. It is concluded that the automatic generation of metadata techniques help in suggesting subjects for robust documents such as a thesis and dissertation, expanding the number of descriptors, in order to facilitate the configuration of taxonomies, filters and facets. This work proposed the generic framework validated by the research model, through the applied case study. This framework can be adapted and applied in any area of knowledge, with the aim of improving and facilitating the search and retrieval of information in digital repositories by users and managers of these collections.

Descriptors: Digital Repositories. Automatic and Semiautomatic Generation. Metadata. Thesaurus. Corpus of knowledge. Indexing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURA E SIGLAS

API – *Application Programming Interface*

BRAPCI – Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação

CTDE – Câmara Técnica de Documentos Eletrônicos

CI – Ciência da Informação

CONARQ – Conselho Nacional de Arquivos

DC – *Dublin Core*

ENANCIB – Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação

ENAP – Escola Nacional de Administração Pública

FCI – Faculdade da Ciência da Informação

IDC – *International Data Corporation*

IBICT – Instituto Brasileiro de Informações em Ciência e Tecnologia

ISTA – *Information Science and Technology Abstracts*

GAM – Geração Automática de Metadados

GSAM – Geração Semiautomática de Metadados

LISA – *Library and Information Science Abstract*

LISTA – *Library, Information Science & Technology Abstracts*

OAI-PMH – *Open Archives Initiative Protocol for Metadata Harvesting*

OED – *Oxford English Dictionary*

ORI – Organização e recuperação da informação

PLN – Processamento de linguagem natural

PRISMA – *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*

RI – Recuperação da Informação

RiUNB – Repositório Institucional da Universidade de Brasília

RSL – Revisão Sistemática de Literatura

SKOS – *Simple Knowledge Organization System*

SRI – Sistemas de Recuperação da Informação

TSV – *Tab Separated Values*

TBCI – Tesauro Brasileiro em Ciência da Informação

TIC's – Tecnologias da Informação e da Comunicação

UNB – Universidade de Brasília

UEL – Universidade Estadual de Londrina

VCGE – Vocabulário Controlado de Governo Eletrônico

SUMÁRIO

1. INTRODUÇÃO

Etimologicamente, o termo informação vem do latim *informatio, onis* (delinear, conceber ideia), podemos entender como um ato de compreensão, valorizando o conteúdo e a forma, além de ser impactada durante o processo de comunicação. Shannon e Weaver (1949), discorrem que uma pessoa recebe informação, quando o que ela conhece se modifica, se altera. Wersig e Neveling (1975) realizam uma abordagem estrutural informacional, onde as estruturas da natureza, apreendidas ou não, constituem informação, concluindo que ela é independente da apreensão humana. Esses autores enfatizam ainda, que o termo informação é usado frequentemente, como sinônimo de mensagem. Entretanto, afirmam que somente o significado da mensagem é informação.

Capurro e Hjørland (2003) relatam que é comum considerar a informação como uma condição básica para o desenvolvimento econômico, juntamente com capital, trabalho e matéria-prima; mas o que torna a informação especialmente significativa no momento é a sua natureza digital.

A humanidade produz informações em volumes e variedades exponenciais diariamente que são armazenadas principalmente em repositórios digitais. Conteúdos físicos legados estão sendo convertidos e mantidos em bancos de dados e unidades de armazenamento lógico, acompanhando a transformação digital atual por qual passa vários países do mundo, abrangendo governos, empresas e sociedade mundial.

Estudo publicado por Reinsel, Gantz e Rydning (2018) para o *International Data Corporation* (IDC), prevê que o crescimento de dados aumentará de 45 *zettabytes* em 2019 para cerca de 175 *zettabytes* até 2025. Essa informação demonstra que em cinco anos, 6 bilhões de pessoas ou 75% da população mundial interagirão com dados todos os dias e cada pessoa conectada terá pelo menos uma interação com dados a cada 18 segundos. De acordo com esses autores, grande parte da economia atual depende de dados, sendo necessário melhorar sua captura, organização, gerenciamento e análise a partir de processos e tecnologias que aumentam a sua qualidade e permitam melhor a exploração de seu valor agregado, fornecendo informação necessária para tomada de decisão.

Nesse contexto, a utilização de abordagens de processos automatizados pode melhorar a eficiência das atividades de organização da informação nesse ambiente digital que está em constante crescimento, disponibilizando em tempo real aquilo que é feito tradicionalmente de forma manual.

Fogl (1979) enfatiza que o aumento da produtividade do trabalho é resultado da atividade cognitiva e avaliativa contida na informação. Na mesma linha de pensamento, Brascher e Café (2008) discorrem que o valor da informação em um sistema de informação depende da compreensão individual atribuído pelo receptor da informação, pois ele a adota segundo seus critérios e objetivos.

Um usuário potencial da informação é capaz de converter sua necessidade de informações em uma lista de referências para documentos armazenados e que contém informações úteis (MOOERS, 1951). Neste contexto, os metadados desempenham um papel essencial, pois descrevem os dados, facilitam sua compreensão e corroboram na eficácia da indexação e sua recuperação.

De acordo com Pomerantz (2015), metadado indica algo que está além dos dados, sendo uma declaração sobre esses dados. Haynes (2018) corrobora com o entendimento de que o metadado tem a função de facilitar o entendimento dos relacionamentos e evidenciar a utilidade das informações obtida dos dados.

Polfreman *et al*. (2008) discorrem que sem os metadados apropriados, os recursos permanecem ocultos e sem utilização, causando desperdício de investimento. Os autores também enfatizam que a baixa qualidade ou metadados inexistentes são igualmente eficazes para tornar os recursos inutilizáveis, pois sem ele um recurso é essencialmente invisível dentro de um repositório ou arquivo morto e, portanto, permanece desconhecido e inacessível.

Crystal e Land (2003) discorrem que para criar metadados para um milhão de documentos deveriam ser alocados 60 empregados/ano para realizar essa tarefa. É considerado um trabalho árduo, lento e caro se executado manualmente, continuam esses autores. Além disso, considerando que o conceito de documento e suas possibilidades de expressão midiática se expandem de forma significativa na era da *web*, torna-se proibitivo imaginar que a catalogação dos documentos seguirá continuamente sendo realizada apenas de forma manual.

Conforme Greenberg (2003), o incremento de metadados com qualidade fornece valor agregado ao conjunto de dados, além de melhorar sua classificação e busca. Os pesquisadores necessitam identificar métodos de produção de metadados mais eficientes e menos dispendiosos. Devido ao alto custo da inserção manual de metadados, é mister o fomento e incentivo em desenvolvimento de ferramentas que possam auxiliar a geração automática ou semiautomática de metadados, melhorando sua escalabilidade.

De acordo com Maratea, Petrosino e Manzo (2012), a geração automática de metadados (GAM) iniciou-se com a introdução de documentos digitais desde os anos de 1950 e diz respeito à sua indexação, abstração e classificação de forma automática.

Park e Brenza (2015) apresentam em seus estudos, ferramentas de geração semiautomática de metadados (GSAM), que dizem respeito ao uso de *softwares* para criação de registros de metadados com graus variados de supervisão por um especialista humano.

De acordo com Greenberg (2003), a geração automática de metadados, baseado no conhecimento sobre indexação automática – associação de termos a documentos –, é mais eficiente, possui menor custo e é mais consistente do que processos executados por seres humanos. Com a ascensão da internet, outra técnica muito importante para esse ambiente é a extração de metadados, pois é um método de geração automática e ocorre quando um algoritmo automaticamente extrai metadados do conteúdo de um recurso de informação exibido através de um navegador *web*. Mesmo com todas essas vantagens da GAM, a autora enfatiza que segundo alguns pesquisadores, os meios mais efetivos para criar metadados é integrar métodos automáticos e semiautomáticos, um complementando o outro.

Pesquisas voltadas para criação e integração de ferramentas de geração automática e semiautomática de metadados são muito importantes para fornecer auxílio às pessoas e profissionais em gerenciar quantidades e tipos cada vez maiores de dados e metadados dos recursos de informação.

Nesse contexto, o trabalho realizado executou o levantamento de ferramentas de geração automática e semiautomática de metadados para melhor compreensão de suas aplicações e, propõe, um *framework* genérico conceitual para geração automática de assuntos, indexação e busca facetada em repositórios digitais, propiciando uma melhor organização da informação e facilitando sua recuperação.

1.1. PROBLEMA

Devido ao aumento exponencial de informações nos repositórios digitais (REINSEL, GANTZ E RYDNING, 2018), surge o debate e a tentativa de propor soluções automáticas (MARATEA, PETROSINO e MANZO, 2012; VERBOGH et *al*., 2012; COSTA et *al*., 2013; YANG e PARK, 2018; AUDICHYA e SAINI, 2019) ou semiautomáticas de metadados (PARK e BRENZA, 2015) com a finalidade de organizar a informação; da mesma forma que também, aumentam as demandas por processos que forneçam uma melhor qualidade e

eficiência na recuperação dessa informação (SUOMINEN, 2019; LAPPALAINEN et *al*., 2021).

Os gestores de repositório digitais são responsáveis por uma quantidade enorme de metadados relacionados a diferentes tipos de documentos que geralmente são indexados por títulos, assuntos e descritores para que possam ser recuperados posteriormente (SUOMINEN, 2019). Entretanto, nem todos os usuários de sistemas de biblioteca e repositórios digitais executam a entrada correta e completa de metadados, o que dificulta a recuperação do objeto de informação. O processo manual de geração de metadados de documentos é um trabalho árduo, oneroso e dependendo do volume de dados é humanamente impossível de ser realizado.

A problemática identificada neste estudo está na dificuldade de encontrar na literatura um *framework* para organização e recuperação da informação que possa ser aplicado em um contexto geral e em qualquer área de conhecimento. Mesmo com o crescimento de soluções automáticas e semiautomáticas para geração de metadados, a variedade de técnicas e métodos disponíveis na literatura, corroboram na compreensão sobre os desafios para o desenvolvimento de um *framework* geral. Por isso, essa pesquisa busca provocar reflexões acerca das ferramentas de geração automática e semiautomática de metadados com o objetivo de melhorar a organização e recuperação da informação nos repositórios digitais. Diante das argumentações apresentadas, o problema de pesquisa foi delimitado com a seguinte questão: "**Como a técnica de geração automática/semiautomática de metadados pode apoiar usuários e gestores de repositórios digitais na melhoria da organização da informação, visando facilitar a busca e a recuperação da informação em seus acervos?**"

A partir da inquietação representada na pergunta acima, elaboraram-se os objetivos (geral e específicos) que possibilitassem uma relação com o problema de pesquisa, corroborando com as ações necessárias que delinearam as diversas atividades desse estudo.

1.2. OBJETIVO GERAL

Essa pesquisa tem por objetivo, propor um *framework* genérico com um conjunto de atividades e técnicas para executar a geração e indexação automática de assuntos em um repositório digital, visando a organização e a recuperação da informação.

1.3. OBJETIVOS ESPECÍFICOS

Os seguintes objetivos específicos contribuem ao objetivo geral:

- Identificar técnicas, tecnologias e metodologias de geração automática/semiautomática de metadados;
- Identificar para o estudo de caso, um repositório digital e caracterizar seus recursos atuais de organização da informação que possam ser melhorados para busca e recuperação pela aplicação de técnicas de geração automática/semiautomática de metadados;
- Identificar instrumentos de tratamento descritivo e/ou tratamento temático da informação que possam ser utilizados para melhorar a busca e recuperação da informação do repositório digital escolhido como estudo de caso;
- Analisar e identificar um conjunto de técnicas específicas de geração automática/semiautomática de metadados que possam ser aplicadas a dados reais de um repositório digital;
- Aplicar as técnicas escolhidas a dados reais, comparar resultados e analisar criticamente as contribuições e desafios em relação às possibilidades de melhoria dos recursos de busca e recuperação da informação;
- Identificar recomendações a gestores de repositórios digitais para a implementação de técnicas de geração automática/semiautomática de metadados.

1.4. JUSTIFICATIVA

A organização e a recuperação da informação nos repositórios digitais têm sido impactadas devido a diversos problemas relacionados aos metadados e a representação da informação nesses acervos. Alguns exemplos de pesquisa relatando essa problemática podem ser observados a seguir.

Café e Muñoz (2016) investigaram a usabilidade do Repositório Institucional da Universidade de Brasília – RiUnB no processo de recuperação da informação e interação com usuários pós-graduandos dessa instituição. Observaram que os usuários da RiUnB enfrentaram dificuldades na recuperação da informação, devido à falta de conteúdo ou problemas com o sistema de busca, o que gerou descontentamento desse público com a solução disponibilizada.

Santos (2017) pesquisou a representação temática da informação no contexto dos repositórios digitais, especificamente dos conteúdos dos documentos na Biblioteca Digital de Monografias da Universidade Federal do Rio Grande do Norte (BDM/UFRN). A autora destacou que a representação do campo assunto, ocorreu de forma livre e sem padronização, tais como: emprego de termos com polissemia, erros ortográficos, descritores abrangentes e com múltiplos sentidos, abreviações e uso de siglas, o que implicou diretamente no processo de representação e recuperação dos documentos em ambiente digital.

Sales, Rocha e Cavalcanti (2017) identificaram metadados obrigatórios dentro do repositório institucional do Instituto de Engenharia Nuclear (IEN) compostos por recursos informacionais registrados nos mais diversos formatos e mídias. Entretanto, problemas como a presença de sinônimos, erros de digitação ou diferenciação de letras maiúsculas e minúsculas, provocavam erros quando os objetos digitais necessitavam ser recuperados.

Diante do contexto de ocorrências de variados tipos de problemas apresentados nesses repositórios digitais, percebe-se que várias organizações e bibliotecas possuem a necessidade claramente identificada de automatizar a geração de metadados através de ferramentas práticas, integradas aos sistemas de informação, que ofereçam boa qualidade neste processo (SUOMINEN, 2019; LAPPALAINEN et *al*., 2021). O intuito é propiciar a recuperação mais assertiva da documentação armazenada. Essa atividade, se executada manualmente, é cara, lenta e dispendiosa, além de onerar a administração e organização pelo gestor de repositórios digitais.

Pesquisas demonstram a necessidade de implementação de novas tecnologias para geração automática de metadados (MARATEA, PETROSINO e MANZO, 2012; VERBOGH et *al*., 2012; COSTA et *al*., 2013; YANG e PARK, 2018; AUDICHYA e SAINI, 2019), geração semiautomática de metadados (PARK e BRENZA, 2015) e uso conjunto destas ferramentas (GREENBERG, 2003).

Conforme será desenvolvido na fundamentação teórica, a geração automática e semiautomática de metadados é uma atividade essencial para auxiliar os gestores de repositórios digitais nas atividades de organização e recuperação da informação em acervos com grande volume e variedade de dados. A pesquisa acadêmica e seu pragmatismo corroboram para análise do problema desta investigação. Tendo em vista a lacuna de soluções tecnológicas que mitiguem os problemas relatados, além da ausência de um *framework* para tratar de um contexto geral.

De acordo com Kivunja (2018), *framework* pode ser compreendido como uma estrutura e pode ser de dois tipos: o primeiro, denominado conceitual, existe quando o

framework está associado a pensamentos, planos e prática subjacentes, além da implementação de todo o projeto de pesquisa; e o segundo, denominado de teórico, abarca a base acadêmica para toda a compreensão do significado contido nos dados, fornecendo oportunidade de discussão das descobertas de forma clara, fundamentado nas teorias existentes. Nesta investigação está sendo utilizado tanto a estrutura conceitual quanto teórica.

1.5. RESULTADOS ESPERADOS

- Compreensão de como as técnicas, tecnologias e metodologias de geração automática/semiautomática de metadados podem contribuir para organização e recuperação da informação em repositórios digitais;
- Apresentar bases conceituais dos referenciais teóricos que permitam o desenvolvimento de uma estrutura de implantação e sustentação de um *framework* genérico para auxiliar a organização e recuperação da informação;
- Validar o *framework* proposto, através de um estudo de caso com um Repositório Digital real, melhorando a organização e recuperação da informação, através da: construção de vocabulário controlado, criação das taxonomias, geração automática de assuntos, indexação em repositório digital e configuração de buscas facetadas;
- Exemplificar o uso de técnicas de aprendizagem de máquina com o uso de algoritmos léxicos e associativos para geração automática de assuntos, treinando o modelo, capaz de aprender os termos de uma área de conhecimento e realizar o tratamento temático da informação, auxiliando na sua recuperação;
- Analisar resultados da aplicação do *framework* proposto, realizar os ajustes necessários e apresentar a estrutura de implantação e sustentação do modelo.

1.6. ESTRUTURAÇÃO DA PESQUISA – TESE

O Capítulo 1 consiste na apresentação e introdução do trabalho, a questão de pesquisa, seus objetivos (geral e específicos). Discorre também, sobre a justificativa da pesquisa, os resultados esperados e a estruturação da pesquisa, tecendo um panorama geral da tese.

O Capítulo 2 expõe os fundamentos teórico-metodológicos, realizando Revisão Sistemática da Literatura – RSL (específica sobre ferramentas de geração automática e semiautomática de metadados) e Revisão de Literatura de Conveniência – RLC. Esta última, traz reflexões sobre os conceitos de informação, ciência da informação, organização, representação e recuperação da informação, repositórios digitais; a definição de vocabulário controlado, listas de termos e assuntos, arquivos de autoridade, taxonomias e tesauros; conceitos sobre a inteligência artificial e corpus de conhecimento; abordagens de ferramentas analisadas à posteriori da RSL; a solução ANNIF; e o *framework* genérico proposto.

O Capítulo 3 se dedica à metodologia de pesquisa para direcionar as atividades a serem executadas. São explicadas sua caracterização, os procedimentos da pesquisa-ação adotada e suas etapas: inicialmente, a seleção das atividades e soluções tecnológicas utilizadas na pesquisa, baseado nos fundamentos teórico-metodológicos do Capítulo 2; e a aplicação de um estudo de caso.

O Capítulo 4 demonstra a análise e a discussão dos resultados. São descritos os resultados do processo de pesquisa relacionados à aplicação da revisão sistemática e da validação do *framework* genérico através do estudo de caso. A análise se faz em ordem cronológica de ocorrência dos resultados obtidos.

O Capítulo 5 remete aos aprendizados e conclusões, que direcionam para possibilidades futuras de aprimoramento do *framework* e utilização de outras ferramentas para coleta, geração automática de assuntos e repositório digital, além de outras fontes e origens de dados para compor o vocabulário controlado e taxonomias. Nesse capítulo, enfatiza-se o atendimento dos objetivos (geral e específicos) propostos para a tese; o problema de pesquisa; as contribuições para a ciência da informação; proposições de pesquisas futuras; e as considerações finais.

2. FUNDAMENTOS TEÓRICO-METODOLÓGICOS

De acordo com Pettigrew e McKechnie (2001), a seriedade e a respeitabilidade de uma pesquisa são alcançadas através da teoria, pois ela auxilia na organização e comunicação dos dados, além de simplificar sua complexidade. Esses autores, também discorrem que as teorias inspiram e orientam as ações práticas, afirmando que elas são uma construção mental e uma marca da maturidade de sua disciplina acadêmica.

O *Oxford English Dictionary* conceitua "teoria" como um conjunto de ideias que explicam algo, fundamentados em princípios gerais, independente dos fatos e fenômenos a serem explicados.

Baseado nesses argumentos, este capítulo aborda e discute os pressupostos teóricos-metodológicos, executando as seguintes atividades:

- Desenvolvimento de uma Revisão Sistemática da Literatura (RSL), focando na identificação do estado da arte das ferramentas automáticas e semiautomáticas de geração de metadados;

- Elaboração de uma Revisão de Literatura por Conveniência (RLC), relacionada com a fundamentação sobre a informação, ciência da informação, a organização, representação e recuperação da informação, além de análise à posteriori de outras ferramentas de geração automática e semiautomáticas de metadados identificadas na literatura.

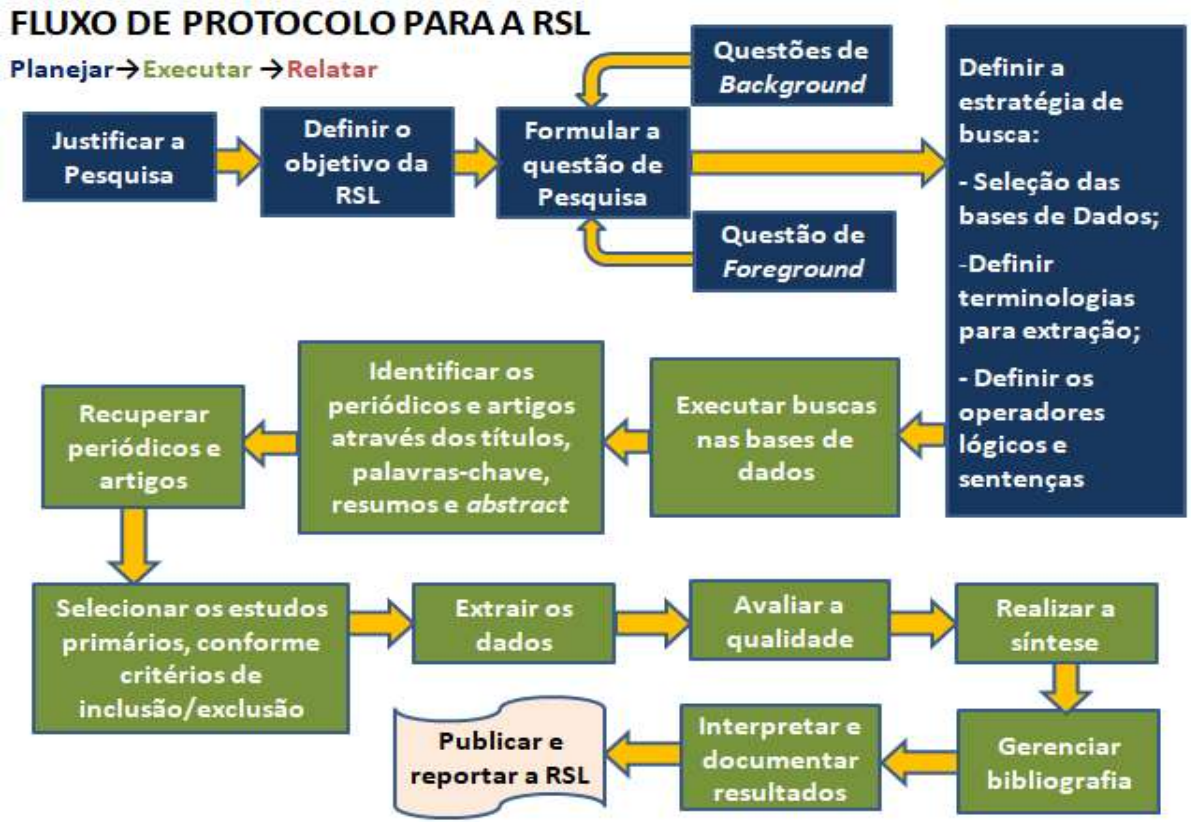
2.1. REVISÃO SISTEMÁTICA DA LITERATURA – RSL

A RSL consiste em um método de pesquisa científica através de um processo rigoroso e transparente com intuito de identificar, selecionar, obter dados, análise e descrições importantes de uma determinada investigação (FERENHOF e FERNANDES, 2016).

Galvão e Ricarte (2019) discorrem que a RSL se preocupa com a reprodutibilidade do método científico por outros pesquisadores, ou seja, promove a descrição explícita de cada passo do processo, demonstrando todas as atividades realizadas.

Neste estudo foi desenvolvido um protocolo de revisão, conforme Figura 1, onde foram executadas atividades alocadas em três fases, conforme proposta descrita por Kitchenham (2004): Planejar, executar e relatar a RSL.

Figura 1 – Etapas da revisão sistemática da literatura.



Fonte: o Autor, adaptado de Kitchenham (2004)

2.1.1. Planejar a RSL

O ponto de partida para a RSL é a delimitação da questão orientadora, onde é definida a população ou o problema, a intervenção, comparação e o resultado (GALVÃO e RICARTE, 2019). Inicialmente, elencaram-se questões de *background* para fornecer a compreensão básica e conceitual do tema. Em seguida, para melhor definição do escopo, estabeleceram-se questões de *foreground*, conforme pode ser visualizado através da Tabela 1:

Tabela 1 - Questões de *background* e *foreground*.

QUESTÕES DE *BACKGROUND*	QUESTÕES DE *FOREGROUND*
O que são metadados? Para que servem? O que é geração automática e semiautomática de metadados?	Quais técnicas, características, funções e ferramentas de geração automática e semiautomática de metadados?

Fonte: Elaborado pelo autor (2021).

Com o entendimento dos conceitos e da abrangência, formulou-se a seguinte questão orientadora para a investigação da RSL: “Quais as aplicabilidades e limitações das

ferramentas de geração automática e semiautomática de metadados para o gestor de repositórios digitais? ”

Da questão orientadora elaborada, podem-se evidenciar os seguintes componentes:

Tabela 2 – Descrição e componentes da pergunta.

DESCRIÇÃO	COMPONENTES DA PERGUNTA
População	Gestor de repositórios digitais
Intervenção	Geração automática e semiautomática de metadados
Comparação	Ferramentas, técnicas, características e funções.
Desfecho	Aplicabilidades atuais das ferramentas

Fonte: Elaborado pelo autor (2021).

Após a delimitação da questão orientadora, consultou-se profissional especializado da área de biblioteconomia da Universidade de Brasília para sugestão de bases consolidadas no contexto da Ciência da Informação, sendo elencadas:

- Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI). Possui indexação de artigos publicados nas revistas científicas e profissionais das áreas desde 1972 até o momento atual;

- *Library and Information Science Abstract* (LISA). Abrange a literatura internacional na área de Ciência da Informação desde 1969;

- *Library, Information Science & Technology Abstracts* (LISTA). Abrange a literatura internacional nas áreas de Ciência e de Tecnologia da Informação desde meados da década de 1960;

- *Emerald Publishing Limited*. Abrange revistas e livros acadêmicos nas áreas de administração, negócios, educação, estudos de bibliotecas, assistência médica e engenharia, desde 1967;

- *Information Science and Technology Abstracts* (ISTA)*;*

- *Wiley Online Library*;

- *Web of Science*. Plataforma que realiza indexação controlada em diversas bases multidisciplinares com mais de 100 anos de conteúdo indexado; e

- Scopus. Compreende várias áreas do conhecimento, incluindo: análise bibliométrica, história, educação, psicologia, direito, religião, linguística e literatura;

Foram definidos os seguintes termos para compor a expressão de busca em português e inglês, no singular e plural: Geração automática de metadado, *Automatic metadata generation*, Geração semiautomática de metadado, *Semi-automatic metadata generation*, Ferramenta, *Tool*, Técnica, *Technique*, Característica, *Feature*, Função, *Function*, Aplicação e *Application*.

Elaborou-se as sentenças de buscas utilizando os seguintes operadores lógicos:

- Sentença em Português: ((("Geração automática de metadado") *OR* ("Geração semiautomática de metadado")) *AND* (Ferramenta *OR* Técnica *OR* Característica *OR* Função *OR* Aplicação));

- Sentença em Inglês: ((("*Automatic metadata generation*") *OR* ("*Semi-automatic metadata generation*")) *AND* (*Tool OR Technique OR Feature OR Function OR Applications*)).

Realizou-se atividades de pré-testes nas bases científicas para verificar se as sentenças deveriam passar por um processo de readequação, o que foi confirmado. Algumas bases não retornaram informações, sendo executadas alterações nas sentenças elaboradas, conforme orientação de busca/ajuda do próprio periódico e revista. Essa documentação pode ser consultada em <https://cutt.ly/pWiqjrN>.

Importante salientar que a escolha do termo “geração automática de metadado” e “geração semiautomática de metadado” ao invés de “indexação automática e semiautomática de metadado” se deu por questões de abordagem mais abrangente, descritiva e temática que se queria obter em relação a identificação das técnicas, tecnologias e metodologias das ferramentas pesquisadas. Concluído o pré-teste, finalizou-se a fase de planejamento e após a aprovação por especialista, passou-se a etapa de execução da RSL.

2.1.2. Executar a RSL

Nesta etapa executou-se buscas nas bases científicas, utilizando os termos, sentenças e os operadores lógicos definidos na etapa de planejamento. Ao acessar a página de cada periódico, realizou-se o preenchimento dos filtros de busca de forma a delimitar a recuperação da informação. Na Tabela 3, demonstra-se a aplicação dos seguintes critérios de inclusão e exclusão para obter as publicações:

Tabela 3 - Critérios de inclusão e exclusão.

CRITÉRIOS DE INCLUSÃO	CRITÉRIOS DE EXCLUSÃO
Trabalhos científicos publicados entre os anos de 2010 e 2020	Trabalhos científicos publicados antes do ano de 2010
Descrição de estudo de caso, experimentos ou *survey*	Título do trabalho não condizente com a proposta do projeto
Resumo ou *abstract* condizente com a proposta de pesquisa	Resumo ou *abstract* com fuga ao tema proposto na pesquisa
Artigos e periódicos	Publicações sem cunho científico
Publicações em inglês e português com disponibilidade completa e suporte em meio eletrônico	Disponibilização de partes da pesquisa, textos incompletos

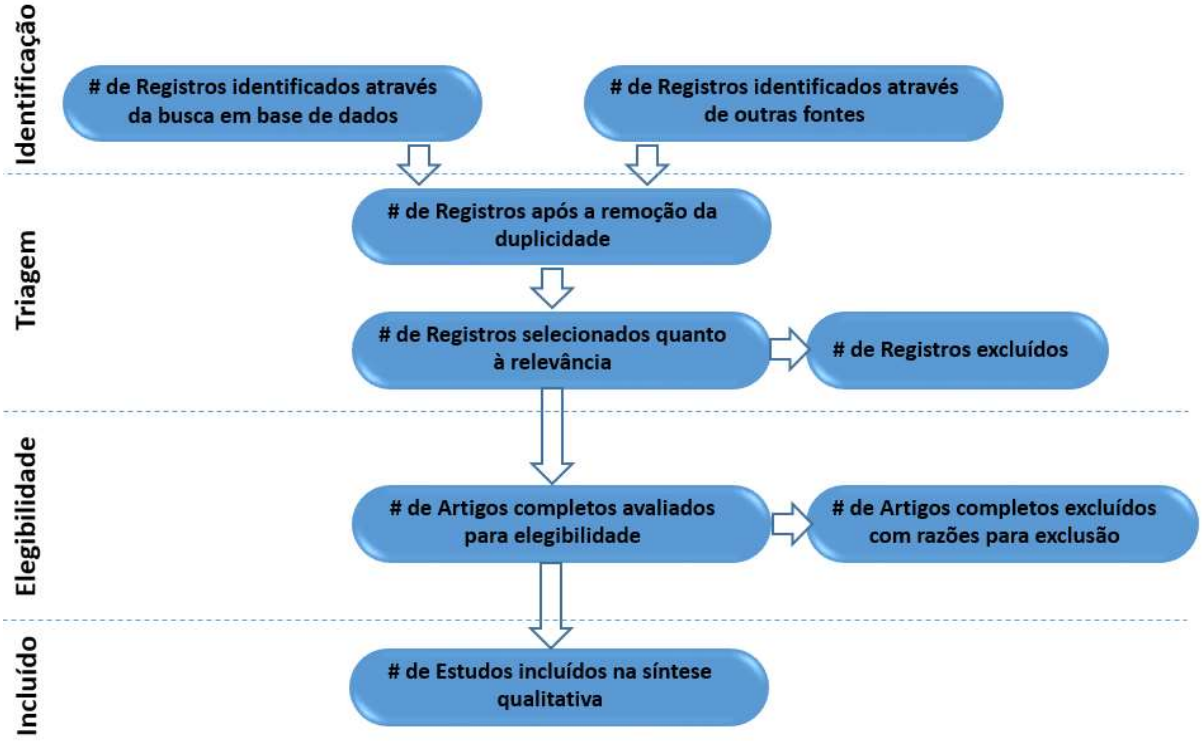
Fonte: Elaborado pelo autor (2021).

Duas bases de dados não retornaram resultado com as sentenças definidas na etapa de planejamento, sendo elas a BRAPCI e *Wiley Online Library*. Resolveu-se desmembrar as sentenças compostas em termos simples de busca, mas obteve-se zero resultado na recuperação de informação nesses dois repositórios. Interessante comentar que a BRAPCI por ser uma base referencial em Ciência da Informação com artigos indexados desde 1972, não possui investigações relacionadas com o tema geração automática e semiautomática de metadados. Foi constatado que os filtros do formulário dessa plataforma apresentavam falhas. Ao utilizar qualquer termo de busca como teste, o critério de inclusão denominado: Trabalhos científicos publicados entre os anos de 2010 e 2020; retornava consultas desde 1972 e não no período selecionado. Vale registrar que essa atividade foi realizada no período de 6 a 16 de maio de 2020.

As outras bases de dados pesquisadas retornaram o total de 49 trabalhos científicos, utilizando as sentenças definidas no planejamento.

Para analisar os critérios mínimos de qualidade foi utilizado um conjunto de itens com base em evidências, baseado em um *framework* elaborado por Moher *et al.* (2009) denominado *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA), adaptado para esta RSL, obedecendo o seguinte fluxograma na Figura 2:

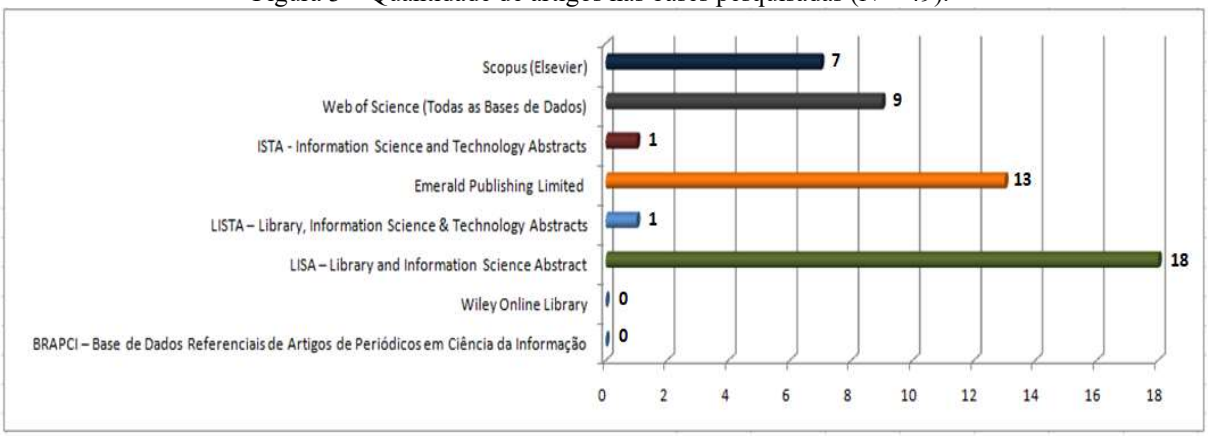
Figura 2 – Fluxo de seleção de artigos, baseado no *framework* PRISMA.



Fonte: Moher *et al.* (2009) com adaptações.

A primeira etapa (identificação) consistiu na documentação da quantidade de registros recuperados nas bases de dados selecionadas para a pesquisa, conforme a Figura 3:

Figura 3 – Quantidade de artigos nas bases pesquisadas (N = 49).



Fonte: Elaborado pelo autor (2021).

Na segunda etapa (triagem), executou-se a ação de verificar se o mesmo documento é indexado em várias bases de dados (ocorrendo duplicidade). Após a leitura do título e descrição dos registros identificados através da busca nas bases de dados, identificou-se 15 artigos duplicados, ou seja, na busca executada retornaram resultados em mais de uma base de dados

para o mesmo artigo. Em seguida, passou-se a leitura de resumo, *abstract* e palavras-chave, sendo excluídos 16 (dezesseis) artigos, devido não serem úteis para continuidade da pesquisa.

A terceira etapa (elegibilidade) consistiu na leitura completa de 18 (dezoito) estudos selecionados para análise. Entretanto, 6 (seis) artigos completos foram excluídos após uma leitura preliminar, pois o conteúdo dos trabalhos não correspondia ao objeto desta RSL.

Por fim, na quarta etapa (incluído) executou-se a ação de documentar a quantidade de registros incluídos na síntese qualitativa, que neste estudo culminou no total de 12 (doze) artigos.

A documentação deste processo pode ser acessada através do endereço <https://cutt.ly/tWiDg3k>. Todas as tabulações das publicações foram realizadas utilizando planilha eletrônica *Microsoft Excel*® para apoiar a organização deste estudo, categorizando as ferramentas descritas em cada artigo como: ferramenta de geração automática de metadados e ferramentas de geração semiautomática de metadados.

Executou-se *download* do arquivo e armazenado em repositório para realização da síntese. Utilizou-se a nomenclatura: Ano_de_Publicação – Autores – Título do Documento.

2.1.3. Contextualização sobre as técnicas para geração de metadados

Polfreman et *al.* (2008) elencam seis técnicas para a geração de metadados: colheita de *metatags*, extração de conteúdo, indexação ou classificação automática, mineração de textos e dados, *folksonomia* ou marcação social e geração automática de metadados extrínsecos. A seguir, será realizada uma explanação sobre cada uma dessas técnicas com a finalidade de auxiliar na classificação das ferramentas que serão descritas a partir dos artigos selecionados para a síntese qualitativa.

a) Colheita de *metatags*

A colheita de *metatags* é definida como processo de computação em que os valores para os campos de metadados são identificados e preenchidos por meio de um exame de *tags* de metadados em um documento ou anexado a ele, sendo a forma mais comum e simples de coleta através de *metatags* existentes em HTML e *tags* <meta>, onde várias soluções utilizam essa abordagem (PARK e BRENZA, 2015). A limitação se restringe na qualidade das *metatags* do documento, o que impacta sua efetividade. Esse tipo de ferramenta não é útil

para gerar valores automaticamente de metadados para propriedades que ainda não foram descritas, devendo recorrer a outras soluções para atender essa necessidade (POLFREMAN et *al*., 2008).

b) Extração de conteúdos

A extração de conteúdo é uma técnica que aborda a captura de palavras e frases do corpo de um recurso de informação e as utilizam para fornecimento de metadados estruturados (rótulos) com a finalidade de representar o objeto (GREENBERG, 2003). Ferramentas desenvolvidas para fazer uso dessa abordagem, utilizam uma mistura de técnicas baseadas em regras e estatísticas. Esses autores discorrem que técnicas baseadas em regras as utilizam de forma pré-determinadas para decidir como e onde as palavras e frases extraídas devem ser aplicadas às propriedades de metadados. As abordagens baseadas em estatísticas requerem treinamento com um conjunto inicial de documentos, pode-se citar como exemplo o uso de algoritmo de aprendizagem de máquina *Naive Bayes*. Ferramentas baseadas em estatísticas podem melhorar seu desempenho ao longo do tempo, à medida que mais recursos são processados, oferecendo um caminho promissor de geração automática de metadados de boa qualidade. Park e Brenza (2015) enfatizam que a principal vantagem desse tipo de técnica é que a extração de metadados pode ser feita independentemente da qualidade dos metadados associados a qualquer recurso de informação;

c) Indexação ou Classificação Automática

Gil Leiva (2009) afirma que o alcance da qualidade na recuperação da informação é obtido através do processo de indexação de documentos.

Suominem (2019) discorre que a indexação manual de documentos é uma tarefa intelectual que demanda bastante tempo, sendo que muitos destes artefatos estão em formato digital, tornando possível a automatização do trabalho de indexação a partir do texto completo ou certas partes de documentos, como títulos e resumos.

De acordo com Bandim e Correa (2019), a indexação é “um dos processos de análise documentária realizada com a finalidade de determinar para cada documento um conjunto de palavras-chave ou assuntos”. Esses autores afirmam que a organização e a recuperação da informação são materializadas via processo de indexação e sua automatização tem sido

adotada, visando a aplicação em um volume cada vez mais crescente de artigos científicos e da necessidade de elaboração de índices de busca que facilitem sua recuperação.

Silva e Correia (2020) contribuem com a reflexão de que a indexação automática pode se dividir em dois tipos: por atribuição ou por extração:

- A extração de termos dos textos dos documentos, fornecendo-lhe pesos e selecionando aqueles mais expressivos, representando seu conteúdo temático denomina-se de indexação automática por extração (LANCASTER, 2004, p. 18-19);

- Enquanto que atribuir termos ao documento através de outra fonte, tal como o emprego de termos extraídos de um vocabulário controlado, denomina-se de indexação automática por atribuição.

A indexação ou classificação automática envolve o uso de aprendizado de máquina e algoritmos baseados em regras para extrair valores de metadados dos próprios recursos de informação, em vez de depender do conteúdo das *metatags* aplicadas aos recursos, conforme Park e Brenza (2015). No entanto, estes autores afirmam que a técnica também abrange o mapeamento de termos de metadados extraídos para vocabulários controlados. Enfatizam que os pesquisadores utilizam algoritmos de classificação e agrupamento para extrair metadados relevantes dos textos. Demonstram também que são utilizadas estatísticas de frequência de termo ou documento IF.IDF em oposição à sua relativa não frequência em documentos relacionados. Conforme Polfreman *et al*. (2008), as tecnologias de última geração podem combinar os dois processos de extração e atribuição de conteúdo a categorias, criando automaticamente a taxonomia (denominada de ontologia) com base nos conceitos extraídos.

d) Mineração de textos e dados

Mineração de textos e dados – uso de aprendizado de máquina, análise estatística, técnicas de modelagem e tecnologia de banco de dados, para processar grandes quantidades de dados e identificar padrões recorrentes (POLFREMAN *et al*., 2008). Park e Brenza (2015) discorrem que esta é uma técnica complexa de implementação porque depende da qualidade e quantidade dos dados para desenvolver um modelo e usá-lo para treinar o sistema. Devido a

essa característica, poucas ferramentas foram totalmente desenvolvidas para aplicação em repositórios digitais;

e) *Folksonomias* ou Marcação Social

Folksonomias ou marcação social, conforme Polfreman et *al*. (2008), são funcionalidades desenvolvidas nas aplicações que confiam nas etiquetas geradas pelo autor e pelo usuário (na verdade, nos termos do assunto) para classificar os recursos. À medida que os recursos são acessados e compartilhados por outras pessoas, o vocabulário é usado e adicionado de maneira colaborativa. Verifica-se algumas aplicações atuais utilizando essa técnica em ferramentas de redes sociais, tais como *Facebook* e *Instagram*. Os autores discorrem que apesar do seu valor do ponto de vista do usuário, é improvável que as *folksonomias* substitua inteiramente os vocabulários controlados por causa de sua falta de precisão ou autoridade.

f) Geração automática de metadados extrínsecos

Geração automática de metadados extrínsecos, conforme Park e Brenza (2015), é o processo de extrair metadados sobre um recurso de informação que não está contido no próprio recurso, ou seja, um exemplo seria extrair metadados técnicos, como o formato e tamanho do arquivo, mas também pode incluir a extração de recursos mais complicados, como o nível de nota de um recurso educacional ou o público-alvo de um documento.

2.1.4. Síntese Qualitativa

A síntese qualitativa dos 12 (doze) artigos selecionados na RSL compreende a atividade de leitura, interpretação e compreensão das experiências e uso das ferramentas de geração automática e semiautomática de metadados, analisando suas técnicas, funções, aplicações e limitações.

2.1.4.1. Ferramentas de geração automática de metadados (GAM)

Após a introdução de documentos digitais a partir dos anos de 1950, o setor de biblioteconomia tem presenciado o potencial uso da geração automática de metadados (GAM)

como meio de simplificar o processo de descrição dos recursos de informação (KLEPPE et *al.*, 2019).

Ferramentas de geração automática de metadados não necessitam de intervenção humana, pois os algoritmos se encarregam de realizar a ação de geração de metadados automaticamente, conforme as regras de negócio implementadas no *software* com o uso de inteligência artificial e técnicas de aprendizagem de máquina.

Apresenta-se na Tabela 4, o produto da Revisão Sistemática da Literatura sobre as ferramentas utilizadas para geração automática de metadados, ordenado pela cronologia de publicação do trabalho. Em seguida, realizou-se a análise de cada uma delas, conforme as técnicas empregadas.

Tabela 4 – Pesquisas correlatas com uso de ferramentas de geração automática de metadados.

Autores	Publicação	Local	País	Origem	Técnicas
Kovaevic *et al*. (2011)	*Automatic extraction of metadata from scientific publications for CRIS systems. Electronic Library and Information Systems Vol. 45 No. 4, pp. 376-396*	Novi Sad	Sérvia	*Novi Sad University*	Extração de Conteúdo
Maratea, Petrosino e Manzo (2012)	*Automatic Generation of SCORM Compliant Metadata for Portable Document Format Files. International Conference on Computer Systems and Technologies – CompSysTech*	Nápoles	Itália	*Parthenope University*	Mineração de textos e dados
Verborgh *et al.* (2012*)*	*Enabling context-aware multimedia annotation by a novel generic semantic problem-solving platform. Multimed Tools Appl 61, 105–129*	Ghent	Bélgica	*Ghent University*	*Folksonomia* ou marcação social
Sah e Wade (2012)	*Automatic metadata mining from multilingual enterprise contente. Web semantics: Science, services and agents on the world wide web, Vol 11, p. 41-62*	Dublin	Irlanda	*Trinity College Dublin*	Colheita de *metatag*
Costa *et al.* (2013)	*EURAC SDI: A Near Real Time and Offline Automatic Metadata Generation Processing Chain. GI Forum, Conference Proceedings, volume 1*	Bozen	Itália	*Eurac Research*	Mineração de textos e dados
Vlachidis *et al.* (2013)	*Automatic Metadata Generation in an Archaeological Digital Library: Semantic Annotation of Grey Literature. In: Przepiórkowski A., Piasecki M., Jassem K., Fuglewicz P. (eds) Computational Linguistics. Studies in Computational Intelligence, vol 458. Springer, Berlin, Heidelberg*	Londres	*United Kingdom*	*University College London*	Extração de Conteúdo
Rafferty, Nugent e Liu (2015)	*Automatic Metadata Generation Through Analysis of Narration Within Instructional Videos. Transaction Processing Systems, J MedSyst nº 39, 94*	Belfast	Irlanda do Norte	*Ulster University*	Extração de Conteúdo

Gonzalo *et al.* (2018)	*ScienceSearch: Enabling Search through Automatic Metadata Generation. Conferência: 14th IEEE International Conferenceon E-Science (E-Science), p. 93-104*	Califórnia	EUA	*Berkeley University*	Indexação ou classificação automática
Yang e Park (2018)	*Automatic Extraction of Metadata Information for Library Collections. International Journal of Advanced Culture Technology, Vol.6, nº 2, p. 117-122*	Filadélfia e Mokpo	EUA e Coréia do Sul	*Drexel University e Mokpo University*	Extração de Conteúdo
Audichya e Saini (2019)	*Computational linguistic prosody rule-based unified technique for automatic metadata generation for Hindi poetry. 1st International Conference on Advances in Information Technology*	Gujarat	Índia	*Gujarat Technologic al University*	Indexação ou classificação automática
Morris (2020)	*Automated Language Identification of Bibliographic Resources. Cataloging & Classification Quarterly, 58:1, 1-27*	Wetherby	*United Kingdom*	*British Library*	Geração automática de dados extrínsecos

Fonte: Elaborado pelo autor (2021).

Observa-se a partir dos artigos listados, que o assunto pesquisado é relevante e de interesse ao redor do mundo, pois se verifica investigações sobre o tema por diversos pesquisadores de várias universidades localizadas nos Estados Unidos, Europa e Ásia.

A síntese dos artigos foi executada, destacando-a por meio da técnica utilizada:

Técnica: colheita de *metatag*

Sah e Wade (2012) investigam a utilização de ferramentas de geração automática de metadados para fornecer informações avançadas de conteúdos acessados pelos usuários, fomentando aspectos de personalização do cliente, fazendo com que eles permaneçam mais no site, incentivando-os a retornar ao provedor de serviços. Técnicas de seleção e navegação adaptáveis melhoraram a experiência do usuário em termos de assistência à tarefa e satisfação do usuário, tornando-os mais motivados em ler e se envolver mais com o sistema. Os autores desenvolveram uma ontologia *DocBook* e ontologia de tipo de recurso para extrair metadados estruturais e descritivos dos documentos *DocBook* no formato RDF. A ontologia do *DocBook* é independente do domínio e pode ser usada por outros aplicativos do *DocBook*. Além disso, propuseram um algoritmo para extrair semiautomaticamente uma ontologia de tópicos dos índices do *DocBook* usando o Sistema de Organização do Conhecimento Simples (SKOS). Finalmente, um novo sistema de granulação de informação e inferência difusa *Mandani* foi proposto para metadados cognitivos.

Técnica: extração de conteúdo

O trabalho de Kovaevic *et al*. (2011) apresenta um método para a extração automática de metadados de artigos científicos em formato PDF, que é projetado como parte integrante do sistema de informação para monitorar a atividade de pesquisa científica. O método é implementado como um complemento à entrada manual de metadados, no sentido de que os resultados da extração são oferecidos ao curador para inspecionar e corrigir antes de armazená-los no repositório. O sistema é baseado nos métodos de aprendizado de máquina (ML), ou seja, classificação. Os metadados são classificados em oito categorias pré-definidas: título, autores, afiliação, endereço, e-mail, resumo, palavras-chave e nota de publicação. As experiências foram realizadas por modelos de classificação padrão, conforme abaixo:

- Árvore de Decisão: árvore cujos nós internos são marcados com recursos e nos quais os ramos que vinculam esses nós são marcados pelos valores desses recursos, enquanto as folhas contêm rótulos de classe. A classificação é feita descendo a árvore ao longo do caminho ditado pelos valores do recurso até que uma folha seja alcançada. Esse classificador foi aplicado com sucesso em tarefas da extração da informação, como: classificação de nomes pessoais, extração de informações da web, análise de sentenças;

- *Naive Bayes*: pertence a classificadores probabilísticos que vêem o problema de categorização do ponto de vista das probabilidades condicionais determinadas pelo teorema de Bayes. Pode ser usado para: extrair metadados de rótulos de amostras e classificar texto usado na construção automática de portais de internet, por exemplo;

- *K-neighbors* mais próximos (KNN): pertencem à categoria de métodos de classificação lenta. Esses métodos não criam modelos de classificação explícitos, mas executam a categorização calculando a similaridade da nova instância com instâncias já em conjunto de treinamento. O rótulo da classe da nova instância é determinado pelos rótulos de k instâncias mais semelhantes;

- *Support Vector Machines* (SVM): Modelo de classificação com bons recursos de generalização e capacidade de lidar com dados de alta dimensão. Esta técnica foi desenvolvida para resolver problemas de classificação binária. O SVM resolve o problema de classificação encontrando o hiperplano de margem máxima que separa os vetores de recurso pelo rótulo de classe.

Kovacevic *et al.* (2011) discorrem que a tarefa de extração é formalizada como um problema de classificação realizada pelos métodos de aprendizado de máquina e destacam também que os classificadores utilizados na pesquisa foram avaliados, usando a validação cruzada de cinco vezes, em um corpus anotado manualmente de 100 trabalhos científicos em formato PDF, de várias conferências, periódicos e páginas pessoais dos autores. Oito modelos de SVM separados alcançaram os melhores desempenhos e, com base nisso, foram escolhidos como modelo de classificação para o sistema proposto em seus estudos. Todos os oito modelos SVM tiveram bom desempenho. O *F-measure* ultrapassou 85% para quase todos os classificadores e mais de 90% para a maioria deles. Os melhores resultados foram alcançados para o título (*F-measure* de 98,77 por cento) e e-mail (*F-measure* de 98,41 por cento). O desempenho das categorias resumo, autores, filiação e endereço foi um pouco menor, com *F-measure* de 91,52%, 92,13%, 90,37% e 87,80 por cento, respectivamente. Desta forma, compreende-se que a utilização do modelo de classificação SVM foi a melhor técnica empregada nos estudos de Kovacevic *et al.* (2011).

Vlachidis *et al*. (2013) realizam investigações sobre bibliotecas digitais, em especial a Europeana que possui algo em torno de 6 milhões de itens digitais do domínio cultural e patrimonial. O objetivo da pesquisa dos autores foi fornecer geração automática de metadados com enriquecimento semântico significativo para seus objetos digitais vinculados, através do *Europeana Data Model* (EDM), que resume o CIDOC *Conceptual Reference Model* (CRM) entre outros modelos de metadados. Foi empregando o kit de ferramentas de Arquitetura Geral para Engenharia de Texto (GATE). O processo de enriquecimento semântico é dividido em três fases amplas, cada uma subdividida em várias subtarefas (*pipelines*). A fase inicial processa previamente a literatura cinza e os recursos de vocabulário, enquanto a segunda fase identifica os conceitos de domínio no contexto. A fase final transforma anotações GATE em documentos semanticamente enriquecidos em forma de anotações XML, juntamente com os relatórios da literatura cinza e as representações RDF desacopladas de metadados. O estudo demonstrou a capacidade dos métodos baseados em CRM para impulsionar a geração automática de metadados avançados em bibliotecas digitais de domínio específicos. Esses

metadados podem ser expressos em formatos interoperáveis, como gráficos XML e RDF, que podem ser explorados pelos sistemas de bibliotecas digitais para permitir a funcionalidade de pesquisa cruzada entre recursos diferentes.

A pesquisa de Rafferty *et al*. (2015) apresenta um mecanismo de geração de metadados a partir de análises de clipes de vídeo, sendo que esses metadados devem ser usados no suporte ao fornecimento de instruções dinâmicas dentro de um paradigma *Smart Home*. De acordo com os pesquisadores, os metadados para os videoclipes eram fornecidos manualmente, o que poderia levar a registros incompletos ou incorretos. A geração automática de metadados com base em vídeo envolve principalmente a análise de interações de objetos em uma cena, realizando análises de elementos textuais em vídeos ou análises de conteúdo usando análise estatística. A geração de metadados baseada em áudio concentra-se principalmente na análise de som, com trabalho limitado, incorporando sistemas de reconhecimento de fala automatizados (ASR) que convertem fala em texto. Esses modelos, de acordo com Rafferty *et al*. (2015), não fornecem um método adequado para produzir anotações de atividades, pois atualmente não podem identificar um conjunto de ações de objetivo nesses clipes de vídeo. Os autores utilizaram um método de anotação capaz de gerar metadados enriquecidos para videoclipes, sendo criado e implementado dentro de uma plataforma de avaliação chamada *Audio BaSEd Instruction ProfiLer* (ABSEIL). Esta plataforma destina-se a trabalhar em conjunto com o repositório de vídeo gerado pelo projeto *Personal IADL Assistant* (PIA). O objetivo do projeto PIA é ajudar os idosos, oferecendo orientação com atividades instrumentais da vida diária, tais como: preparação de refeições, como utilizar o controle remoto de uma TV, como se barbear, limpar e manter uma casa, etc.

Os estudos de Yang e Park (2018) têm como objetivo apresentar um mecanismo de extração automática de metadados para atenuar problemas relacionados à aplicação inconsistente de metadados e à interoperabilidade semântica entre as coleções digitais. Os autores discorrem que os fenômenos linguísticos (sinônimos, homônimos e polissemia) podem gerar confusão no sentido de que comunidades diferentes podem usar formas de palavras variadas para fornecer conceitos idênticos ou semelhantes, ou podem usar as mesmas formas para transmitir conceitos diferentes. Os autores enfatizam que para obter metadados de qualidade e interoperabilidade semântica, é essencial um mecanismo de mediação que forneça relações contextuais entre os elementos de metadados e suas definições e uso correspondentes. Eles sugerem a construção de gráficos conceituais, pois eles têm um bom potencial para facilitar a interpretação adequada dos conceitos de metadados e o uso preciso e consistente dos elementos de dados. Um gráfico conceitual é uma das linguagens formais que

representam os significados das sentenças da linguagem natural. O gráfico conceitual pode ser utilizado como um mecanismo de mediação que aprimora a qualidade dos metadados ao desambiguar ambiguidades semânticas causadas pelo isolamento de um elemento de metadados e sua definição correspondente do contexto relevante. Portanto, é benéfico usar gráficos conceituais como uma linguagem para descrever a fonte formal de conhecimento de que as informações de metadados para coleções de bibliotecas podem ser extraídas automaticamente. Os autores demonstram um mecanismo de extração automática de informações de metadados para coleções de bibliotecas chamado ExMETA que foi projetado utilizando gráficos conceituais como representação interna. A ferramenta é capaz de analisar sentenças em linguagem natural e gerar metadados descritivos e estruturais, permitindo a eliminação de intervenções humanas (ou seja, catalogadoras). Os pesquisadores acreditam que isso contribuirá para melhorar a velocidade de geração de metadados, bem como a qualidade dos metadados e a interoperabilidade semântica.

Técnica: indexação ou classificação automática

Gonzalo *et al.* (2018) apresenta um estudo sobre o *ScienceSearch*, uma infraestrutura de pesquisa escalável generalizada que utiliza o aprendizado de máquina para capturar metadados de dados, contexto e artefatos circundantes. A implementação se concentrou no conjunto de dados do Centro Nacional de Microscopia Eletrônica, unidade do Departamento de Energia do Laboratório Nacional *Lawrence Berkeley*. Os dados deste instituto possuem milhões de micrografias produzidas por centenas de cientistas. Os autores identificaram, nesse contexto, vários artefatos que cercam esses dados, incluindo propostas de projetos, publicações e a estrutura do sistema de arquivos. Eles discorrem que o *ScienceSearch* permite a busca eficiente de dados com base em metadados gerados automaticamente. Os cientistas podem expressar suas necessidades de dados como uma consulta de texto em uma interface da web e receber uma lista de micrografias, propostas e publicações relevantes em segundos. A problemática identificada no estudo foi que os arquivos de micrografia gerados pelos cientistas, raramente incluem metadados além das configurações de captura do microscópio (por exemplo: exposição, contraste, tensão do sinal). Os usuários armazenam os metadados de diversas maneiras: cadernos de papel físico, convenções de nomenclatura elaborada de arquivos e diretórios ou simplesmente pela memória, sendo que falta de metodologia não facilita a preservação do conhecimento sobre os dados. Gonzalo *et al*. (2018) informam que o *ScienceSearch* gera novos metadados analisando quatro fontes principais de conhecimento:

- Estrutura de armazenamento de arquivos, capturando a lógica do usuário para organizar os dados;

- Propostas e publicações relacionadas às imagens, descrevendo a finalidade e o objetivo das micrografias;

- Dados de imagens, que podem ser modelados para identificar padrões comuns entre eles; e

- *Feedback* do usuário, que pode adicionar informações detalhadas a uma imagem de micrografia.

De acordo com Gonzalo *et al*. (2018), a arquitetura do *ScienceSearch* possui quatro componentes principais para captura de dados: importação ou ingestão de dados, extração de metadados, mecanismos de pesquisa e *feedback* do usuário. A avaliação de desempenho mostrou que o *ScienceSearch* é capaz de executar consultas simples em um único nó, em mais de 11 milhões de *tags* de metadados em menos de cinco segundos. Os testes iniciais realizados por usuários especializados indicaram que a qualidade da busca corresponde aos resultados esperados para suas consultas de teste. Os autores informam que trabalhos futuros melhorarão o modelo de pesquisa e explorarão métodos para avaliar algoritmos dos resultados da qualidade da pesquisa.

A pesquisa de Audichya e Asini (2019) teve como objetivo estruturar e padronizar adequadamente o conhecimento disperso sobre a prosódia, denominada de *Hindi Poetries*, disponível de maneira deficiente ou contraditória em diferentes fontes de informação. Foi utilizada a técnica de linguística computacional e o trabalho de pesquisa também se concentrou em moldar, um conjunto de regras padronizadas, para a geração automática de metadados com base nessas regras unificadas padrão de prosódia. Os autores testaram um gerador de metadados em 3026 entradas que incluem diferentes poemas, parte de poemas que cobriam mais de 30 "*Chhands*" (quadra/estrofe usada nas tradições poéticas do Norte da Índia e Paquistão), além de cobrir também sua classificação e subclassificação. O resultado do trabalho de pesquisa foi suficiente para provar a robustez da metodologia e do mecanismo técnico do gerador de metadados, que alcançou 98,09% de taxa de sucesso, juntamente com

1,91% de falha devido a erros de formatação no texto, ausência de delimitador ou uso excessivo e irregular de uso de delimitadores.

Técnica: mineração de textos e dados

Maratea, Petrosino e Manzo (2012) utilizaram algoritmos com técnicas de processamento de linguagem natural para geração automática de metadados para conteúdos de aprendizagem. Foi utilizada como padrão uma coleção de especificações para o *e-learning* baseado na Web amplamente adotado em todo o mundo, denominado Modelo de Referência de Objeto compartilhável para conteúdo (SCORM). Segundo os autores, o objetivo deste modelo é permitir a interoperabilidade, fácil acesso e reutilização de unidades de aprendizagem baseadas na Web para indústria, governo e universidade. A principal vantagem do SCORM é a reutilização de objetos de aprendizagem (OA), ao custo de uma anotação completa que geralmente pode ser garantida apenas por um especialista humano. Eles utilizaram 14 artigos sobre *e-learning* escritos em 7 idiomas diferentes (inglês, francês, alemão, espanhol, italiano, português e polonês) para testar o método. Em todos os casos testados, as técnicas propostas classificam corretamente a linguagem, com uma boa margem, o título e o resumo, o local e isolou um bom conjunto de palavras-chave.

Costa *et al*. (2013) apresentam em seu estudo a abordagem para geração automática de metadados através de um método baseado em regras, codificado manualmente, implementados como *plug-ins* que geram metadados em formato padronizado extraído de um conjunto heterogêneo de dados geoespaciais. Os autores discorrem no artigo que a estação receptora EURAC recebe diariamente dados brutos das missões da NASA Aqua, Terra e Suomi NPP. A pesquisa do instituto lida com muitos dados de satélite diferentes: LANDSAT, RapidEye, ENVISAT e *Quickbird*. Como a quantidade de dados cresce rapidamente, os autores discorreram sobre a necessidade de automatizar o tratamento de dados e a geração de metadados. Eles usaram o conceito de ingestão e manuseio de dados que foi implementado por meio do desenvolvimento de um servidor principal, denominado de *Data Exchange Server* (DES), que pode ser considerado um aplicativo geral multitarefa que executa qualquer tipo de tarefa ou trabalhos configuráveis. A tecnologia usada para a implementação da arquitetura geral e da cadeia de processamento de dados foi Java, Perl, XML e XSLT.

Técnica: *folksonomia* ou marcação social

Verborgh *et al.* (2012) descrevem, em seus estudos, como o conhecimento e as tecnologias da Web Semântica podem fornecer um contexto para apresentar algoritmos de extração, gerando anotações multimídia que os algoritmos não podem descobrir individualmente. Apresentam uma plataforma genérica de solução de problemas semânticos, que combina automaticamente os serviços da Web para realizar uma tarefa predefinida e usa a Web Semântica como fonte de conhecimento para iniciar e manter ativamente o contexto da tarefa. Os autores realizaram a aplicação através de um caso de uso de anotação de imagem. Como exemplo, cita o caso de um editor de uma revista de assuntos atuais que possui um arquivo de fotos digitais e que precisa ser anotado. Além dos dados de bitmap da imagem, nenhuma informação adicional estava disponível. Como primeiro passo, queriam identificar as pessoas nas fotografias. As anotações devem ser vinculadas às entidades correspondentes da DBpedia para permitir pesquisas semânticas. Obtiveram resultados satisfatórios quanto à eficiência e escalabilidade, sendo utilizada a ferramenta de raciocínio *Eye* para resolução de problemas semânticos e capaz de criar composições holísticas. Os autores indicaram a importância das informações contextuais na anotação multimídia e demonstraram como a plataforma proposta pode oferecer contexto aos algoritmos de extração de recursos multimídia.

Técnica: geração automática de metadados extrínsecos

Morris (2020) investigou se os códigos de idioma podem ser atribuídos automaticamente aos registros do MARC e avaliar a precisão de qualquer método viável de fazê-lo. A autora enfatiza que nos registros bibliográficos do MARC21, o conteúdo de idioma de um recurso de informação é registrado nas posições do campo 008 35–37, usando um código de três posições de uma lista controlada. Entretanto, uma análise do catálogo da *British Library* em outubro de 2018 revelou que esse código de idioma não era preenchido em quase 4,7 milhões de registros. Desses, 78% também não possuíam um código para o local de publicação no campo 008 posições 15–17.

2.1.4.2. Ferramentas de geração semiautomática de metadados (GSAM)

O artigo de Park e Brenza (2015), intitulado "*Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art*" e publicado na revista *Information Technology and Libraries*, Volume 34, Ed. 3, p. 22-42; merece um destaque. A publicação foi a mais indexada nas bases pesquisadas. Eles examinam uma variedade de ferramentas de geração semiautomáticas de metadados (N=39), analisando suas técnicas, recursos e funções.

Esses autores desenvolveram uma matriz caracterizando cada ferramenta de geração semiautomática de metadados analisada: nome, local online, técnicas usadas para geração de metadados (GREENBERG, 2003 e POLFREMAN *et al*., 2008), além de breve descrição das funções e recursos da ferramenta. Os levantamentos realizados por Park e Brenza (2015) podem ser visualizados na Tabela 5:

Tabela 5 – Ferramentas de geração semiautomática de metadados.

Ferramentas	Técnicas Empregadas	Características e Funções
ANVL/ERC *Kernel Metadata Conversion Toolkit* https://metacpan.org/pod/File::ANVL	Colheita de metatags	Um utilitário que pode automaticamente converter registros de formatos ANVL (*A Name Value Language*) para outros formatos, tais como XML, JSON (*JavaScript Object Notation*), *Turtle* e *Plain*, dentre outras.
Apache POI – *Text Extractor* http://poi.apache.org/download.html	Extração de conteúdo Colheita de metatags Geração automática extrínseca	O Apache POI fornece extração básica de texto para todos os formatos de arquivo suportados pelo projeto. Além do texto (sem formatação), ele pode acessar os metadados associados a um determinado arquivo, como título e autor.
Apache Stanbol (movido para *Apache Attic*, desativado) https://attic.apache.org/projects/stanbol.html	Extração de conteúdo; Indexação automática	Extrai metadados semânticos de arquivos PDF e de texto. Pode aplicar termos extraídos a ontologias.
Apache Tika http://tika.apache.org/	Extração de conteúdo Colheita de metatags Geração automática extrínseca	Construído no Apache POI, o Apache Tika toolkit detecta e extrai metadados e conteúdo de texto de vários documentos.
Ariadne Harvester https://sourceforge.net/projects/ariadnekps/	Colheita de metatags	Uma colheitadeira de registros compatíveis com OAI-PMH que pode ser convertida em vários outros esquemas, como LOM (Metadados de Objetos de Aprendizagem).
Bibframe Tools https://www.loc.gov/bibframe/implementation/	Colheita de metatags	O BIBFRAME oferece várias ferramentas para a conversão de documentos MARCXML em documentos BIBFRAME. Serviço da Web e *software* para download estão disponíveis.
Biblio Citation Parser https://metacpan.org/release/MJEWELL/Biblio-Citation-Parser-1.10	Extração de conteúdo;	Um conjunto de módulos para análise de citações.

CatMDEdit https://inspire-reference.jrc.ec.europa.eu/tools/catmdedit https://sourceforge.net/projects/catmdedit/	Extração de conteúdo;	O CatMDEdit permite a criação automática de metadados para coleções de recursos relacionados, em particular séries espaciais que surgem como resultado da fragmentação de recursos geométricos em conjuntos de dados de tamanho gerenciável e escala semelhante.
CrossRef https://doi.crossref.org/simpleTextQuery	Extração de conteúdo;	Este serviço da Web retorna Identificadores de Objetos Digitais para referências inseridas
Data Fountains https://sourceforge.net/projects/datafountains/	Extração de conteúdo; Indexador automático; Colheita de metatag; Geração automática extrínseca	Digitaliza documentos HTML e extrai primeiro as informações contidas nas metatags. Se as informações não estiverem disponíveis nas metatags, o programa usará outras técnicas para atribuir valores. Inclui um rastreador da web focado que pode segmentar sites sobre um assunto específico.
Digital Record Object Identification (DROID) https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/	Geração automática extrínseca	DROID é uma ferramenta de software desenvolvido pelo Arquivo Nacional (Reino Unido) para execução automática em lote de identificação de formatos de arquivo.
Dublin Core Meta Toolkit https://sourceforge.net/projects/dcmetatoolkit/	Colheita de metatags	Transforma dados coletados por diferentes métodos em metadados compatíveis com Dublin Core (DC)
Dspace https://duraspace.org/dspace/	Colheita de metatags Geração automática extrínseca Marcação Social	Extrai automaticamente informações técnicas sobre o formato e tamanho do arquivo. Também pode extrair algumas informações das metatags.
Editor-Converter Dublin Core Metadata http://library.kr.ua/dc/dceditunie.html	Colheita de metatags Geração automática extrínseca	Digitaliza documentos HTML, coletando metadados de tags e convertendo-os em Dublin Core (DC).
Embedded Metadata Extraction Tool (EMET) https://sourceforge.net/projects/emet/	Extração de conteúdo; Colheita de metatag; Geração automática extrínseca	O EMET é uma ferramenta projetada para extrair metadados incorporados em arquivos JPEG e TIFF; Detecta corrupção de arquivo de imagem e exporta resultados em arquivo Excel.
Firefox Dublin Core Viewer Extension https://www.splintered.co.uk/experiments/73/ (Plugin descontinuado)	Colheita de metatag; Geração automática extrínseca	Digitaliza documentos HTML, coletando metadados de tags e exibindo-os no Dublin Core
FreeCite (descontinuado e substituído por AnyStyle) https://anystyle.io/	Extração de conteúdo;	Ferramenta de análise gratuita para a extração de informações de referência. Pode ser baixado ou usado como um serviço web.
General Architecture for Text Engineering (GATE) https://gate.ac.uk/overview.html	Extração de conteúdo; Indexação automática	Processador de linguagem natural e extrator de informações
JHove https://jhove.openpreservation.org/	Geração automática extrínseca	Extrai metadados sobre o formato e tamanho do arquivo, bem como validando a estrutura do formato do arquivo identificado
Kea http://community.nzdl.org/kea/	Extração de conteúdo; Indexação automática	Analisa os textos completos dos recursos e extrai frases-chave. As frases-chave também podem ser mapeadas para ontologias personalizadas ou vocabulários controlados para atribuição de termos do assunto.

MarcEdit https://marcedit.reeset.net/	Colheita de metatag;	Coleta dados compatíveis com OAI-PMH e os converte em vários formatos, incluindo DC e MARC.
MetaGen https://www.codeproject.com/Articles/41910/MetaGen-A-Project-metadata-Generator-for-Visual-St	Extração de conteúdo; Indexação automática	O modelo de texto T4 para Visual Studio é uma mistura de blocos de texto e lógica de controle (fragmentos de código em Visual C# ou no Visual Basic) que pode gerar um arquivo de texto de qualquer tipo, como uma página da Web, um arquivo de recurso ou um código-fonte do programa em qualquer idioma. O MetaGen usa o T4 para gerar metadados para projetos Silverlight.
MetaGenerator https://extensions.joomla.org/extension/site-management/seo-a-metadata/seo-generator/	Extração de conteúdo;	Um plug-in que gera automaticamente descrição e metatags de palavras-chave puxando texto do conteúdo do joomla. Com este plugin é possível controlar algumas opções de título e adição de URL metatags.
Metatag Extractor Software https://meta-tag-extractor.soft112.com/	Colheita de metatag;	Permite recursos personalizáveis de extração, coleta de metatags e informações de contato de sites
My Meta Maker http://old.isn-oldenburg.de/services/mmm/	Colheita de metatag;	Permite converter dados inseridos manualmente em DC.
National Library of New Zealand – Metadata Extraction Tool http://meta-extractor.sourceforge.net/	Geração automática extrínseca	Desenvolvido pela Biblioteca Nacional da Nova Zelândia para extrair programaticamente metadados de preservação de vários formatos de arquivo, como documentos PDF, arquivos de imagem, arquivos de som, documentos do Microsoft Office e outros
Omeka http://omeka.org/	Geração automática extrínseca Marcação Social	Extrai automaticamente informações técnicas sobre o formato e tamanho do arquivo
Ont-O-Mat (descontinuado)	Extração de conteúdo;	Auxilia o usuário na anotação de sites compatíveis com a Web Semântica. Pode incluir um recurso que sugere automaticamente partes do site a serem anotadas.
Open Text Summarizer https://open-text-summarizer.soft112.com/	Extração de conteúdo;	Extrai frases pertinentes de um recurso para criar uma descrição de texto livre.
ParsCit https://parscit.comp.nus.edu.sg/	Extração de conteúdo;	Pacote de análise de cadeia de código aberto para a extração de informações de referência de artigos acadêmicos.
Photo RDF-Gen http://www.webposible.com/utilidades/photo_rdf_generator_en.html	Colheita de metatag;	Gera saída Dublin Core e *Resource Description Framework* (RDF) a partir da entrada inserida manualmente.
PyMarc https://pypi.org/project/pymarc/	Colheita de metatag;	Biblioteca python para trabalhar com dados bibliográficos codificados em MARC21. Ele fornece uma API para leitura, gravação e modificação de registros MARC
RepoMMan http://www.ukoln.ac.uk/repositories/digirep/index/RepoMMan (Não disponível em 23/07/2020)	Extração de conteúdo; Colheita de metatag; Geração automática extrínseca	Extrai automaticamente vários elementos para documentos enviados ao Fedora, como autor, título, descrição e palavras-chave, entre outros. Os resultados são apresentados ao usuário para revisão.
Sherpa/RoMEO (descontinuada em julho de 2020) https://v2.sherpa.ac.uk/api/	Colheita de metatag;	Uma API (*Application Program Interface*) de máquina para máquina que permite a pesquisa e importação automáticas de editores e revistas.

Simple Automatic Metadata Generation Interface (SamgI) – Não está disponível para acesso em 23/07/2020.	Extração de conteúdo; Geração automática extrínseca	Um conjunto de ferramentas capazes de extrair automaticamente elementos de metadados como frase-chave e idioma dos documentos a partir do contexto em que um documento existe.
URL and Metatag Extractor https://metatagsextractor.com/	Colheita de *metatag*;	Permite a pesquisa direcionada de sites e extrai URLs e metatags desses sites
Termine http://www.nactem.ac.uk/software/termine/	Extração de conteúdo;	Extrai palavras-chave de textos através da análise do *C-Value* e Acrônimos, um identificador de sigla e dicionário. Disponível gratuitamente serviço web para uso acadêmico.
Yahoo Content Analysis API (descontinuado em 30/06/2020)	Extração de conteúdo; Indexação automática	Detecta entidades, conceitos, categorias e relacionamentos dentro de conteúdo não estruturado. Classifica as entidades/conceitos por sua relevância.

Fonte: Park e Brenza (2015) com atualizações e adaptações do autor

Realizou-se acesso às ferramentas descritas no período de 11/07/2020 a 26/07/2020, sendo identificados alguns *links* que estavam quebrados – na Tabela 5 – foram todos atualizados. Algumas ferramentas foram descontinuadas, outras possuem link para a documentação, mas não ao projeto e ao *software* para *download* (*Apache Stanbol, Ont-O-Mat*, *FreeCite* foi substituído por *AnyStyle*, *RepoMMan*, *Sherpa/Romeu* passou para o projeto *Sherpa* na versão 2, *Simple Automatic Metadata Generation Interface – SamgI,Yahoo Content Analisys* API e o *plugin* do *Firefox Dublin Core Viewer Extension*).

Park e Brenza (2015) discorrem que apesar das ferramentas de geração semiautomática de metadados oferecerem muitos benefícios, especialmente no que se refere à racionalização do processo de criação de metadados, existem barreiras significativas à adoção e implementação generalizadas delas. Um fator é que muitas são desenvolvidas localmente para atender as necessidades específicas de um determinado projeto ou como parte da pesquisa acadêmica. Esse ambiente altamente focado para um contexto específico significa que a aplicabilidade geral das ferramentas é potencialmente diminuída. Outro fator, discorrido pelos autores, é o alto nível de conhecimento técnico exigido para seu desenvolvimento e sua incorporação aos fluxos de trabalho diário no processo de criação de metadados. Por fim, ferramentas de metadados semiautomáticas não são testadas em cenários do mundo real, tais como: pequenos tamanhos de amostra, escopo restrito de domínios de projeto e experimentos que não têm objetividade, conforme relatado por Polfreman *et al*. (2008).

2.1.4.3. Possibilidades de uso das ferramentas de GAM e GSAM

As ferramentas de geração automática e semiautomática de metadados descritas na síntese possuem diversas aplicabilidades na qual citamos:

a. Executar a produção de metadados para conteúdo de aprendizagem com o objetivo de facilitar sua busca e recuperação nos acervos digitais, além de permitir a interoperabilidade (MARATEA, PETROSINO e MANZO, 2012);

b. Resolver problemas semânticos nos repositórios, por meio do uso das ferramentas para geração automática de metadados, propiciando a execução de seu enriquecimento semântico (VLACHIDIS *et al.*, 2013);

c. Fornecer informações avançadas para personalização de um sistema para o cliente, ampliando sua capacidade de uso e melhorando a experiência do usuário (SAH e WADE, 2012);

d. Disponibilizar sistema de catalogação espacial e ferramenta de geração automática de metadados geoespaciais para tomada de decisão e compartilhamento de dados entre instituições (COSTA *et al*., 2013). A solução possui possibilidade de ser aplicada em diversas áreas, tais como: previsão do tempo para a agricultura, desastres naturais, desmatamento florestal, defesa e segurança pública, inteligência, comunicação, aquecimento e efeito estufa, saúde, monitoramento dos recursos hídricos e ambientais, tráfego urbano, aéreo e rodovias etc.;

e. Melhorar o significado dos metadados de objetos vinculados em um repositório por meio da anotação multimídia, aperfeiçoando a recuperação da informação (VERBORGH *et al*., 2012);

f. Fornece instruções dinâmicas a partir de clipes de vídeo, facilitando a compreensão e acessibilidade à informação do usuário ao recurso de informação (RAFFERTY *et al*., 2015);

g. Conversão de dados, documentos e arquivos; extração de metadados de textos e arquivos de imagens; aplicação de metadados em ontologias; análise de citações; criação de metadados para coleções; indexação de documentos etc. (PARK e BRENZA, 2015).

h. Gerar e incluir metadados ausentes em repositórios com grandes volumes de dados, objetivando a completude dos dados armazenados (MORRIS, 2020);

i. Auxiliar na atenuação de problemas de interoperabilidade semântica entre as coleções digitais (YANG e PARK, 2018);

j. Propor e sugerir metadados de escritos antigos e apoiar sua catalogação, classificação e indexação, objetivando a preservação histórica e cultural (AUDICHYA e ASINI, 2019);

k. Atribuir automaticamente os metadados em um recurso de informação e nos repositórios digitais (KOVAEVIC *et al*., 2011).

Observam-se as diversas aplicações realizadas pelas ferramentas de geração automática e semiautomática de metadados, gerando eficiência nas atividades de extração de dados, geração de metadados, enriquecimento semântico, indexação e catalogação, auxiliando na gestão de grandes volumes de registros armazenados nos repositórios digitais.

2.1.4.4. Limitações das ferramentas de GAM e GSAM

As limitações encontradas nas ferramentas e as técnicas empregadas na geração automática e semiautomática de metadados analisadas foram:

a. O alto grau de especialização dos algoritmos para extração de metadados faz com que eles desconheçam o contexto em que operam, inclusive os que contém informações valiosas e muitas vezes necessárias. É necessário aperfeiçoar o tratamento de informações imperfeitas, tais como a incerteza e a incompletude (VERBORGH *et al*., 2012);

b. A mineração automática de metadados cognitivos é um desafio, uma vez que é muito difícil compreender automaticamente o conhecimento intelectual subjacente sobre o documento (SAH e WADE, 2012);

c. Desafio em gerenciar grandes volumes de dados espaciais heterogêneos, assim como melhorar sua organização, busca e prevenir a duplicidades (COSTA *et al*., 2013);

d. As ferramentas são desenvolvidas localmente para atender necessidades específicas, altamente focadas para um contexto particular. Além disso, exigem alto grau de habilidade técnica para sua implementação (PARK e BRENZA, 2015);

e. Limitação à descrição correta e completa de *metatags*, registros *DC* e *MARC* para extração e transformação (MORRIS, 2020);

f. Dependência da seleção do vocabulário controlado, da qualidade dos campos de metadados extraídos de *strings* de referência e de um conjunto bem estruturado dos documentos PDF (MARATEA, PETROSINO e MANZO, 2012);

g. Escalabilidade e desempenhos impactados por variáveis externas: *hardware*, *software*, sistema operacional, ambiente/plataforma computacional, funcionalidades na própria ferramenta ativadas ou não (GONZALO *et al*., 2018);

h. Indicação de que, em vários casos, os metadados extraídos automaticamente não podem ser inseridos na base de dados, pois dependem de controle da curadoria (KOVACEVIC *et al.*, 2011);

i. Ausência de geração de meta-conhecimento ou mapa de conhecimento estruturado, como forma de explicitação do conhecimento por meio da documentação, impactando o repasse e o compartilhamento de uso da ferramenta (YANG E PARK, 2018);

j. Necessidade de avaliação da ferramenta em larga escala para analisar o desempenho de extrações de metadados, considerando ambiguidades lexicais e a precisão de anotação do termo (VLACHIDIS *et al*., 2013);

k. Ferramenta depende totalmente da entrada correta e completa dos dados, de acordo com as regras gramaticais, tais como o uso de acentos, de caracteres comuns e especiais, além dos sinais de pontuação (AUDICHYA e SAINI, 2019).

2.1.5. Interpretar e documentar resultados

Baseado nas possibilidades de uso e nas limitações das ferramentas apresentadas nas seções anteriores, é mister haver pesquisas com a finalidade de desenvolver, implementar, melhorar, customizar e adequar ferramentas de geração automática e semiautomática de metadados para auxiliar os gestores de repositórios digitais a realizarem as atividades de geração de metadados e assuntos, sua classificação e indexação.

As aplicabilidades das ferramentas GAM e GSAM são as mais diversas possíveis e são soluções que podem auxiliar e melhorar a representação, organização, armazenamento, captura e recuperação da informação de forma eficiente.

Identificou-se o uso de diversas técnicas empregadas tanto para as ferramentas GAM e GSAM, no qual cito: colheita de *metatags*, extração de conteúdo, indexação ou classificação automática, mineração de textos e dados, *folksonomia* ou marcação social e geração automática de metadados extrínsecos.

Diversas áreas do conhecimento ampliaram os seus repositórios em formatos digitais devido ao avanço tecnológico. Desta forma, o uso de ferramentas GAM e GSAM são imprescindíveis para automatizar o processo manual executado em várias bibliotecas, fornecendo o apoio necessário ao gestor destes acervos e na melhoria de sua gestão e controle dos documentos digitais.

A identificação de lacunas e limitações no uso de diversas soluções apresentadas, indica a possibilidade de ampliar as investigações no sentido de analisar as características positivas de cada ferramenta GAM ou GSAM para implementação conjunta, diminuindo os *gaps* elencados.

Vive-se em um mundo globalizado e soluções tecnológicas desenvolvidas para atender apenas um determinado idioma não são abrangentes. É necessária a implementação de

soluções com pelo menos o idioma nativo mais o inglês (língua universal). O ideal é que as ferramentas fossem multilíngues, suportando diversos idiomas.

Percebe-se ferramentas dependentes de vocabulário específico. As soluções deveriam ser implementadas, propiciando ao usuário realizar a carga de qualquer vocabulário controlado em formato aberto, realizando o treinamento de seu modelo, utilizando diversos algoritmos que realizem a atividade de aprendizagem de máquina.

É necessária a integração dos algoritmos de aprendizagem de máquina com as soluções de catalogação e indexação. Essa abordagem é bastante interessante, pois as ferramentas GAM e GSAM precisam ser integradas nas soluções que façam indexação de documentos para prover uma melhor organização do repositório digital, facilitando a recuperação de documentos com eficiência.

Identificou-se o uso de diversas tecnologias para soluções de problemas específicos e não foi identificado um padrão de utilização ou consenso de aplicação para uso geral.

De acordo com Cerrao e Castro (2018), a Ciência da Informação objetiva idealizar padrões e modelos de estrutura que proporcionem maior qualidade e confiabilidade na guarda e propagação da informação, por isso é uma área que estuda os métodos e estruturas que possam ser empregadas na recuperação da informação. Desta forma, esta pesquisa busca contribuir para a melhoria dos recursos de busca e na recuperação da informação.

2.1.6. Publicar e reportar a RSL

A revisão sistemática forneceu a compreensão das técnicas empregadas e as ferramentas tecnológicas utilizadas para geração automática e semiautomática de metadados.

Para atendimento da comunicação científica da pesquisa, foi elaborado um artigo sobre a RSL aplicada neste estudo e publicado no XXI Encontro Nacional de Pesquisa em Ciência da Informação (XXI Enancib[1]), no período de 25-29 de outubro de 2021 no Rio de Janeiro. O título da publicação foi "Geração automática e semiautomática de metadados: uma revisão sistemática da literatura". O documento está disponível na BRAPCI, na seguinte URL https://brapci.inf.br/index.php/res/v/192450.

Observação: A publicação inicial foi realizada como revisão sistemática. Entretanto, um dos revisores para o XXI ENANCIB pediu para modificar o termo "revisão sistemática" por

[1] XXI Enancib, Grupo de Trabalho 8, Comunicação 7, disponível em <https://enancib2021rio.ibict.br/programa-do-gt-8/>, acesso em 21/02/2023.

“revisão bibliográfica” e alterar as citações referenciadas. Executou-se as modificações, conforme solicitado pelo revisor e, em seguida submeteu-se o artigo final que continuou com o título de “revisão sistemática” na chamada, mas como revisão bibliográfica no conteúdo do texto. Para todos os efeitos, considera-se que o método realizado nessa seção da tese foi uma revisão sistemática da literatura.

2.2. REVISÃO DE LITERATURA

A revisão de literatura por conveniência, conforme Galvão, Pluye e Ricarte (2017), são estudos que não demonstram os elementos de sua construção e não são reproduzíveis. Galvão e Ricarte (2019) enfatizam que nessa modalidade de pesquisa, o investigador executa a junção de diversos trabalhos científicos, julgando-os relevantes para tratar de um assunto, mas não apresenta os critérios de como a investigação foi projetada.

Marconi e Lakatos (2003, p. 158) discorrem que a revisão compreende a investigação em estudos já realizados, revestidos de importância, nas seguintes fontes de conhecimento: artigos, periódicos e *journals*, fornecendo dados atuais e relevantes relacionados com o tema.

Esta seção aprofundará a fundamentação teórica, executando uma revisão de literatura relacionada com os conceitos sobre: informação, ciência da informação, organização, representação e recuperação da informação, repositórios digitais; a definição de vocabulário controlado, listas de termos e assuntos, arquivos de autoridade, taxonomias e tesauros; conceitos sobre a inteligência artificial e corpus de conhecimento; abordagens de ferramentas analisadas à posteriori da RSL, a solução ANNIF; e o *framework* genérico proposto.

2.2.1. Informação

Retomamos aqui, o primeiro parágrafo da introdução desta tese onde discorre sobre o significado em latim do termo informar ou informação, que se relaciona com a finalidade de “colocar em forma” e a valorização do conteúdo.

Shannon e Weaver (1949) publicaram a obra intitulada *A Mathematical Theory of Communication*, destacando que a transmissão da informação era um fenômeno quantificável e estatístico. Os autores, afirmavam que a informação é a capacidade de liberdade de escolha, diante de um processo de apuração de uma mensagem, não possuindo relação ao que é expresso, mas ao que poderá ser expresso. Preocupa-se na redução do grau de incerteza ao colher uma reposta de uma pergunta proferida, não se preocupando com questões de

semântica. Entretanto, os autores afirmam ainda, que uma pessoa recebe informação quando o que ela conhece se altera, se modifica.

Goffman (1970) discorre que o termo informação é usado em diversos contextos diferentes e que uma única definição precisa englobar todos os aspectos, por isso não pode ser formulado ou não é uma linha útil de investigação. A Ciência da informação que Goffman propõe é aquela que estuda todos os fenômenos relacionados à informação, ao invés da informação em si mesma.

Wersig e Neveling (1975) descrevem seis abordagens para delinear a informação:

1) Ela é constituída das estruturas da natureza e independe da assimilação pelo ser humano;
2) O conhecimento construído fundamentado na apreensão das estruturas da natureza é considerado informação;
3) Ela é utilizada de forma análoga à mensagem (conteúdo);
4) Apenas o significado da mensagem é considerado informação;
5) Ela é o resultado de um efeito específico de um processo específico; e
6) Ela não é um componente do processo, mas é o próprio processo.

Belkin e Robertson (1976), afirmam que a informação é um fenômeno que possibilita o ser humano a modificar suas estruturas cognitivas, através do processamento dessa informação obtida de variadas fontes, através da percepção dos estímulos do ambiente, da experiência e as características individuais. Em 1978, Belkin realiza sua própria definição de informação, como sendo "a estrutura conceitual modificada do gerador (por propósito, intenção, conhecimento do estado de conhecimento do destinatário) que está por trás da estrutura da superfície (por exemplo, linguagem) desse texto".

Mikhailov (1980) discorre que a informação é consequência das ações sociais de construção do conhecimento, atuando como agente de mudança da realidade relacionadas com a sociedade humana.

Brookes (1980) afirma que a informação é o insumo necessário que compõe a estrutura do conhecimento, enfatizou que é necessário ter o conceito de informação para compreender a equação fundamental:

$$\Delta I = (S + \Delta S) - (S)$$

Brookes explica que o estado de conhecimento (S) é afetado por algum incremento de informação ΔI, resultando em novo estado de conhecimento (S + ΔS). A equação serve para enfatizar o quão pouco sabemos sobre as maneiras pelas quais nosso conhecimento cresce. Brookes finaliza discorrendo que a informação pode depender da observação sensorial, mas os dados dos sentidos assim recebidos devem ser interpretados subjetivamente por uma estrutura de conhecimento para se tornarem informações.

Buckley (1983) enfatiza que o significado só é possível de ser construído através das interações sociais entre os indivíduos.

De acordo com Buckland (1991), há uma ambiguidade do termo "informação" e sua utilização em diversos contextos distintos. Um aspecto importante para o uso de informação denota conhecimento transmitido, outro como um processo de informação. O autor delimita a informação em três usos principais:

- Informação-processo: informação como o ato de informar, quando alguém é informado, o que eles sabem é alterado. Pode ser compreendido através do relato de um acontecimento, de um fato;

- Informação como conhecimento: designa aquilo que é percebido na informação como processo, conhecimento comunicado sobre determinado fato, assunto ou evento. É intangível, não pode ser tocado, é pessoal, subjetivo e conceitual. Para comunicar o conhecimento, este tem que ser expresso, descrito ou representado de forma física, como sinal, texto ou comunicação.

- Informação como coisa: utilização atributiva para objetos, como os dados e documentos que são referenciados como informação – tem qualidade de transmissão de conhecimento ou comunica informação, instrutivo.

Seracevic (1995a) correlaciona o conceito de recuperação da informação associada a disponibilização efetiva da informação para o usuário, ou seja, se o objeto recuperado não possuir relevância, não se considera o objeto como uma informação.

Miksa (1999), apresenta em seu segundo paradigma – movimento da informação como um sistema de comunicação humana (anos 1950) – sofreu influências da teoria da cibernética e engenharia da comunicação. O foco em sistemas de comunicação humana, onde os documentos são recuperados como respostas às perguntas feitas – forte influência também da teoria

matemática de Shannon, pois os atuais bancos de dados utilizam *scripts*, parâmetros e cálculos *booleanos* para recuperação de informação nos repositórios de dados. A informação é mensurável, quantificável, considerada um fenômeno físico.

Floridi (2004), afirma que a palavra informação recebeu significados diferentes por vários escritores do campo geral da teoria da informação. O autor discorre que a informação pode ser vista a partir de três perspectivas: informação como realidade (por exemplo, como padrões de sinais físicos, que não são verdadeiros nem falsos), também conhecido como informação ecológica; informação sobre realidade (informação semântica, dialeticamente qualificável); e informação para a realidade (instrução, como informação genética).

Zins (2007) enfatiza a inter-relação entre dado, informação e conhecimento, discorrendo que muitos pesquisadores afirmam que esses elementos fazem parte de uma ordem sequencial. Entretanto, a partir de um painel internacional apresentado por 57 acadêmicos renomados na área da Ciência da Informação, elaborou-se um documento antropológico com 130 definições para dado, informação e conhecimento. Nesse contexto, Zins discorre que muitas fundamentações estão bem embasadas, mas outras carecem de completude e consistência, além de apresentarem problemas filosóficos. A comunidade acadêmica fala sobre o assunto de diversas maneiras diferentes. Mesmo assim, o estudo mapeia várias abordagens conceituais para definição de dado, informação e conhecimento no contexto da ciência da informação. O modelo mais comum utilizado, representando que a base lógica da CI possui o foco em explorar dados e informações (fenômenos externos), não explorando o elemento conhecimento (fenômeno interno).

Para finalizar as reflexões sobre informação, será apresentado o resumo de notas da Disciplina Fundamentos da Informação, compilando e realizando o cotejamento dos conceitos, sendo representado pela Tabela 6:

Tabela 6 – Conceitos de dado, informação e conhecimento

DADO	INFORMAÇÃO	CONHECIMENTO
Simples observações sobre o estado do mundo	Dados dotados de relevância e Propósito	Informação valiosa da mente humana. Inclui reflexão, síntese, contexto
Facilmente estruturado	Requer análise	De difícil estruturação
Facilmente obtido por máquinas	Exige consenso em relação ao Significado	De difícil captura por máquinas
Frequentemente quantificado	Quantificado e qualificado	Frequentemente tácito
Facilmente transferível	Exige necessariamente a mediação humana	De difícil transferência

Fonte: Fundamentos em Ciência da Informação, disponível em http://lillianalvares.fci.unb.br/ciencia-da-informacao, acesso em 22/02/2023

Baseado nas fundamentações teóricas, depreende-se que dado é um fato ou acontecimento registrado, quantificável, transferível, processado em grande volume por computadores. Enquanto a informação, seria a agregação, contextualização, interpretação e análise humana, influenciada pelos modelos mentais (discurso histórico, biologia, genética, fisiologia, crença, valores etc.), seu significado depende de interação, negociação e consenso. Já o conhecimento, é a estruturação lógica da informação, vivenciada e experienciada, podendo ser interna (tácita, na mente das pessoas) ou externa (quando a pessoa explicita o seu conhecimento através da disponibilização de documentos, vídeos, normas, áudios etc.).

Nessa pesquisa, utilizamos muitos dados em forma de "termos" que possam identificar um assunto ou temática para compor um vocabulário controlado e construção de taxonomias na área da Ciência da Informação, além de dados e metadados de base científica para construção de um corpus para treino de um modelo automatizado. Em seguida, ao se coletar teses e dissertações, serão gerados novos metadados, relacionando-os às taxonomias dentro de um repositório digital, realizando a agregação destes dados e metadados (propiciando o surgimento de uma informação para o usuário, recuperar o recurso de informação através de buscas facetadas). Ao documentar todas as atividades desse processo através da tese, experienciando como foi executar e realizar as atividades, gera-se conhecimento que será compartilhado através da proposição de um *framework*, produto desta investigação. O intuito é disponibilizar o conhecimento para que seja adaptado e utilizado em outras temáticas, assuntos e campos de pesquisa.

2.2.2. Ciência da Informação

O *Oxford English Dictionary* (OED) fornece informações valiosas sobre etimologia do termo "ciência da informação" e como os autores a usam no decorrer dos anos, em diferentes disciplinas (ex.: busca de informação, sistemas de informação e serviços de informação). Nesse contexto, o termo ciência da informação foi registrado pela OED em 1958.

A Ciência da Informação é:

> [...] a disciplina que investiga as propriedades e o comportamento informacional, as forças que governam os fluxos de informação, e os significados do processamento da informação, visando à acessibilidade e a usabilidade ótima. A Ciência da Informação está preocupada com o corpo de conhecimentos relacionados à origem, coleção, organização, armazenamento, recuperação, interpretação, transmissão, transformação, e utilização da informação. Isto inclui a pesquisa sobre a

> representação da informação em ambos os sistemas, tanto naturais quanto artificiais, o uso de códigos para a transmissão (BORKO, 1968).

Brookes (1980) faz uma crítica sobre a ciência da informação na introdução do seu artigo, discorrendo que essa área não possui fundamentos teóricos, apenas fragmentos dispersos, sem integração e coerência, flutuando em um limbo filosófico. Entretanto, observando vários autores da CI, percebe-se que esse campo ainda está sendo delineado, construído e que há várias contribuições de pesquisadores ao longo da história contribuindo para consolidação dessa ciência. Brookes enfatiza que dentre todas as ciências sociais, a CI é a que se encontra intimamente preocupada com as interações entre os processos mentais e físicos ou entre os modos objetivos e subjetivos de pensamento. Para fundamentar o seu discurso, o autor realiza uma análise metafísica para a compreensão da CI, utilizando a teoria dos três mundos de Popper: Mundo 1 ou M1 sendo o físico, da matéria, palpável; o Mundo 2 ou M2 sendo o subjetivo, das ideias, pensamento humano, cognitivo, pelo qual o indivíduo experimenta o M1; e o Mundo 3 ou M3 sendo o objetivo ou simbólico, compartilhando o conhecimento individual, explicitando o conhecimento tácito, produto da mente humana, registrada nas línguas, artes, ciência, cultura, tecnologias. Brookes discorre que é justamente no M3 que a CI deve focar o seu campo de estudo no intuito de coletar e organizar os registros, além de estudar as interações do Mundo 2 com o Mundo 3, descrevendo-as e tendo de explicá-las. O autor finaliza seu estudo enfatizando que os cientistas da informação precisam reconhecer a oportunidade e aceitar a pesada responsabilidade de compreender melhor o Mundo 3 construído pelo Homem.

De acordo com Giddens (2007, p. 642), a definição de um campo de pesquisa como ciência necessita cumprir três critérios: uma metodologia, onde os pesquisadores desenvolvem métodos sistemáticos para investigar os fatos, explorar seus padrões e emitir opiniões após a realização da análise; evidências que possam ser objeto para fundamentação, fornecendo subsídios para sustentar ou rejeitar as teorias e hipóteses propostas; e por último revisar o estudo, confirmando ou refutando a informação, seja na comprovação da hipótese ou descartando-a e apoiando a pesquisa em novas evidências e argumentos.

Seracevic (1995b), discorre que a ciência da informação abordou problemas e métodos que foram utilizados para a proposição de soluções no decorrer do tempo e devido a essa característica, podemos compreendê-la como uma ciência. A argumentação de Seracevic vai ao encontro da proposta por Giddens, pois qualquer campo de pesquisa, segundo esse autor, ao utilizar métodos sistemáticos é considerado uma ciência.

Capurro e Hjørland (2003) discorrem em seu trabalho, que na década de 1950, surgiram estudos para definir a informação como uma ciência e seu papel interdisciplinar. O desenvolvimento da Tecnologia da Informação e seus impactos de nível mundial fizeram surgir o que ficou cunhado como sociedade da informação. A ciência da informação se voltou mais para os fenômenos de relevância e interpretação como aspectos básicos do conceito de informação.

Wang e Pontes (2007) discorrem que é necessário definir se ciência da informação é uma ciência, uma profissão ou uma disciplina. Após o exame de uma variedade de conceitos e história da CI, concluem que ela é de fato uma ciência. Para se chegar a essa conclusão, os autores discutem ao longo do artigo os conceitos de informação e da CI, observando com maiores lentes a evolução da CI como um campo científico. Concluem ao final da pesquisa que o uso de metodologias em ciência da informação a caracteriza como uma Ciência, corroborando com a interpretação de Giddens e Seracevic.

Vega-Almeida *et al.* (2009) fornece uma visão abrangente sobre os paradigmas, histórico e epistemológico da ciência da informação. Utilizam suas análises baseadas em uma metodologia qualitativa e dividem de forma cronológica três estágios de paradigmas: **físico** (1945-197? – possui seu argumento filosófico fundamentado no empirismo, racionalismo e positivismo. A informação é tida como mensagem ou sinais expressos por algoritmos e probabilidade, também é considerada externa, objetiva, tangível e mensurável), **cognitivo** (1980-199? – possui seu argumento filosófico fundamentado no cognitivismo e mentalismo. A informação deve refletir a percepção subjetiva de conhecimento e informação do usuário. Além disso, a informação é tida como algo subjetivo, aquela que afeta ou muda o estado da mente, o da mensagem é produzida pelo receptor) e **social** (199? – possui seu fundamento filosófico baseado no historicismo e construtivismo social. A informação é abrangente e envolve mensagens que são processadas cognitivamente considerando o contexto). A CI do ponto de vista do paradigma social é concebida a partir da sociologia da ciência, hermenêutica, semiótica e análise do discurso.

Robinson e Karamuftuoglu (2010) consideram em seu artigo a natureza da ciência da informação como disciplina e profissão. Eles se fundamentam na análise conceitual da literatura da CI e em perspectivas filosóficas. Esses autores argumentam que é discutível a natureza da CI, principalmente pela falta de consenso sobre seu conceito, além de seu objeto de estudo, sua relação com outras disciplinas e se há atuação exclusiva de profissionais da área. Os autores tentam compreender a CI sob duas perspectivas: cadeia de comunicação (processo dividido em seis etapas: criação, disseminação, organização, indexação, armazenamento e uso da

informação) e análise de domínio (identificação de objetos, operações e as relações entre os temas que os especialistas em uma determinada área do conhecimento percebem como importante). Discorrem também sobre os modelos quantitativos, abrangendo: a teoria matemática de Shannon, a teoria situacional, regras de dedução (geral → particular), regras de indução (particular → geral), regras de abdução (criação de hipóteses a partir de evidências incompletas). Eles ainda afirmam que o progresso pode ser promovido através do desenvolvimento de teorias gerais que são aplicáveis a vários domínios. No entanto, a abordagem no domínio analítico sugere que teorias específicas de um campo de estudo sejam construídas com base em metateorias generalistas, que são construídas sobre pressupostos filosóficos. Dessa forma, compreende-se que: as filosofias são a base para o desenvolvimento de paradigmas e metateorias, que fornecem subsídios para formulação de teorias, e por sua vez dará a sustentação aos artefatos e a prática (pesquisa aplicada).

Depreende-se dos textos estudados que a maioria dos autores concordam que a epistemologia da CI ainda está em processo de construção. Observa-se que muitas investigações estão produzindo trabalhos de qualidade em CI no Brasil e no mundo, utilizando justamente uma metodologia científica, proposição de hipóteses, evidências e observação de um determinado fenômeno da CI, o que corrobora com a compreensão dela como ciência pura e aplicada. Este trabalho realiza uma pesquisa exploratória na área da ciência da informação, seguindo um método com protocolo rigoroso de revisão sistemática, além do estabelecimento de um conjunto de atividades em um processo que pode ser adaptado e aplicado por outro pesquisador. Contribui para corroborar com o papel científico da CI em desenvolver pesquisas e ampliar suas áreas de investigação.

2.2.3. Organização, Representação e Recuperação da Informação

A organização, representação e recuperação da informação (ORI) está muito relacionado com a capacidade de guarda, armazenagem, transmissão, comunicação e obtenção da informação. Serão apresentadas definições de alguns autores que auxiliaram na evolução de compreensão histórica da ORI.

Bush (1945) realiza a proposição de desenvolver um computador analógico denominado Memex, com a finalidade de propiciar ao usuário a capacidade de armazenar e recuperar livros, arquivos, documentos e comunicações, tendo como base o microfilme. Esse pesquisador identificou a necessidade de organização da informação e a criação de uma teoria capaz de descrever essa ciência.

Mooers (1951) propõe a utilização de operações com álgebra booleana, tais como OR, AND e NOT, executando comandos para a recuperação da informação (RI). Esse pesquisador definiu o conceito de RI, definindo-a como "aspectos intelectuais da descrição da informação e suas especificações de busca, bem como qualquer sistema, técnica ou instrumento utilizado na operação". Desenvolveu o conceito de "termo único" ou "palavra-chave", designando o assunto que expressa o conteúdo do documento. Com as proposições de Mooers, posteriormente surgiu a criação de uma linguagem para indexação e manipulação de texto denominada TRAC, resultando em um tesauro documental. A Lei de Mooers descreve que um sistema de recuperação da informação "tenderá a não ser usado sempre que for mais penoso e incômodo para o usuário, ter informações do que não as ter". Isso significa que o mais importante é a qualidade da informação e quanto maior a relevância da informação recuperada, maior importância e essencial ela é para o usuário.

Taube, Gull e Wachtel (1952) desenvolveram um sistema de indexação com o objetivo de organizar e recuperar a informação de forma facilitada. Denominaram as partículas ou unidades de informação como unitermos, contribuindo da mesma forma que Mooers para a consolidação da ideia de um tesauro.

Egan e Shera (1952) definiram o termo "controle bibliográfico" relacionado a organização da informação, delineando uma teoria que abordasse a armazenagem e a recuperação da informação (classificação e indexação), sendo considerada uma base fundamental da ciência da informação.

Wiener (1954) foi o autor da disciplina de cibernética, propiciando a criação da ciência que enfatiza o controle, cognição e comunicação. Desenvolveu a teoria da cibernética, onde os computadores e os mecanismos que operam automaticamente pudessem ser desenvolvidos.

Luhn (1960) foi o primeiro autor que desenvolveu a indexação automática e o processamento de textos completos através de um sistema de palavras autorizadas através de um tesauro.

Percebe-se que nos estudos da ORI houve o estabelecimento de sinergias entre diversas ciências, dentre elas a de comunicação (envolvendo a cibernética, inteligência artificial, sistemas de organização e estudos de automação) e a recuperação da informação. Fica nítido esse entendimento com as pesquisas de Lancaster (1968), que dividiu a recuperação da informação em dois subsistemas: de entrada (seleção de documentos, indexação e vocabulário controlado) e de saída (busca, comparação e interação entre o usuário e o sistema).

Os aspectos teóricos da ORI auxiliam na compreensão da evolução histórica dos conceitos alinhadas à sua operacionalização através de sistemas de informação que propiciam a

armazenagem, classificação, indexação e a recuperação da informação. Nesse contexto, visto que as informações no ambiente digital crescem exponencialmente, faz-se necessário a implementação de soluções tecnológicas que propiciam a melhoria da organização e recuperação da informação pelo usuário.

Os repositórios digitais têm ampliado a capacidade de digitalização dos acervos físicos, propiciando seu acesso online e independente de proximidade geográfica. Essa característica tem propiciado a denominação do termo "sociedade da informação", transformando aquilo que é físico e analógico em recursos de informação digital.

2.2.4. Repositórios Digitais

Para melhor delimitar o conceito de **repositórios digitais**, apresenta-se duas definições, sendo a primeira definida pela Câmara Técnica de Documentos Eletrônicos do Conselho Nacional de Arquivos (CONARQ) e a segunda realizada por especialistas na área da Ciência da Informação, através de Nota Explicativa dentro do Tesauro Brasileiro em Ciência da Informação (TBCI):

> **Plataforma Tecnológica** que apoia o gerenciamento dos **materiais digitais**, pelo tempo que for necessário, e é formado por elementos de ***hardware*****,** ***software*** **e metadados**, bem como por uma infraestrutura organizacional e procedimentos normativos e técnicos (CONARQ/CTDE, 2020, p. 42). [grifo do autor]

> Mecanismos para administrar, armazenar e preservar **conteúdos informacionais em formato eletrônico**, e que podem ter como foco um assunto (repositórios temáticos) ou a produção científica de uma instituição (repositórios institucionais). Muitos permitem o acesso universal e gratuito a seus conteúdos, que variam de acordo com a política de cada instituição. São **coleções digitais de documentos** de interesse para a pesquisa científica e, no caso dos institucionais, representam a sua memória científica (PINHEIRO e FERREZ, 2014, p. 195). [grifo do autor]

Percebe-se nas palavras grifadas nas duas definições, que os "repositórios digitais" representam um papel importante de integrar em uma mesma solução, a operacionalização de elementos descritos na ciência da informação, na arquivologia, suportada pelos recursos computacionais associados à tecnologia.

De acordo com Pavão et *al.* (2015), as tecnologias de informação e comunicação (TIC) são cada dia mais utilizadas pelas instituições de ensino e pesquisa com o intuito de oferecer

informações sobre sua coleção de documentos, preservando seu conteúdo informacional em meio digital, fazendo uso dos repositórios institucionais. Esses autores enfatizam a importância da padronização, normalização e enriquecimento dos metadados para fortalecimento da qualidade dos registros nos repositórios digitais. Estes são atributos extremamente importantes que garantem a descrição e a identificação do documento, auxiliando na obtenção dos resultados em buscas executadas nos sistemas automatizados com o objetivo de aumentar a satisfação do usuário. Para isso, os metadados são uma ferramenta essencial para a melhoria da representação da informação.

Shintaku e Vidotti (2016) discorrem que a disponibilização das informações através da internet (web 2.0) nas últimas décadas, propiciou o advento da mudança do físico para o digital, surgindo assim, os ambientes de informação digital, que nesta tipologia se encontram os repositórios digitais.

Gusmão et *al.* (2017) discorre que no paradigma sócio-tecnológico contemporâneo, os indivíduos produzem informações e as ofertam em diversos ambientes de informação digital, tendo como objetivo a promoção da interação entre esses indivíduos, dos indivíduos com as instituições e entre as instituições.

Araújo, Maia e Vechiato (2018) discorrem que o termo "repositório" se origina da palavra em latim *repositorium* e significa "um local onde os objetos poderiam ser armazenados e coletados". Os autores enfatizam que os repositórios digitais surgiram como uma resposta espontânea às dificuldades e custos associados para divulgação dos periódicos científicos, pela evolução da TIC e da necessidade de armazenar e disseminar o patrimônio intelectual de várias instituições.

Cerrao e Castro (2018) afirmam que para utilizar os repositórios digitais de forma adequada e funcional, deve-se realizar ações de representação e descrição dos recursos informacionais de forma padronizada para auxiliar na busca e recuperação da informação (RI).

O Conarq/Ctde (2020, p. 25) definiu o termo documento digital como a "informação registrada, codificada em dígitos binários, acessível e interpretável por meio de sistema computacional" e, quando este é reconhecido e tratado como um documento arquivístico, seu conceito geral é denominado de documento arquivístico eletrônico. Em 2022, o Conarq publicou a 2ª versão do e-Arq Brasil, contendo o modelo de requisitos para sistemas informatizados de gestão arquivística de documentos. Essa preocupação se deve ao fato de que há no Brasil um crescimento de acervos de documentos digitais e que vem sendo tratado por diversos técnicos e especialistas de várias áreas, como a arquivologia a tecnologia da informação (CONARQ/CTDE, 2022, p. 10).

Martins, Lemos e Andrade (2021) afirmam que o ArXiv foi o primeiro repositório digital e foi implantado em 1991, mas esse tipo de plataforma ganhou impulso após os anos 2000, com a implementação de protocolos para coleta de metadados denominados de *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH). Esse protocolo será mais bem abordado nas seções subsequentes desse capítulo.

Em relação aos requisitos técnicos, um repositório digital armazena os dados em banco de dados relacionais, tais como *MySQL* e *PostGreSQL* e necessita de uma aplicação para organizar as coleções digitais e recuperar as informações.

Para delimitar o escopo de análise, foram selecionadas três ferramentas utilizadas para gestão e organização dos repositórios digitais: Omeka, Dspace e o Tainacan.

O **Omeka**[2] - Figura 4, é um *software* de publicação na *web*, possui código aberto, desenvolvido pela *Corporation for Digital Scholarship*, *Roy Rosenzweig Center for History and New Media e George Mason University*, tendo sido lançado no ano de 2008 e se tornado uma importante alternativa para museus, bibliotecas e arquivos como ferramenta de apoio a publicação de coleções digitais.

Figura 4 – Uso do Omeka na Biblioteca Graciliano Ramon.

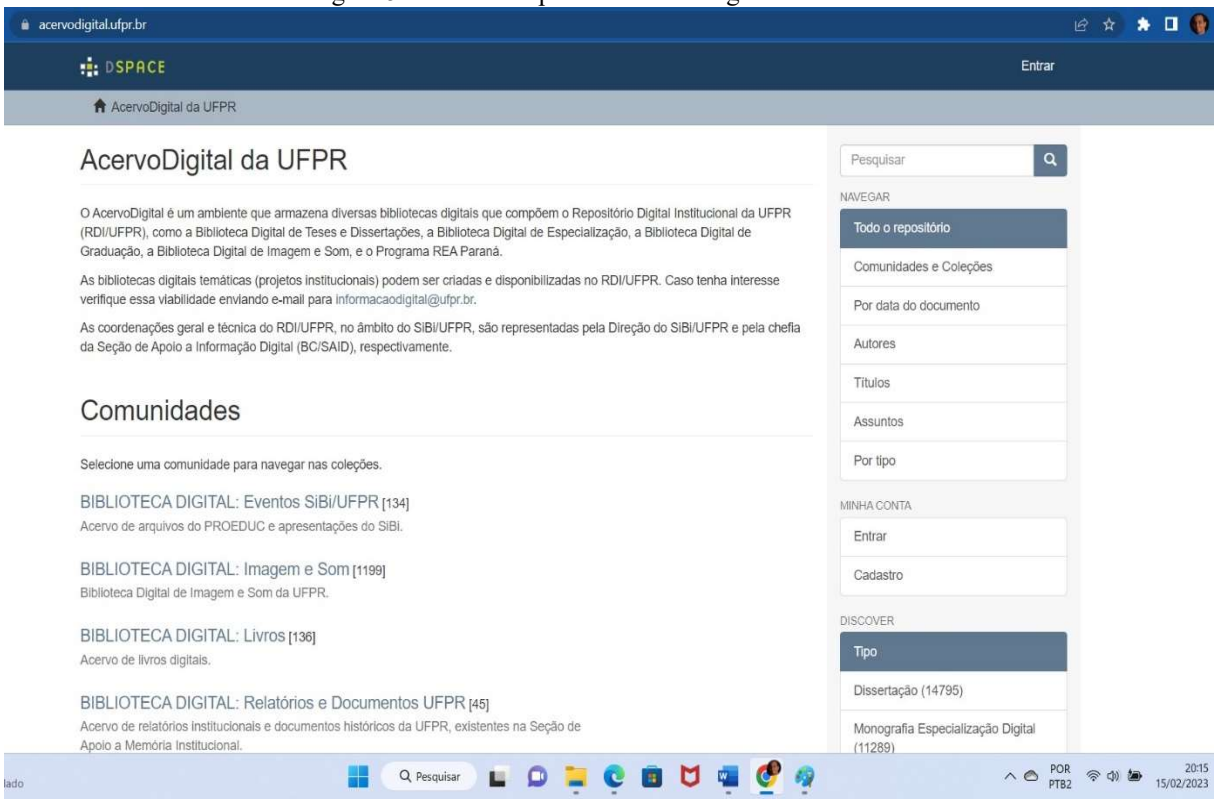
Fonte: ENAP, disponível em <https://biblioteca.enap.gov.br/index.php>, acesso em 22/02/2023

[2] Disponível em <https://omeka.org/>, acesso em 22/02/2023.

Atualmente o *software* possui duas categorias: OMEKA S para instituições, estando disponível na versão 4.0.0; e OMEKA Classic para projetos individuais e educacionais, na versão 3.1. Possibilita o compartilhamento de coleções digitais e acesso *online* e permite trabalhar com diversos tipos de mídias. O Instituto Brasileiro de Informação em Ciência e tecnologia (Ibict) e a Escola Nacional de Administração Pública (Enap) executaram um projeto de pesquisa que objetivou a melhoria do sistema de informação da Biblioteca Graciliano Ramos. Para isso, desenvolveram o Guia do Usuário e implantaram a solução Omeka durante o projeto. O Omeka é utilizado por várias instituições: Acervo Digital Regional de Extrema em Minas Gerais, Museu Ferroviário de Bauru, Centro de Memórias da Faculdade de Ciências Farmacêuticas da USP, Portal Interatrilhas criado pelo Museu de Ciências da USP, Festival de Dança de Joinville/SC, Fundação Dorina Nowill, Exposição Curitiba ontem e hoje, dentre vários outros.

O **Dspace**[3] - Figura 5, é um *software* desenvolvido para a plataforma *web*, possui um sistema de gerenciamento de conteúdo e de documentos, além de arquivos digitais e, foca no armazenamento de longo prazo, facilitando o acesso e a preservação dos acervos. É uma aplicação gratuita e de simples instalação, personalizável para atender particularidades de uma instituição.

Figura 5 – Uso do Dspace – Acervo Digital da UFPR.



Fonte: UFPR, disponível em <https://acervodigital.ufpr.br/ >, acesso em 22/02/2023

[3] Disponível em <https://dspace.lyrasis.org/>, acesso em 22/02/2023.

A primeira versão do Dspace foi lançada em 2002, através de esforço conjunto entre os programadores do *Massachusetts Institute of Technology* (MIT) e *Hewlett-Packard Labs*. Mais tarde foi fundado o *Dspace Foundation*, uma organização sem fins lucrativos que fornecia segurança e suporte, sendo atualmente realizado pela *DuraSpace* e *Lyrasis*. A aplicação encontra-se atualmente na versão 7.4. De acordo com Shintaku *et al.* (2018), a ferramenta Dspace não disponibiliza a funcionalidade de *streaming*, executa essa atividade incorporando outras soluções, o que pode fornecer um nível de complexidade ao projeto de implantação. Esses autores discorrem que a ferramenta apresenta coleções estáticas e não possibilita dar destaque a determinados documentos. Entretanto, muitas instituições a utilizam, tais como: Biblioteca Digital do Ministério da Justiça; Universidade Federal do Paraná, Universidade Mackenzie, Fundação Getúlio Vargas, Superior Tribunal Militar, dentre outros.

O **Tainacan**[4] - Figura 6, é um *software* livre, flexível para criação de acervos digitais em *WordPress*. O Tainacan foi desenvolvido pelo Laboratório de Inteligência de Redes da Universidade de Brasília, com apoio da Universidade Federal de Goiás, Instituto Brasileiro de Informação em Ciência e Tecnologia e do Instituto Brasileiro de Museus.

Figura 6 – Uso do Tainacan: Museu do Índio – Funai.

Fonte: FUNAI, disponível em <http://tainacan.museudoindio.gov.br/>, acesso em 22/02/2023

[4] Disponível em <https://tainacan.org/>, acesso em 22/02/2023.

O Tainacan auxilia na preservação, comunicação, gestão e compartilhamento dos acervos. Executa funções de catalogação, organização, armazenagem e compartilhamento de informações, além de customizar a criação de coleções, metadados, itens, filtros, possibilitando a parametrização de buscas facetadas.

Além dessas funcionalidades, o Tainacan permite gerenciar os vocabulários controlados através de parametrização de taxonomias. Outro fator importante do Tainacan é que ele implementa uma *Application Programming Interface* (API) RESTful, permitindo a interoperabilidade e interações de outras aplicações com o repositório configurado. As coleções podem ser expostas através de diferentes formatos: JSON, JsonLD, OAI-PMH, além de permitir mapear padrões de metadados como o *Dublin Core*. Diversas instituições utilizam o Tainacan, tais como: museus mantidos pelo Instituto Brasileiro de Museus (Museu Victor Meirelles, Museu do Índio da Funai, Museu do Diamante, Museu da Inconfidência, Museu Villa Lobos); Universidade de Brasília, Universidade Federal de Minas Gerais, Universidade Federal do Rio Grande do Sul, além de instituições privadas e do primeiro setor.

2.2.4.1. Estudos comparativos entre Omeka, DSpace e Tainacan

Estudos comparativos entre o *software* Omeka e Dspace foram realizados por Shintaku *et al.* (2018), apontando a solução Omeka como uma melhor escolha para implementação de projetos de repositórios digitais por conseguir implementar *streaming* e coleções dinâmicas.

Martins, Lemos e Andrade (2021), retratam em seu artigo a comparação do uso do *software* Tainacan e Omeka. Como resultado, apresentou-se que o *software* Omeka exige esforço 25% maior e conhecimentos especializados de TI para sua implementação, se comparado com o Tainacan. Os autores também discorrem que o Tainacan se alinha às tendências da web 2.0, propicia fácil instalação do *plugin* por usuários comuns e configuração facilitada de características relacionadas à Web Semântica, sendo uma delas a configuração de taxonomias e filtros, possibilitando a implementação de buscas facetadas.

Nesse contexto, a partir dos estudos prévios apresentados, a ferramenta de gestão e organização do repositório digital escolhido para criação das coleções a serem indexadas nesta pesquisa foi o **Tainacan**.

Compreende-se que os repositórios digitais são estruturas de organização da informação e possui relação direta com a busca e a recuperação da informação. Ressalta-se, portanto, a importância de investigações que enfatizem a representação da descrição dos recursos

informacionais, suportando de forma adequada a recuperação da informação e a satisfação dos usuários desses repositórios digitais.

2.2.4.2. Interoperabilidade dos repositórios digitais

Santarém Segundo, Silva e Martins (2019) discorrem que muitos acervos digitais foram idealizados de forma isolada, e com o passar dos anos há a idealização da integração desses vários repositórios. Os autores afirmam ainda que, a construção de novos acervos digitais, atualmente, contempla o requisito de interoperabilidade.

Os dados e informações estão armazenados e publicados em diversos repositórios, e bibliotecas digitais. Cada um desses elementos são soluções tecnológicas desenvolvidas em diferentes linguagens de programação, plataformas, arquiteturas e banco de dados. Muitas vezes é necessário que um sistema interaja com outro para realizar o consumo, a inclusão, a atualização e a criação de uma nova base de dados.

A interoperabilidade é definida na seção 3.1.5 da ISO/IEC 17788:2014, como a:

> Capacidade de dois ou mais sistemas ou aplicativos trocarem informações e usarem mutuamente as informações que foram trocadas.

No contexto da Ciência da Informação, a interoperabilidade é um requisito essencial para executar diversas atividades, tais como: o compartilhamento de dados e arquivos, geração de metadados e assuntos, a construção de taxonomias e vocabulários, além das atividades de classificação e indexação nos repositórios digitais.

Santarém Segundo, Silva e Martins (2019) realizaram uma reflexão extensa sobre as diversas possibilidades de interoperabilidade disponíveis para realizar integrações nos acervos digitais. Os autores dividiram os modos de operação dos protocolos em quatro categorias:

1) Agregação (exposição e coleta de metadados, ex: *Open Archives Initiative Protocol for Metadata Harvesting* – OAI-PMH e *Open Archives Initiative Object Reuse and Exchange – OAI-ORE*);

2) Sindicação (distribuição servidor-cliente, ex: *Really Simple Syndication* - RSS e (*Access to Memory* – ATOM);

3) Publicação (recebimento de depósito direto de outros acervos, ex: *Simple Web-service Offering Repository Deposit* – SWORD e *Atom Publishing Protocol* – AtomPUB); e

4) Busca Distribuída (recebimento de buscas federadas e repostas com resultados, ex: Z39.50 protocol, *Search and Retrieve URL* – SRU e *Search/Retrieve Web Service* – SRW).

Dentre os protocolos de interoperabilidade observados para coleta de metadados está o *Open Archives Initiative for Metadata Harvesting* (referido como OAI-PMH) que foi selecionado para essa pesquisa e por isso terá um maior detalhamento de sua estrutura. Esse protocolo possui modelo desenvolvido com a participação de dois atores, segundo Silveira *et al.* (2019):

- Provedores de dados: possui a responsabilidade de administrar os ambientes que implementam o OAI-PMH, com intuito de disponibilizar os metadados;

- Provedores de serviços: utilizadores dos metadados (*harvested*/coletados) através do OAI-PMH, com intuito de ofertar serviços de valor agregado.

A definição para *harverster* está relacionada com a ação de um cliente que envia solicitações OAI-PMH, objetivando coletar metadados das aplicações que exercem o papel de servidor para processar as requisições.

Em geral, as requisições se dividem em partes:

- *Identifier* (identifica um item no repositório);
- *ListMetadataFormats* (recupera formato de metadados);
- *ListSets* (recupera a estrutura de um conjunto de um repositório);
- *ListIdentifier* (recupera apenas cabeçalhos em vez de registros);
- *ListRecords* (coleta registros da aplicação), *GetRecords* (recupera registro de metadados individual).

Após a seleção do protocolo a ser utilizado para executar a interoperabilidade, basta realizar a seleção de ferramenta tecnológica que realize a sua implementação para executar

ações de integração de repositórios ou coleta de metadados através de métodos de importação e exportação de dados.

2.2.5. Vocabulário Controlado

Lancaster (2004) discorre que um vocabulário controlado é uma lista de termos autorizados com uma forma de estrutura semântica (significado), que controlam sinônimos, distinguem homógrafos e agrupam termos afins,

Segundo Harpring (2010, p. 14), um vocabulário controlado é uma ferramenta informacional que possui palavras e frases comumente utilizadas para se referir a uma ideia, características, pessoas, lugares, eventos, assuntos e diversos outros conceitos. Essa ferramenta permite executar a categorização, indexação e recuperação da informação de forma facilitada. A autora enfatiza ainda que enquanto os vocabulários controlados podem funcionar como padrões para valores de dados e serem referenciados em padrões de conteúdo de dados, sua construção deve observar padrões e critérios estabelecidos em conformidade com as normas nacionais e internacionais. Nesse contexto, se uma instituição deseja elaborar um vocabulário controlado para uma temática interna específica de sua organização, ela deve obedecer às normas, caso objetive no futuro realizar a interoperabilidade e integração com outros sistemas para executar a busca e recuperação da informação.

Hedden (2010) discorre que um vocabulário é controlado quando os usuários (catalogadores e indexadores) aplicam os termos da lista para sua área de escopo (o valor do metadado ou campo). A autora continua sua argumentação afirmando que é controlado também porque durante um processo de revisão dos termos dentro de um vocabulário, este pode modificar, crescer ou diminuir, sendo uma responsabilidade do editor do vocabulário controlado ou taxonomista e não de um usuário comum.

A seguir, descrevemos algumas normatizações internacionais para o desenvolvimento de vocabulários controlados:

- ANSI/NISO Z39.19-2005: Diretrizes para a construção, formatação e gerenciamento de vocabulários controlados monolíngue;

- BS 8723-1:2005, BS 8723-2:2005, BS 8723-3:2007, BS 8723-4:2007: Vocabulários Estruturados para Recuperação da Informação;

- ISO 2788:1986: Documentação – Diretrizes para o Estabelecimento e Desenvolvimento de um Tesauro Semântico Aplicado – *Thesa* Monolíngue;

- ISO 5964:1985: Documentação – Diretrizes para o Estabelecimento e Desenvolvimento de Tesauros Multilíngues;

No Brasil, o governo federal definiu em 2004, a Lista de Categorias do Governo (LCG), como uma lista que contemplava todos os assuntos relacionados com a atuação de Governo. Após dois anos, modificou-se para a Lista de Assuntos de Governo (LAG), com um foco em taxonomia de navegação. Em 2010 surgiu o Vocabulário Controlado de Governo Eletrônico (VCGE)[5], cuja expectativa é ser usado para classificar qualquer conteúdo de informação (documentos, bases de dados, mídia eletrônica, documentos em papel etc.) que não seja classificado outra forma mais específica de indexação. O VCGE se alinha com a *Classification of the Functions of Government* (COFOG) que é uma classificação feita pela Organização das Nações Unidas (ONU) para funções de governo. Entretanto, o VCGE possui pouca utilização, apenas instituições isoladas fazem o seu uso, visto que a última versão 2.1.0, data de 2016 e o grupo de discussão é pouco atuante.

De acordo com Harpring (2010, p.16) e Hedden (2010), um vocabulário controlado pode ser constituído das seguintes tipologias:

- Listas de títulos de assuntos;
- Listas ou arquivo de autoridade;
- Taxonomias; e
- Tesauros.

As subseções seguintes abordarão o conceito de cada uma das topologias elencadas.

2.2.5.1. Lista de títulos de assuntos (termos)

Segundo Hedden (2010), a lista de termos é a tipologia mais simplista de um vocabulário controlado, sendo utilizada na descrição de metadados, tipo de conteúdo, idioma, fonte, departamento, livro, artigo, documento, dentre outros.

Harpring (2010, p. 18), afirma que a lista de termos ou assunto são normalmente organizadas em ordem alfabética, possui referências cruzadas entre os termos preferidos, não

[5] Disponível em <https://www.gov.br/governodigital/pt-br/governanca-de-dados/vocabulario-controlado-do-governo-eletronico>, acesso em 15/02/2023.

preferidos e outros títulos relacionados. A finalidade dessas listas é descrever o assunto ou tópico dos textos e agrupá-los com textos que contém assuntos similares.

Os cabeçalhos de assuntos podem combinar vários conceitos em uma linha. Por exemplo: “crucifixo de ouro medieval” – combina um período (medieval), um material/elemento (ouro), tipo de trabalho/objeto (crucifixo de ouro) – em um mesmo título.

A lista de títulos de assuntos inclui listagem com subtítulos padronizados e podem combinar regras, tais como: localização geográfica, traços, parênteses, pontos, dois pontos e travessão (quando utilizar uma pontuação, seja em termos simples ou compostos).

Exemplifica-se através do Quadro 1, uma lista de títulos de assuntos:

Quadro 1 – Lista de títulos de assuntos.

Desfile de escolas de Samba – Brasil
Família dos caninos (mamíferos) – Coleções literárias
Itália. Grupo de Trabalho de Artes, Cultura e Gastronomia
Arquitetura Moderna – Brasília
História da televisão: movimentos e estilos
Desenvolvimento infantil
Crucifixo de ouro medieval
Portugal Descrição e viagens 1500-1600

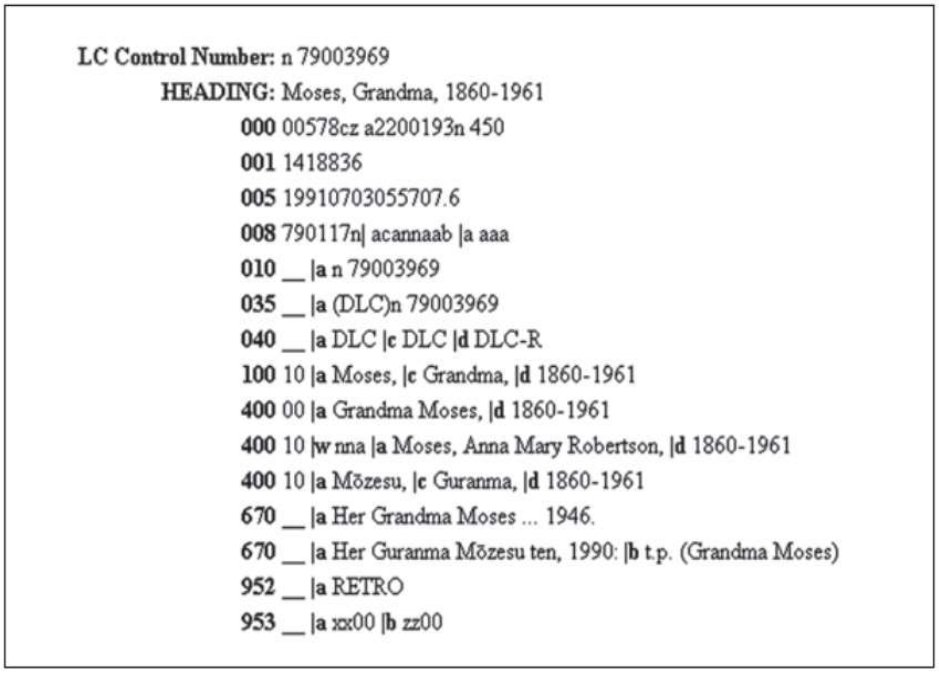
Fonte: adaptado de Harspring (2010)

2.2.5.2. Listas ou arquivo de autoridades

De acordo com Harpring (2010, p. 21), uma lista ou arquivo de autoridade é um conjunto de nomes ou cabeçalhos estabelecidos e referências cruzadas para termos similares. O controle da autoridade se refere tanto à metodologia adotada quanto ao vocabulário controlado específico e, se esse é aceito pela comunidade como uma autoridade ou for utilizado para dar consistência aos dados, exerce então nesse contexto o papel de autoridade.

Hedden (2010) discorre que o arquivo de autoridades inclui sinônimos ou variações para cada termo e estes funcionam como referência cruzada com a finalidade de apoiar e direcionar o usuário de uma variante de termo não preferido para um termo preferido equivalente. A Figura 7 demonstra uma lista de autoridades como exemplo:

Figura 7 – Lista de autoridade (Registro LCNAF para a Gandma Moses, ilustrando o título estabelecido e as referências cruzadas para esta artista)



Fonte: Harpring (2010, p. 22)

2.2.5.3. Taxonomias

Hedden (2010) discorre que o termo taxonomia significa a ciência de classificar as coisas, tornando-se um conceito popular para qualquer classificação hierárquica ou sistema de categorização. Dessa forma, a taxonomia é um vocabulário controlado, onde os termos são pertencentes a uma única estruturação hierarquizada, possuindo relações de termos pai/filho ou outras mais amplas/próximas com outros termos.

Harpring (2010, p. 22) afirma que a taxonomia pode ser compreendida como uma classificação ordenada de um domínio definido, conhecido também como um vocabulário facetado. Difere do tesauro, pois as hierarquias da taxonomia são mais superficiais e tem estrutura menos complicada, não possuindo termos equivalentes (sinônimos e variantes) e termos relacionados (relações associativas).

De acordo com Pontes e Lima (2012), as taxonomias são utilizadas na implementação de mecanismos de consulta, próximo às soluções de busca em portais institucionais e nas bibliotecas digitais. Discorre também que uma taxonomia auxilia na organização da informação, através da alocação, recuperação e comunicação dos conteúdos informacionais de forma lógica, através da navegação nesses ambientes. O Quadro 2 demonstra um exemplo sobre a taxonomia:

Quadro 2 – Recortes de Taxonomia de Direitos Humanos – Julho/2020

1.	**Integridade**
1.1.	Psíquica
1.1.1.	Alienação Parental
1.1.2.	Ameaça Coação
1.1.3.	Constrangimento
1.2.	Patrimonial
1.2.1.	Coletivo
1.2.2.	Cultural
2.	**Liberdade**
2.1.	Liberdade Laboral
2.2.	Liberdade ou direitos individuais
3.	**Vida**
3.1.	Aborto
3.2.	Automutilação
4.	**Direitos Sociais (Estado)**
5.	**Direitos Civis e Políticos (Estado)**
5.1.	Acesso à Informação
5.2.	Cultural

Fonte: Manual da Taxonomia de Direitos Humanos da ONDH[6]

2.2.5.4. Tesauros

O CONARQ/CTDE (2022, p. 50) define tesauro como:

> [...] uma lista controlada de termos ligados por meio de relações semânticas, hierárquicas, associativas ou de equivalência que cobre uma área específica do conhecimento. Em um tesauro, o significado do termo e as relações hierárquicas com outros termos são explicitados.

Na Figura 8, apresenta-se um exemplo da estrutura de um tesauro:

Figura 8 – Termos e estrutura de um Tesauro

acesso remoto
ING: remote access (UF offsite access)
ESP: acceso remoto
UP acesso a distância
TR ftp
telnet
World Wide Web
NE: A sistemas de informação - ASIST, p. 111.
CAT: 5.4 Redes de Comunicação e Informação, Internet, Web

acesso sem fio
ING: wireless access
TR World Wide Web
NE: A sistemas e serviços de informação - ASIST 138.
CAT: 5.4 Redes de Comunicação e Informação, Internet, Web

acesso universal
ING: universal access
TG acesso
TR censura
sociedade da informação
NE: Acesso à informação.
CAT: 6.2 Transferência e Acesso à Informação

acoplamento bibliográfico
ING: bibliographic coupling (UF citation coupling)
UP acoplamento de referências bibliográficas
TG análise de citação
TR citações bibliográficas
cocitação
índices de citações
CAT: 1.4.1 Metrias da informação e comunicação

acoplamento de referências bibliográficas
USE **acoplamento bibliográfico**
CAT: 1.4.1 Metrias da informação e comunicação

Fonte: Tesauro Brasileiro da Ciência da Informação (PINHEIRO E FERREZ, 2014)

[6] ONDH – Ouvidoria Nacional de Direitos Humanos, Disponível em < https://www.gov.br/mdh/pt-br/centrais-de-conteudo/publicacoes/ondh/manual-da-taxonomia-de-direitos-humanos-da-ondh.pdf>, acesso em 16/02/2023.

Harpring (2010, p. 24) discorre que um tesauro combina as características de uma lista de termos de sinônimos e taxonomias, podendo ser monolíngues e multilíngues. Discorrem também que os tesauros podem incluir informações periféricas ou explicativas adicionais sobre um determinado conceito, uma definição, nota de escopo e citações bibliográficas.

Em 2014, as professoras Lena Vania Pinheiro e Helena Dodd Ferrez participaram de um projeto no Laboratório de Pesquisa e Comunicação Científica, patrocinado pela Financiadora de Estudos e Projetos (FINEP). Dentre os subprojetos estipulados, havia a elaboração e publicação do Tesauro Brasileiro de Ciência da Informação (TBCI)[7]. As professoras enfatizaram que um tesauro é um instrumento para garantir a consistência da terminologia e do vocabulário que descreve uma área de conhecimento, auxiliando o trabalho dos profissionais que lidam com as linguagens documentárias. Essas podem ser conceituadas como o conjunto de termos descritos no texto e seus vínculos, utilizadas no processo de indexação e denota sua representação mental, sua imagem. Ao executar os algoritmos que utilizam técnicas de aprendizagem de máquina, as linguagens documentárias orientam os indexadores, fornecendo a compreensão dos termos dentro da lista controlada e estruturada, auxiliando na análise de assuntos e na recuperação de documentos e publicações nas mais variadas áreas de conhecimento humano. As professoras afirmam que o tesauro também se propõe a ser um "instrumento para a recuperação da informação de sua literatura em bibliotecas, bases de dados, repositórios e bibliotecas digitais, entre outros serviços e produtos de informação".

Os campos do Tesauros podem ser delimitados da seguinte forma:

- Descritor: termo escolhido para a representação de um conceito e que será usado na indexação e recuperação da informação. Se existirem outros termos que representam o mesmo conceito, utilizar a sigla USE antes do descritor;

- Não-descritor: termo que não é autorizado na indexação para evitar sinonímias. Antes do termo Não-descritor, utilizar a sigla UP;

- Categoria (CAT): Classificação em Grupo ou Facetas a qual o descritor pertence;

- Nota Explicativa (NE): Apresenta a definição do termo ou seu uso na indexação;

[7] Disponível em <https://www.gov.br/ibict/pt-br/central-de-conteudos/publicacoes/TESAUROCOMPLETOFINALCOMCAPA_24102014.pdf>, acesso em 16/02/2023.

- Termo Genérico (TG): Indica relação hierárquica entre os termos, sendo o descritor com conceito mais abrangente;

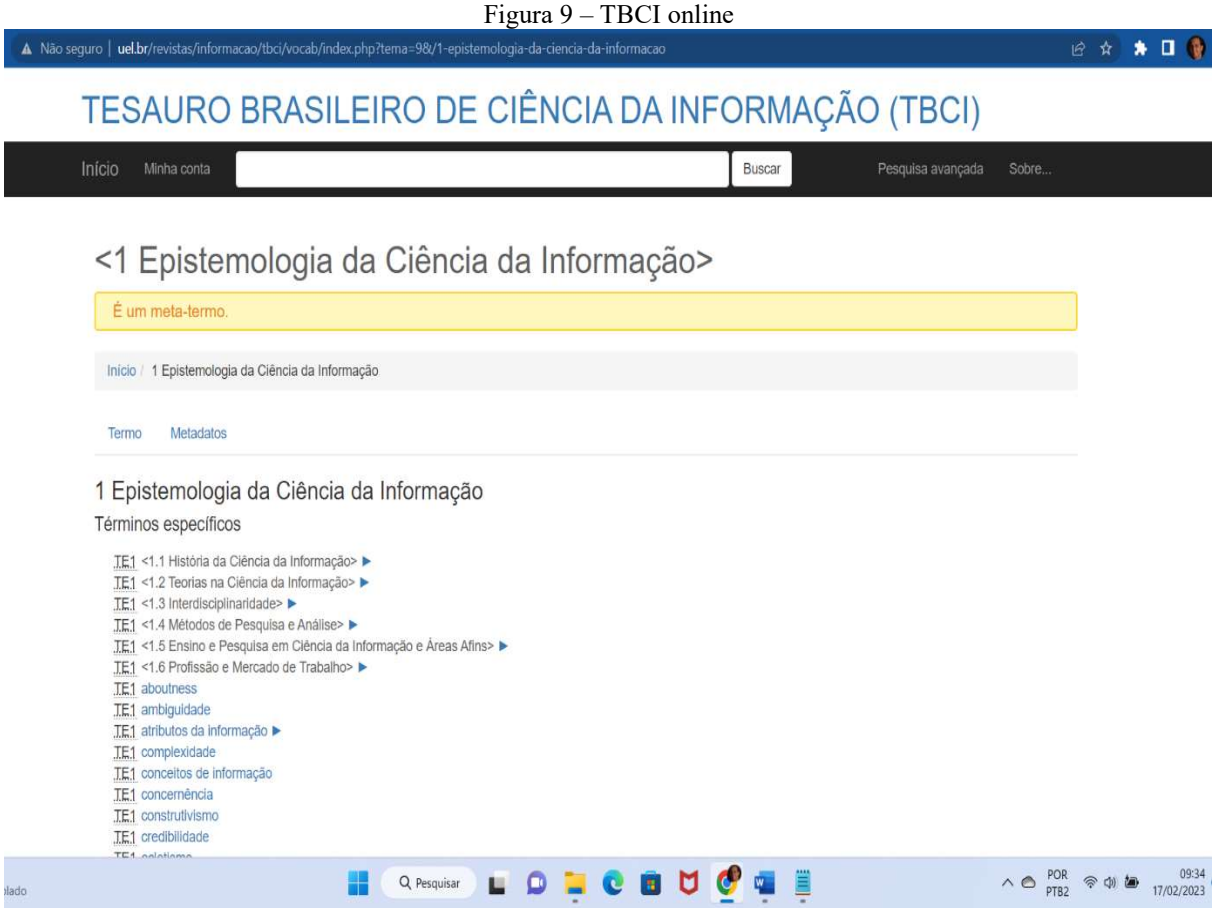
- Termo Específico (TE): Indica os termos específicos e subordinados ao termo genérico;

- Termo Relacionado (TR): Indica relação entre termos e não possuem hierarquia;

- Ao se utilizar termos em idiomas estrangeiros, utiliza-se as siglas ING para termos em inglês e ESP para termos em espanhol.

A pesquisa de Santos, Cervantes e Fujita (2018), apresentado no XIX ENANCIB, propôs a implementação do TBCI em formato eletrônico, realizando a importação na ferramenta TemaTres, disponibilizando o tesauro na *web,* conforme Figura 9:

Figura 9 – TBCI online



Fonte: TBCI (UEL), disponível em <http://www.uel.br/revistas/informacao/tbci/vocab/index.php>, acesso em 17/02/2023.

Para o TBCI online, conforme os autores, foram importados 2.058 termos, 1.828 relacionamento entre termos, 744 termos não preferidos e 336 notas de escopo. O TBCI está hospedado na Universidade Estadual de Londrina (UEL).

O uso de vocabulário controlado através de tesauros melhora o processo de indexação automática de documentos, pois auxilia as ferramentas computacionais no fornecimento de termos comumente utilizados.

2.2.6. Inteligência Artificial

O termo Inteligência Artificial (IA) tem ganhado muitas conotações, impulsionado pelo imaginário humano, muito como consequência da ficção científica e dos filmes futuristas.

De acordo com Cox (2023), a IA é mais bem compreendida como uma ideia de evolução e não como uma única tecnologia. O autor discorre que desde a década de 1950, a IA foi alvo de investigação e instanciada em tecnologias diferentes, passando por momentos de ápice, seguida por momentos de descrença e desconfiança. Entretanto, ele afirma que ultimamente, o termo IA é usado como um guarda-chuva no qual serve como um termo abrangente para diversas técnicas e soluções. O impacto da IA nas atividades de biblioteca será impulsionada por desenvolvimentos técnicos mais amplos que está além do controle dos bibliotecários. Todavia, a biblioteconomia já possui uma visão de que pesquisas apoiadas em IA não elimina a necessidade de alfabetização informacional.

Tarapanoff, Suaiden e Oliveira (2002) já demonstravam há vinte anos, o seguinte conceito para alfabetização informacional: “criar aprendizes ao longo da vida, pessoas capazes de encontrar, avaliar e usar a informação eficazmente para resolver problemas ou tomar decisões”. Esses autores enfatizam que uma pessoa que possui a competência em alfabetização da informação consegue: compreender sua necessidade informacional, organizá-la para que seja aplicada, realiza a junção da nova informação com o conhecimento vivenciado, utiliza a informação para desenvolver soluções, adquire a habilidade de aprendizado contínuo.

Retomando o discurso de Cox (2023), esse autor cita as principais áreas de aplicação dessas novas tecnologias, relacionadas com IA às bibliotecas:

- Pesquisas diárias na *web* em sistemas de bibliotecas existentes em interfaces de pesquisa, através de dispositivos móveis;

- Para descoberta de conhecimento, oferecer coleções de dados;
- Agentes de conversação e assistentes de voz;
- Análise de aprendizagem, bibliotecas e sentimentos;
- Automação de processos; e
- Bibliotecas inteligentes.

Com a transformação digital das bibliotecas e repositórios, os profissionais envolvidos nesses ambientes, trabalham atualmente, muito mais com mineração de dados e textos e utilizando recursos de IA, principalmente no que tange a aprendizagem de máquina para realização de análise de um volume gigantesco de dados. Essa abordagem propicia a construção de modelos que conseguem aprender os termos através de um vocabulário controlado e utilizar técnicas de sugestão de assuntos, criação de taxonomias, além de realizar a classificação, indexação nos ambientes digitais.

2.2.6.1. Corpus de Conhecimento

De acordo com Sinclair (2004):

> Um corpus é uma coleção de pedaços de texto linguístico em formato eletrônico, selecionados de acordo com critérios externos para representar, tanto quanto possível, uma língua ou variedade linguística como fonte de dados para pesquisa linguística (Sinclair, 2004). [tradução nossa]

A concepção de Sinclair se relaciona com a perspectiva linguística do corpus, onde os dados devem estar no formato eletrônico.

Aluísio e Almeida (2006), discorrem que a partir da década de 1990, o corpus ou corpora passaram a ter maior relevância nas pesquisas relacionadas à linguística, visto o desenvolvimento da computação e da linguística computacional, destacando ferramentas para o processamento de linguagem natural (PLN) no idioma português/Brasil e a grande contribuição dessas soluções para o processamento do corpus. As autoras complementam ainda, que por meio do corpus é possível analisar características morfológicas, sintáticas,

semânticas e discursivas; além de tornar claro o emprego de palavras, expressões e as formas gramaticais utilizadas, em suma, consta em detalhes a língua de forma objetiva.

2.3. ABORDAGENS DE SOLUÇÕES TECNOLÓGICAS À POSTERIORI

Na seção 2.1, executou-se a RSL e foi identificado um conjunto de técnicas e ferramentas específicas para geração automática e semiautomática de metadados. Entretanto, não foi possível selecionar dentre as soluções tecnológicas relatadas, aquelas passíveis de serem aplicadas a dados reais de um repositório digital, considerando as limitações apresentadas anteriormente na seção 2.1.4.4.

Realizou-se identificação à posteriori:

- **Ferramenta semiautomática de metadados** denominada ColetadorOAI[8], desenvolvida pelo Laboratório de Inteligência de Redes da Universidade de Brasília (UnB) em parceria com o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Essa solução é de código aberto, desenvolvida em *Python* e de fácil assimilação e implementação. Foi utilizada em projeto do IBICT com a Fundação Nacional de Artes (FUNARTE), no intuito de coleta, busca e recuperação da informação da produção científica implementando um repositório digital real sobre o mundo das artes no Brasil e experimentação de modelos de agregação de acervos digitais de artes de instituições de cultura ligadas ao Governo Federal;

- Após buscas com o termo "indexação automática", outras **ferramentas para geração automática de metadados** em evidência que estão sendo utilizadas nos acervos digitais de bibliotecas nos últimos anos. Pode-se citar a compilação de 10 sistemas de indexação automática executada por Corrêa e Lapa (2015), cuja pesquisa evidenciou as seguintes ferramentas com maior destaque: BIB/DIÁLOGO, SISA, PRECIS, OGMA; e outras soluções que foram utilizadas em estudos com menor frequência: SINTAGMED, ZSTATION, SRIAC, SPIRIT, MISTRAL e SIRILICO.

[8] Disponível em <https://github.com/tainacan/data_science/blob/master/FUNARTE/BIBLIOTECA_DIGITAL/ColetadorOAI-sickle.py>, acesso em 22/02/2022.

Silva, Correia e Gil-Leiva (2020) realizam uma análise comparativa entre os sistemas SISA e MAUI, obtendo bons resultados de precisão na indexação. Na literatura acadêmica é notório o discurso sobre a necessidade de ferramentas para indexação automática de metadados, visto o volume e variedade de dados produzidos ao longo dos anos sem a devida classificação e organização sobre seu contexto, o que dificulta sua indexação, busca e recuperação.

Realizou-se busca exploratória nos sites de várias bibliotecas que contemplassem investigações sobre a geração automática de metadados, sendo encontrado um relatório descrito por membros da Biblioteca Nacional da Holanda, intitulado "*Exploration possibilities Automated generation of metadata*"[9] e publicado em 2019.

Kleppe et *al.* (2019) descreve no relatório a dificuldade da Biblioteca da Nacional da Holanda em realizar a atribuição da descrição dos recursos de informação (conhecido como "geração de metadados" ou "criação de registros bibliográficos"). Essa problemática se deu em parte ao grande crescimento de material eletrônico gerado e armazenado, culminando na necessidade de otimizar a descrição dos recursos de informação realizados até então manualmente. Desta forma, esses autores demonstram o uso de tecnologias inteligentes para analisar e descrever fontes como artigos de notícias, livros, transmissões de televisão, fotografias com o uso de geração automática de metadados.

Kleppe et *al.* (2019) investigaram e utilizaram duas soluções:

- ANNIF, ferramenta existente e desenvolvida pela Biblioteca Nacional da Finlândia. Essa solução oferece módulos existentes para processamento de linguagem natural e aprendizagem de máquina, podendo ser combinado de diversas maneiras, além de ser *open source*. Pode utilizar o classificador *FastText* como um *backend* alternativo;

- ARIADNE, ferramenta desenvolvida pela OCLC (*Online Computer Library Center*), alcançou boas pontuações, mas os pesquisadores não sabiam o bastante sobre a metodologia utilizada no *background* e nem o material para treinar o sistema e isso se deve ao fato de a ferramenta não ser *open source*.

[9] Disponível em http://doi.org/10.5281/zenodo.3375192, acesso em 02/03/2022

O relatório forneceu indicadores interessantes sobre a solução ANNIF, *open source*, escrito em Python e todo o seu código está disponível no GitHUB, além de exemplos práticos de uso, comunidade de discussão ativa <annif-users@googlegroups.com>, artigos publicados discorrendo sobre a sua aplicação em grande conjunto de acervos digitais.

Os testes executados por Kleppe et *al.* (2019) demonstram que a ferramenta ANNIF é robusta, possui nível de qualidade adequada e pode ser customizada para diversas necessidades e aplicações, sendo essa ferramenta selecionada para análise desta tese.

2.4. ANNIF – FERRAMENTA DE INDEXAÇÃO AUTOMÁTICA ESTATÍSTICA

De acordo com Lappalainen et *al*. (2021) a ferramenta ANNIF tem despertado interesse em muitas organizações e a experiência das primeiras implementações na Biblioteca Nacional da Finlândia tem sido promissoras. O ponto de partida é a seleção adequada do vocabulário de assuntos adequado e um corpus de conhecimento para ensinar os modelos de aprendizado de máquina e a combinação de algoritmos para diferentes abordagens para obtenção dos melhores resultados.

ANNIF[10] é uma solução mantida pela Biblioteca Nacional da Finlândia de código aberto e baseada em microserviço. A ferramenta foi apresentada pelo Sr. Osma Suominen, especialista em sistemas de informação na Biblioteca Nacional da Finlândia na *47th LIBER Annual Conference* em 2018 na França, Sessão 10, com o título "*Annif: Feeding your subject indexing robot with bibliographic metadata*".

De acordo com Suominen (2018), o protótipo inicial foi desenvolvido em 2017 e vem sofrendo atualizações constantes e que estão sendo disponibilizadas no repositório do GitHub <https://github.com/NatLibFi/Annif>, sob a licença Apache 2.0. Esta solução foi concebida para executar a indexação automática de assuntos e classificação a partir de diferentes coleções de documentos, dentre artigos científicos, dissertações, livros antigos digitalizados, *e-books* e arquivos (SUOMINEN, 2018).

ANNIF é multilíngue e suporta qualquer vocabulário de assunto tanto em formato *Simple Knowledge Organization System* (SKOS) ou *Tab Separated Values* (TSV). É possível acessar sua interface através de linha de comando, formato *web* ou através de microserviço REST-API. Essa solução combina o uso de ferramentas de processamento de linguagem

[10] Zenodo DOI: https://doi.org/10.5281/zenodo.2578948

natural e aprendizagem de máquina, incluindo os algoritmos Maui, Omikuji, fastText e Gensim.

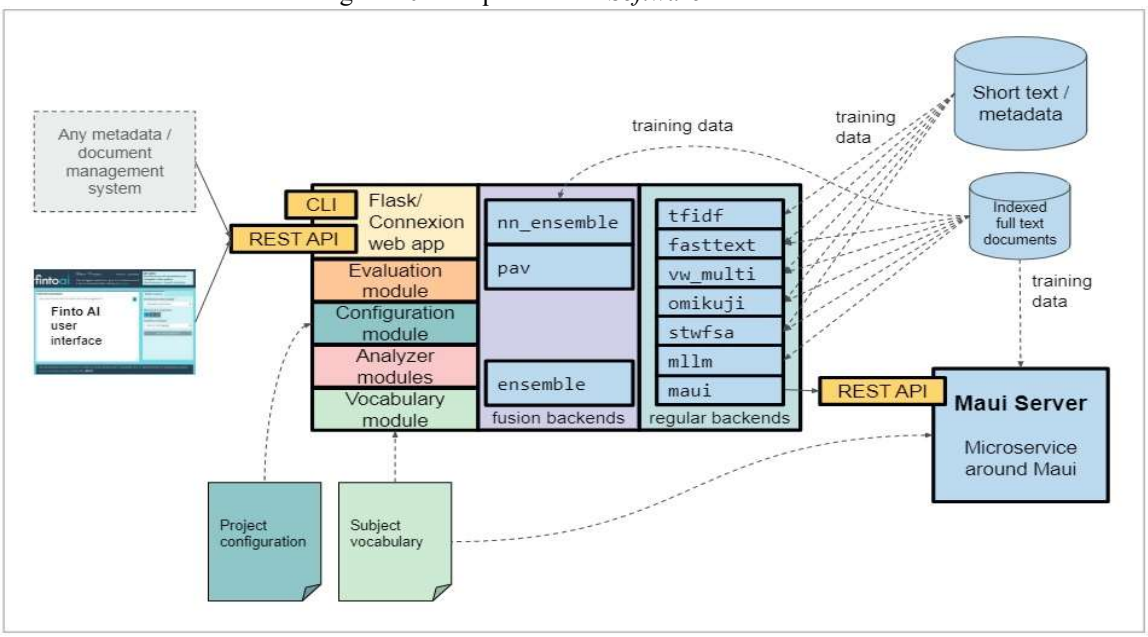
Suominen (2019, p. 21-22) destaca que “as funcionalidades de indexação são gerenciadas por diferentes algoritmos que podem ser usados separadamente ou combinados nos chamados conjuntos”. Dessa forma, cada algoritmo pode ser implementado como módulos separados e novos algoritmos podem ser adicionados posteriormente.

A instalação do ANNIF pode conter um ou vários projetos independentes, onde cada um deles especifica uma série de parametrizações, tais como o *backend*, o idioma e o vocabulário de indexação. Cada projeto é limitado a um único idioma, mas a indexação multilíngue poderá ser executada definindo vários projetos, sendo cada um deles por idioma.

Em cada projeto é definido um número, geralmente grande de assuntos, que espelham a ideia ou significado de um vocabulário de indexação. Os assuntos são criados a partir de um corpus[11] que é extraído dos registros de metadados presentes e/ou documentos indexados. No ANNIF poderá haver vários *backends* independents que auxiliam com sugestões de assuntos. Estes *backends* podem ser integrados ao ANNIF, ou serviços externos que poderão ser consultados via *Application Program Interface* (API), sendo que um projeto pode executar vários *backends* e combinar os seus resultados de análise.

A arquitetura de *software* do ANNIF pode ser visualizada através da Figura 10, conforme Suominen, Inkinen e Lehtinen (2022):

Figura 10 – Arquitetura de *Software* do ANNIF.



Fonte: Suominen, Inkinen e Lehtinen (2022)

[11] Coletânea ou conjunto de documentos sobre determinado tema – Oxford *Dictionaries*.

De acordo com Suominem (2019), o ANNIF possui uma aplicação principal que utiliza os *frameworks Flask* e *Connexion*, que fornecem uma API REST (quando o usuário utiliza o aplicativo através de uma interface web). A interface de linha de comando (CLI) suporta diversas funcionalidades de administração, tais como: gerenciamento do corpus, avaliação de qualidade e demais funções secundárias e auxiliares. A Figura 11 apresenta o acesso à ferramenta através de Rest API (online):

Figura 11 – Acesso ao ANNIF via Rest API.

Fonte: Annif, licença < https://zenodo.org/record/7553653>

ANNIF utiliza tipos diferentes de corpus de documentos e de assuntos, normalmente é utilizado um tesauro, uma classificação ou lista de cabeçalhos de assuntos. A ferramenta não se preocupa com a estrutura interna do vocabulário de assuntos, precisando apenas compreender as URIs e os rótulos de preferência (termos ou descritores) de cada um dos assuntos ou classes ou conceitos. O vocabulário de assuntos TSV simples especifica as URIs e os rótulos de conceitos, sendo um arquivo TSV com valores separados por tabulação,

codificado em UTF-8 e salvos com extensão *.tsv, onde a primeira coluna possui a URI e a segunda coluna o seu rótulo, como no exemplo abaixo:

<http://exemplo.xxx/tesauro/assunto1> inteligencia artificial
<http://exemplo.xxx/tesauro/assunto2> aprendizagem de máquina
<http://exemplo.xxx/tesauro/assunto3> pesquisa em linguagem natural

É possível também utilizar um vocabulário de assuntos como um arquivo SKOS/RDF, incluindo serializações comuns, tais como o RDF/XML, Turtle e N-Triples.

ANNIF pode utilizar um ou vários corpus de documentos para treinar modelos baseados em estatística ou aprendizagem de máquina e avalia o desempenho desses modelos. Uma curiosidade interessante é que possui suporte a dois formatos de corpus, sendo um mais adequado para aqueles documentos robustos ou compridos (textos completos ou resumos extensos) e outro adequado para textos pequenos, tais como o título ou palavras-chave de um artigo.

Em relação aos analisadores que são utilizados para pré-processar, *tokenizar* e normalizar o texto, o ANNIF possui três tipos (*snowball*, *spacy* e *voikko*), sendo que os dois primeiros suportam o idioma português.

2.4.1. *Backends*/Algoritmos suportados pelo ANNIF para indexação de assuntos

O ANNIF utiliza duas abordagens de *backends* (algoritmos) para execução das atividades de indexação de assuntos: algoritmos léxicos e associativos.

Os **algoritmos léxicos** combinam termos de um documento para termos contidos em um vocabulário controlado. Essa abordagem executa comparação utilizando poucos dados de treinamento. Exemplo de algoritmo: *Maui*, *YAKE*, *MLLM*, *STWFSA*.

Os **algoritmos associativos** aprendem quais conceitos estão correlacionados e com quais termos nos documentos, com base nos dados de treinamento. A abordagem associativa precisa de muito mais dados de treinamento para cobrir cada assunto. Exemplos de algoritmos: TF-IDF, *fastText* e *Vowpal Wabbit*.

2.4.2. *Backends* regulares para a indexação automática de assuntos e classificação

O ***Term Frequency – Inverse Document Frequency*** (TF-IDF) é baseado na hipótese de que um termo que não ocorre com frequência em geral (ou seja, em todo o corpus), mas

ocorre com frequência em um determinado documento do corpus, poderia indicar um assunto relevante para o conteúdo do documento. Conforme Suominen (2019), TF-IDF *similarity* é implementada com a biblioteca de código aberto *Gensim*, desenvolvida em *Python* e realiza a comparação de novos documentos com aqueles já conhecidos, sendo uma estatística numérica muito simples que pode ser usada para estabelecer uma linha de base onde os métodos de aprendizado de máquina mais avançados precisam suportar.

De acordo com Medelyan (2009, p. 7) ***Maui*** é um algoritmo de indexação automática construído para múltiplos propósitos e realiza três tipos de indexação: através de vocabulário controlado, utilização de termos da *Wikipedia* e marcação automática – em nível de desempenho similar aos dos seres humanos. Esse algoritmo foi escrito em Java 5.0 é de código aberto, distribuído sob a *GNU General License* e foi desenvolvido na Universidade de *Waikato*, na Nova Zelândia.

Suominen (2021) discorre que o algoritmo ***Maui-like Lexical Matching*** (MLLM) é uma reimplementação em *Python* de vários conceitos utilizados no *Maui*, com algumas adaptações. Esse algoritmo necessita de documentos longos de textos completos e, similar ao Maui, necessita ser treinado com um número relativamente pequeno (centenas ou milhares) de documentos indexados manualmente para que o algoritmo escolha a combinação correta de heurísticas que supra os melhores resultados em uma determinada coleção de documentos. Em comparação com o *Maui*, vários testes executados com o *MLLM* tiveram desempenho tão bom ou melhor em medidas de qualidades comuns (precisão, *recall*, F-*measures*, *NDCG*), mas diferentemente do *Maui*, o *MLLM* necessita de um vocabulário controlado.

O algoritmo ***STWFSA*** é um pacote em torno do *STWFSAPY*, sendo desenvolvido como parte do esforço da automatização da indexação de assuntos (AutoSE[12]) na ZBW – *Leibniz Information Centre for Economics* (Hamburg/Kiel, na Alemanha). Realiza a indexação de assuntos baseado em soluções de aprendizagem de máquina de código aberto, combinando vários métodos associativos e léxicos em uma abordagem de fusão, atingindo nível de desempenho superior.

FastText é um *kit* de ferramentas que foi desenvolvido pela equipe de pesquisadores do Meta *Research*® para aprendizagem de representação de palavras e classificação de textos. De acordo com Bojanowski et *al.* (2016)[13], *fastText* combina alguns dos conceitos de processamento de linguagem natural e aprendizagem de máquina, incluindo a representação

[12] *Leibniz Information Centre for Economics Your partner for research and studies*, disponível em <https://www.zbw.eu/en/about-us/key-activities/automated-subject-indexing>, acesso em 18/08/2022
[13] **fastText**, disponível em <https://research.facebook.com/blog/2016/8/fasttext/>, acesso em 18/08/2022

de sentenças através de um pacote de palavras (*bag-of-words*) e pacote de n-gramas, emprega a técnica de softmax hierárquico, aproveitando a distribuição desequilibrada das classes para acelerar a computação. Esses autores enfatizam que estes conceitos citados estão sendo utilizados para as seguintes tarefas: classificação de texto eficiente e aprendizagem de representações de vetores de palavras.

De acordo com Suominen (2019) o ***Vowpal Wabbit*** é um modelo de aprendizagem de máquina de propósito geral, originalmente criado pelo *Yahoo! Research* e seu desenvolvimento continua através da *Microsoft Research*. Essa estrutura é um pacote em torno de vários algoritmos *Vowpal Wabbit* para execução de classificação multi-classe e multi-rótulo, sendo pouco avaliado, mas se apresenta adequado para classificação de vocabulários pequenos (menor que 1.000 classes/assuntos).

Conforme Khandagale, Xiao e Babbar (2019), o algoritmo ***Bonsai*** utiliza a técnica *Extreme Multi-Label Classification* (XMC) onde o aprendizado supervisionado de um classificador pode rotular automaticamente uma instância com um subconjunto de rótulos relevantes e um conjunto extremamente grande de todos os rótulos de destino possíveis. Código desenvolvido em C++, adaptado do código fonte dos autores do algoritmo ***Parabel***: *Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Aversiting*.

Suominen (2020) discorre que o *backend* ***Omikuji*** é uma implementação bastante eficiente de uma família de algoritmos de aprendizagem de máquina, baseados em árvore. Pode emular os algoritmos *Parabel* e *Bonsai* para classificação multi-rótulo extrema. A exigência de processamento é intensa ao utilizar o *Omikuji*, sendo que por padrão utiliza todos os recursos de CPU disponíveis em paralelo e pode exigir também alocação de grandes quantidades de memória RAM durante o treinamento.

Suominen (2021) afirma que o pacote ***YAKE*** executa a extração automática de palavras-chave não supervisionadas, sendo pesquisadas a partir de um índice que é formado a partir dos rótulos do vocabulário SKOS. Esse autor afirma que YAKE não funciona tão bem quanto aos outros *backends* lexicais descritos anteriormente e as palavras-chave não encontradas no vocabulário são mostradas no log de depuração. Entretanto, a abordagem não supervisionada pode ser útil quando não exigir dados de treinamento.

2.4.3. Conjunto/Fusão – *Backends* que combinam resultados de outros *Backends*

Um dos grandes diferenciais e funcionalidades do ANNIF é a execução da combinação ou fusão dos resultados de vários algoritmos, possibilitando o uso do que há de melhor de cada um dos *backends* utilizados, fornecendo melhor qualidade na indexação automática de assuntos (LAPPALAINEN *et al*., 2021).

De acordo com Suominen (2021), o ***backend ensemble*** deve ser configurado com projetos de origem, tais como os *backends* TF-IDF ou MMLM, sendo ainda possível a parametrização de pesos para cada um desses *backends*. As solicitações para sugestão de assuntos são encaminhadas para os projetos de origem e posteriormente combinados, calculando a média das pontuações devolvidas por cada *backend* de origem para cada conceito. Esse autor informa que o uso do *backend ensemble* é facilitado devido não requerer configuração específica do algoritmo, além da parametrização das fontes.

O ***backend PAV*** implementa um conjunto dinâmico treinável e combina os resultados de vários projetos de maneira inteligente (SUOMINEN, 2021). As solicitações para sugestão de assuntos para o conjunto de *backends* são redirecionados para os projetos de origem e os seus resultados são ponderados novamente, mas usando uma regressão isotônica que tenta converter as pontuações brutas em probabilidades. Essa regressão é uma implementação do algoritmo *PAV* que está disponível na biblioteca *scikit-learn*, sendo realizada separadamente para cada conceito e os resultados são combinados calculando a média das pontuações regredidas, ou seja, probabilidades estimadas para cada conceito. É possível atribuir pesos para os resultados de cada *backend*.

O *backend* ***nn_ensemble***, conforme Suominen (2021), redireciona as solicitações de sugestões de assuntos de *backend* para os projetos de origem, sendo ponderados através do uso de uma rede neural *Keras* e *TensorFlow* 2 e suporta aprendizado online. A configuração dos nós determina o tamanho da rede neural e quanto maiores, ocupam mais memória, mas fornecem melhores resultados. São necessárias algumas experimentações para encontrar os parâmetros e a calibragem ideal.

2.4.4. *Backends* especiais

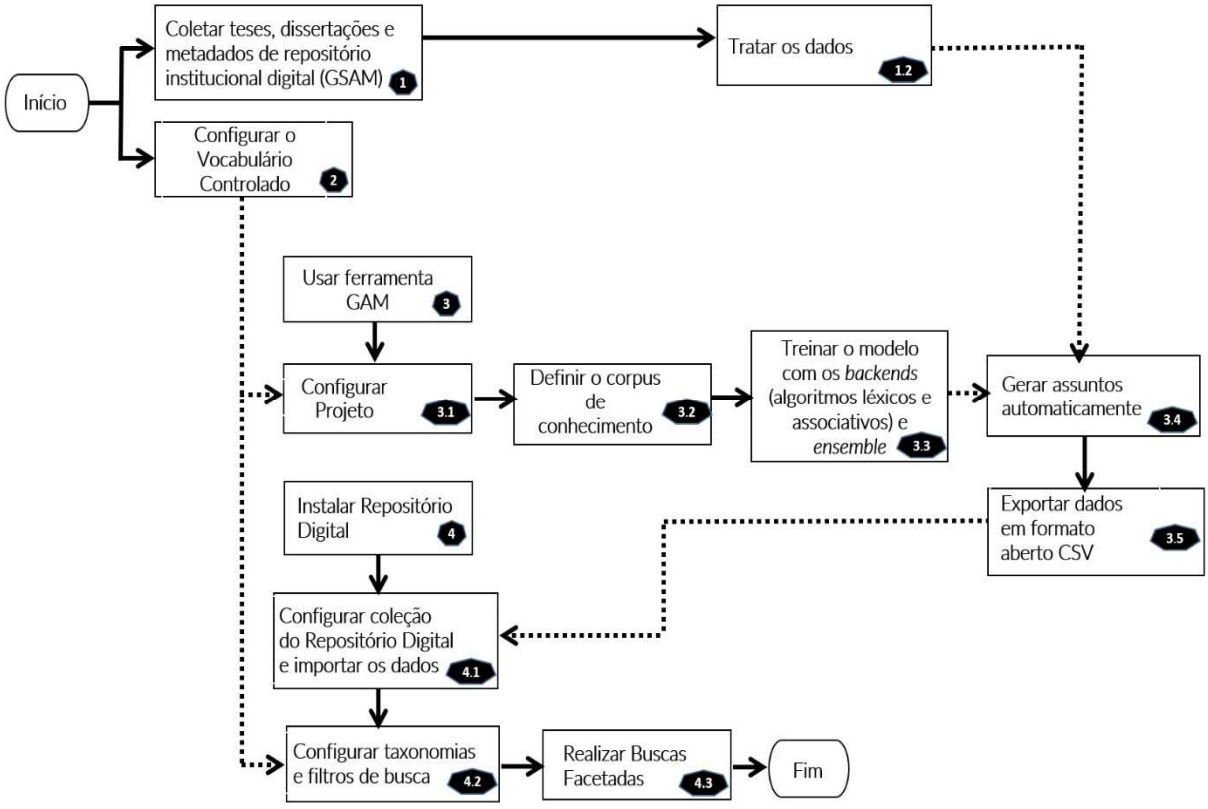
O *backend* ***HTTP*** se comunica com uma API REST que fornece um método de sugestão, podendo ser uma instância do próprio ANNIF ou serviço, tal qual o *MauiServer*. A configuração do endpoint especifica uma URL onde as solicitações de assunto são postadas.

O *backend* ***Dummy*** sempre retorna a mesma URI, sendo útil apenas para execução de testes de unidade.

2.5. *FRAMEWORK* GENÉRICO CONCEITUAL PROPOSTO

A partir das investigações realizadas e os conceitos apresentados neste capítulo de fundamentos teórico-metodológicos, foi proposto o *Framework* Genérico Conceitual, demonstrado através da Figura 12, que serviu como apoio para o estudo de caso executado e a seleção do repositório digital, da base de dados para compor o corpus de conhecimento, do vocabulário controlado, das ferramentas GAM e GSAM a serem utilizadas nessa pesquisa para geração automática de assuntos, indexação e busca facetada em repositórios digitais.

Figura 12 – *Framework* Genérico Conceitual proposto.



Fonte: o Autor

O *Framework* Genérico Conceitual proposto é um processo encadeado que contempla uma série de atividades para execução:

- **(1)** – Selecionar um repositório digital real para coleta de documentos e metadados através de uma ferramenta semiautomática para extração de metadados (GSAM);

- **(1.2)** – Tratar os dados, metadados e documentos obtidos no passo (1);

- **(2)** – Elaborar ou adequar um vocabulário controlado com lista de termos e assuntos relacionados com a temática em análise;

- **(3)** – Selecionar e instalar uma ferramenta de geração automática de metadados (GAM);

- **(3.1)** – Configurar ou parametrizar o projeto na ferramenta GAM;

- **(3.2)** – Definir o corpus de conhecimento (qual será a fonte de dados na temática em análise para treinar o modelo: ex. revistas, periódicos, base de dados);

- **(3.3)** – Treinar o modelo utilizando *backends*/algoritmos léxicos, associativos e em conjunto (*ensemble*). O objetivo é treinar o modelo, extraindo a eficácia que cada algoritmo pode fornecer a partir do corpus de conhecimento (3.2) e do vocabulário controlado desenvolvido (2);

- **(3.4)** – De posse das teses e dissertações colhidas e tratada no passo (1.2) e do modelo treinado com os algoritmos no passo (3.3), agora é gerar automaticamente a sugestão de assuntos para cada documento;

- **(3.5)** – Exportar os dados gerados no passo (3.4) em padrão aberto, preferencialmente com a extensão CSV;

- **(4) –** Selecionar e instalar a ferramenta tecnológica para implantação do repositório digital;

- **(4.1)** – Configurar coleção no repositório digital e importar os dados gerados no passo (3.5) para essa coleção;

- **(4.2)** – Configurar no repositório digital as taxonomias e filtros de busca, a partir do vocabulário controlado executado na atividade (2);
- **(4.3)** – Testar o repositório digital criado executando buscas facetadas através dos filtros configurados.

3. METODOLOGIA

A pesquisa procura atingir um determinado objetivo e, para que isso se concretize, é executada uma investigação científica que está sujeita a uma série de processos, atividades e procedimentos cognitivos e técnicos que denominamos de métodos científicos.

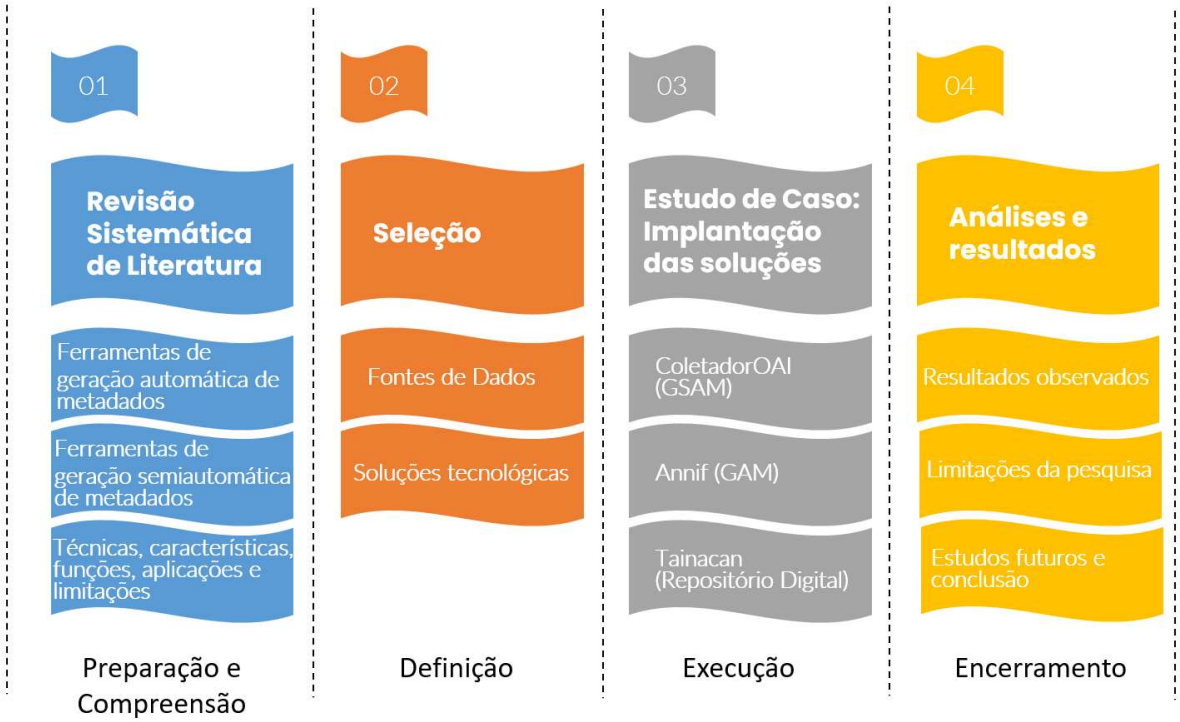
Todos os procedimentos realizados estão descritos através das seguintes seções:

a) Fases da pesquisa;
b) Classificação da pesquisa;
c) Definição do escopo da pesquisa;
d) Revisão Sistemática da Literatura (RSL);
e) Cronograma de atividades;

3.1. FASES DA PESQUISA

Para se atingir os resultados propostos desta pesquisa, elaborou-se a Figura 13, com o intuito de organizar a investigação, sendo composta de quatro fases: preparação e compreensão, definição, execução e encerramento.

Figura 13 – Fases da Pesquisa.
Validação do *framework* genérico proposto



Fonte: o Autor

3.2. CLASSIFICAÇÃO DA PESQUISA

Inicialmente, aplicou-se o método de revisão sistemática da literatura no capítulo de fundamentos teórico-metodológicos, considerada um estudo secundário, obtendo sua fonte de dados a partir de estudos primários (artigos, periódicos, *journals*, livros técnicos, de origem nacional e internacional), com intuito de identificar as ferramentas de geração automática e semiautomática de metadados, suas técnicas, características, funções, aplicações e limitações.

A técnica de pesquisa desta tese foi realizada por intermédio de método misto, utilizando procedimentos quantitativos e qualitativos.

A análise quantitativa contribuiu para o desenvolvimento de conhecimento (através do número de publicações analisadas, estudos correlatos, percentual de publicações obtidas nas bases de dados pesquisadas, adequação de quantitativos de termo/assuntos para vocabulário controlado e taxonomias, quantidade de registros para coleta de metadados e arquivos completos, geração automática de assuntos de grande volume de dados, número de registros para importação/exportação em formato aberto – interoperabilidade, indexação em repositório digital) empregando estratégias de pesquisa (desenvolvimento de estudo de caso, implantando e configurando as soluções tecnológicas selecionadas), resultando em dados estatísticos e informação.

A análise qualitativa baseia-se na compreensão da revisão sistemática da literatura realizada, além da interpretação e análise dos resultados, através das soluções tecnológicas implementadas, utilizando as fontes de dados selecionadas, apresentando as limitações e propondo sugestões de pesquisas futuras sobre o tema abordado.

É uma pesquisa aplicada, visando resolver um problema referenciado na questão de pesquisa: "**Como a técnica de geração automática/semiautomática de metadados pode apoiar usuários e gestores de repositórios digitais na melhoria da organização da informação, visando facilitar a busca e a recuperação da informação em seus acervos**".

É uma pesquisa descritiva, pois apresenta particularidades de uma população específica: gestores e utilizadores de repositórios digitais.

A estratégia de pesquisa se deu através de um estudo de caso. Essa abordagem visa observar e contextualizar os fenômenos da organização e recuperação da informação, identificando as lacunas e descrevendo processo e atividades, apoiada por tecnologias que implementem técnicas de inteligência artificial, capazes de implementar a geração automática de assuntos/metadados, indexar em repositório digital e propiciar a busca facilitada da informação. Yin (2015) afirma que caso o pesquisador pretenda investigar o como e o porquê

de uma série de fatos contemporâneos ocorre, o estudo de caso é apropriado. Esse autor enfatiza que essa estratégia de pesquisa é uma investigação empírica que permite o estudo de um fenômeno contemporâneo dentro de seu contexto da vida real.

A organização e recuperação da informação é um fenômeno que é estudado há décadas, mas o tema se torna contemporâneo quando se observa que nas últimas décadas, o volume e variedade da informação aumentou consideravelmente, sendo necessário mecanismos automatizados para auxiliar os gestores de repositórios digitais na organização e recuperação da informação nesses acervos.

3.3. ESCOPO DA PESQUISA

Essa investigação possui seu escopo delimitado nos seguintes aspectos:

a) Identificar e selecionar as ferramentas tecnológicas e descrever o processo e as atividades para sua implantação, a partir da proposição do *Framework* Genérico Conceitual na seção 2.5. O objetivo é melhorar a organização e recuperação da informação e que possa ser utilizado em qualquer área do conhecimento;

b) Executar um estudo de caso, aferindo sua efetividade para responder à questão-problema e validar a execução do *Framework* Genérico Conceitual proposto.

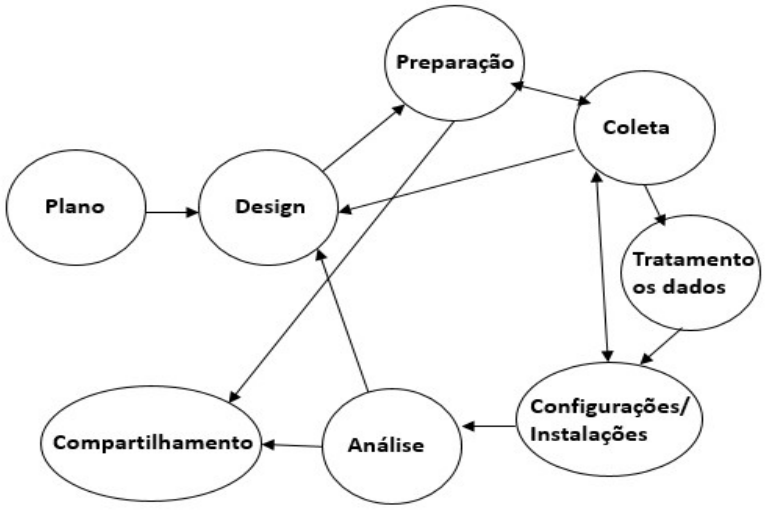
3.4. ESTUDO DE CASO

A ciência da informação, em termos institucionais, é considerada uma ciência social aplicada, conforme classificação de áreas de conhecimento elaborado e publicado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Sob o aspecto da estratégia metodológica, o estudo de caso é utilizado em diversas áreas, inclusive em pesquisas nas ciências sociais aplicadas. Conforme Yin (2015, p. 17), essa metodologia é válida, principalmente naquelas circunstâncias em que a questão de pesquisa a ser respondida é do tipo “como? ” ou “por que? ”, onde o investigador possui baixo controle sobre os acontecimentos onde permeiam os fenômenos complexos e contemporâneos, sob o contexto da vida real.

De acordo com Yin (2015, p.28), o estudo de caso deve ser planejado e seguir uma metodologia, com sequência de passos definidas e que serão aplicáveis à validação do estudo de caso, composto pelas seguintes etapas demonstradas através da Figura 14:

Figura 14 – Protocolo para estudo através de estudo de caso.



Fonte: adaptado de Yin (2015)

3.4.1. Plano

Conforme Yin (2015, p. 5), o pesquisador necessita identificar fato relevante para pesquisa, através de um estudo de caso. Para isso, deve responder as seguintes indagações:

- Como definir o "caso" que será investigado?
- Como determinar a relevância dos dados coletados?
- Após a coleta, o que fazer com os dados?

A questão de pesquisa elaborada para esta tese foi: "Como a técnica de geração automática/semiautomática de metadados pode apoiar usuários e gestores de repositórios digitais na melhoria da organização da informação, visando facilitar a busca e a recuperação da informação em seus acervos? "

Observando a questão elaborada, percebe-se que o "caso" definido no contexto desta tese será o uso de ferramentas de geração automática e semiautomática de metadados, suas

características, técnicas, funções, aplicações e limitações. Essas ferramentas podem contribuir para melhorar a organização e a recuperação da informação nos repositórios digitais.

Os dados obtidos através da revisão sistemática realizada, fornecem relevância alta, visto que foi elaborado um protocolo rígido para coleta, seleção, tratamento, análise e publicação dos dados. A partir dos estudos que embasaram a construção teórica foram gerados conhecimentos que propiciaram a fundamentação adequada para realização da pesquisa prática.

Finalizada as análises, os dados devem ser compartilhados através da comunicação científica, relatando e disseminando o projeto de pesquisa e o estudo de caso que validará o *Framework* Genérico Conceitual proposto nesta tese.

3.4.2. Design

Yin (2015, p. 28) discorre que a fase de projeto/*design* é o momento da definição da teoria, a verificação dos assuntos relacionados com a finalidade de direcionar o estudo de caso e propiciar a generalização dos resultados.

Na seção 1.4 desta tese, abordou-se na justificativa da pesquisa, problemas identificados no **Repositório Institucional da UnB (RiUNB)**, explicitados por Café e Muñoz (2016). O foco era a usabilidade dos usuários que passavam dificuldades ao recuperar a informação no repositório digital. Nesse contexto, selecionou-se o repositório da RiUNB, em especial, a comunidade de Pós-graduação da Faculdade da Ciência da Informação (FCI)[14] para ser o objeto de análise.

Para coletar os dados, metadados e documentos completos de dissertações e teses da RiUnB, selecionou-se a ferramenta GSAM denominada **ColetadorOAI** que implementa o protocolo OAI-PMH descrito na seção 2.3, pois é de fácil implementação e considerada uma solução simples e leve, além de ser uma ferramenta semiautomática de metadados, que implementa a técnica de extração de conteúdo/colheita de *metatags,* conceitos abordados nas seções 2.1.3.

Para adequação de um vocabulário controlado sob a temática da Ciência da Informação, escolheu-se o **Tesauro Brasileiro em Ciência da Informação (TBCI)**, contendo os termos e assuntos comumente utilizados na área abordada (seção 2.2.5.4). Esse

[14] Disponível em < https://repositorio.unb.br/handle/10482/5363>, acesso em 19/02/2023.

tesauro, também será utilizado para a construção das taxonomias para configuração dos filtros no repositório digital para propiciar as buscar facetadas.

O corpus de conhecimento utilizado para o treino do modelo com a temática da Ciência da Informação foi construído baseado em artigos científicos publicados na **Base de Dados em Ciência da Informação (BRAPCI)**.

A análise dos repositórios digitais abordados na seção 2.2.4, apontou para a seleção da solução **Tainacan**. Essa ferramenta foi instalada e configurada para receber a indexação dos dados e metadados, além de ser o *front-end* de acesso dos usuários, para realização das buscas e consultas facetadas da coleção disponibilizada.

E por fim, selecionou-se como ferramenta de indexação automática estatística, a solução **ANNIF**, conceituada na seção 2.4, para executar a geração automática de assuntos.

O Quadro 3, resume os elementos a serem utilizados neste estudo de caso:

Quadro 3 – Elementos utilizados no estudo de caso.

Elemento	Função	Atividade a ser executada
03 Máquinas Virtuais Ubuntu 64-bit em ambiente VirtualBox	Suportar os serviços de: coleta de metadados, sugestão automática de assuntos e repositório digital	Instalar o virtualizador e as três máquinas virtuais Ubuntu (Linux), além de parametrizar e configurá-las.
Tesauro Brasileiro da Ciência da Informação – TBCI	Vocabulário controlado, contendo lista de termos da Ciência da Informação	Adequar o vocabulário controlado para uso no ANNIF e criação das taxonomias no repositório digital Tainacan.
Repositório Institucional da Universidade de Brasília - RiUNB	Repositório com coleções de teses e dissertações da Pós-graduação em Ciência da Informação	Compreender as coleções, *datasets* e identificadores para serem carregados no coletador com a finalidade de extrair metadados de teses e dissertações.
Coletador	*Software* em phyton para coletar metadados na RiUNB	Instalar o *software* no servidor dedicado a este serviço, além das dependências de bibliotecas necessárias para seu funcionamento.
Base de Dados em Ciência da Informação – BRAPCI	Base de dados que fornece metadados de trabalhos científicos na área da Ciência da Informação	Adequar o corpus de conhecimento, levando-se em consideração o vocabulário controlado com os termos da CI.
ANNIF	Geração automática de assuntos das teses e dissertações coletadas	Instalar a ferramenta, configurar o projeto, carregar o vocabulário controlado, treinar modelos com o corpus de conhecimento e sugerir assuntos para os documentos a partir do modelo de treinamento.
Tainacan	Repositório digital que propicia a criação de taxonomias, coleções, filtros.	Instalar e configurar Wordpress, MySQL, PHP, servidor apache. Posteriormente, importar dados para a coleção, configurar as taxonomias e os filtros para disponibilização das buscas facetadas.

Fonte: o Autor.

3.4.3. Preparação

Conforme Yin (2015, p. 76), a preparação para o estudo de caso deve levar em consideração:

- As habilidades e valores: realizado através da seleção do tema, questão de pesquisa, objetivos e justificativa da pesquisa, definidos no Capítulo 1;

- O treinamento para o estudo de caso: realizado através dos fundamentos teórico-metodológicos, executado no Capítulos 2;

- O desenvolvimento de um protocolo de estudo: foi descrito na seção 3.4;

- A triagem dos candidatos ao caso: nesta pesquisa delimitou-se as ferramentas GAM e GSAM, além de atividades e repositório digital selecionado, descritas na seção anterior;

- A condução do estudo de caso piloto.

Para testar a ferramenta ANNIF que foi selecionada para geração automática de assuntos, realizou-se caso-piloto para aferir a funcionalidade de geração de termos e assuntos a partir do corpus de conhecimento e vocabulário controlado desenvolvido.

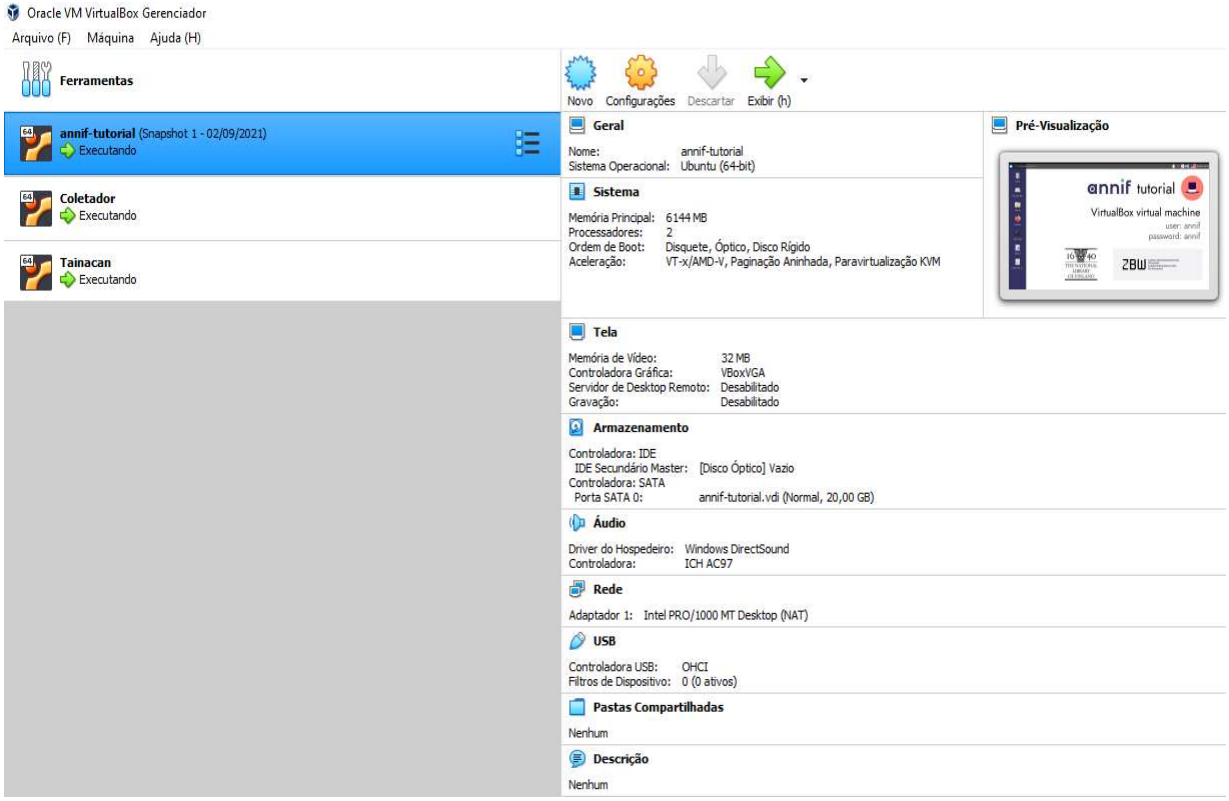
Utilizou-se o Tesauro da Ciência da Informação para adequar um vocabulário controlado para uso no ANNIF. Em seguida, gerou-se um corpus de conhecimento com 52 artigos da BRAPCI.

Após o processo de treinamento do modelo realizou-se teste preliminar de indexação automática estatística sobre uma Tese Completa armazenada no Repositório Institucional da Universidade de Brasília (RiUnB) gerando a recomendação de assuntos/descritores. Os termos atribuídos pelo ANNIF foram comparados com as palavras-chave da tese da RiUnB, obtendo boa similaridade.

Concluiu-se que o uso do ANNIF, utilizando a técnica de indexação automática estatística contribuiu para automatização da tarefa, obtendo desempenho satisfatório. A comunicação científica deste caso-piloto foi publicada no XXII ENANCIB[15].

Posteriormente, providenciou-se a infraestrutura tecnológica que sustentou os elementos deste estudo de caso, através da plataforma *Oracle VM VirtualBox*, criando-se três máquinas virtuais: Coletador (Coleta de dados, metadados e documentos), ANNIF (geração automática de assuntos) e Tainacan (repositório digital), conforme Figura 15:

Figura 15 – Plataforma tecnológica implementada para o Estudo de Caso.



Fonte: dados da Pesquisa.

As subseções seguintes demonstrarão a execução de cada atividade proposta no *Framework* Genérico Conceitual, abordado na seção 2.5.

3.4.4. Coletar teses, dissertações e metadados da RiUNB

O repositório institucional da Universidade de Brasília (RiUNB) é um serviço disponibilizado pela Biblioteca Central para disponibilização da produção científica da

[15] BRITO, J. C. B; MARTINS, D. L. Geração automática de metadados: estudo de caso utilizando a técnica de Indexação automática estatística com a ferramenta ANNIF. XXII Enancib, Porto Alegre, 07-11 novembro, 2022. Disponível em <https://enancib.ancib.org/index.php/enancib/xxiienancib/paper/viewFile/777/719>

universidade. O acervo digital é composto de 37 comunidades de interesse de diversas áreas de conhecimento. Entretanto, como a pesquisa a ser realizada nesta tese se relaciona a área da Ciência da Informação, será selecionada a comunidade "FCI – Programa de Pós-Graduação"[16]. As coleções disponíveis nessa comunidade são:

- FCI – Doutorado em Ciência da Informação (Teses) [216];
- FCI – Mestrado em Biblioteconomia e Documentação (Dissertações) [75];
- FCI – Mestrado em Ciência da Informação (Dissertações) [382].

Executou-se coleta de metadados e arquivos completos, com acesso aberto, das dissertações e teses das coleções, pois a RiUNB oferece um modelo de interoperabilidade, compartilhando o formato de coleta em um padrão de metadados que pode ser acessado via protocolo *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH). A ferramenta utilizada é o **ColetadorOAI**, desenvolvida pelo laboratório de inteligência de redes da UnB, que será adaptado para essa pesquisa.

O ColetadorOAI é desenvolvido em *Python* e está disponível no GitHub: https://github.com/tainacan/data_science/blob/master/FUNARTE/BIBLIOTECA_DIGITAL/ColetadorOAI-sickle.py.

Para utilização do código é necessário instalar as bibliotecas *Streamlit*, *Sickle* e iniciar uma sessão *Tmux* no *Linux Ubuntu*, conforme Quadro 4:

Quadro 4 – Instalação das bibliotecas *Streamlit*, *Sickle* e sessão no *Tmux*.

Instalar o pip
 # sudo apt install python3-pip -y
Instalar o *streamlit*
 # pip install streamlit
Instalar o *Sickle*
 # pip install sickle
Instalar o *tmux* no Ubuntu
$ *sudo apt-get update && sudo apt-get -y install tmux*
Para criar uma sessão tmux, executar o comando:
 # tmux new -s *StreamlitSession*
Rodar o ColetadorOAI no *Tmux* (localhost:8501)
 # streamlit run ColetadorOAI-sickle.py

Fonte: o Autor.

[16] FCI – Programa de Pós-Graduação, disponível em <https://repositorio.unb.br/handle/10482/5363>, acesso em 21/02/2023.

Provavelmente, caso não haja todos os pacotes dependentes para instalação das bibliotecas citadas, atualizar os pacotes conforme orientações explicitadas no *bash*.

Passo seguinte é identificar o nome do provedor, sua *Uniform Resource Locator* (URL) e o conjunto de itens da organização hierárquica do repositório digital. No site <https://repositorio.unb.br/oai/request?verb=ListSets> estão descritos todos os registros e identificadores do repositório digital da RiUnB, mas delimitou-se nessa investigação o conjunto "FCI – Programa de Pós-Graduação". Elaborou-se o seguinte arquivo CSV com as informações das coleções (*arquivo_coleta_unb.csv*), conforme Quadro 5:

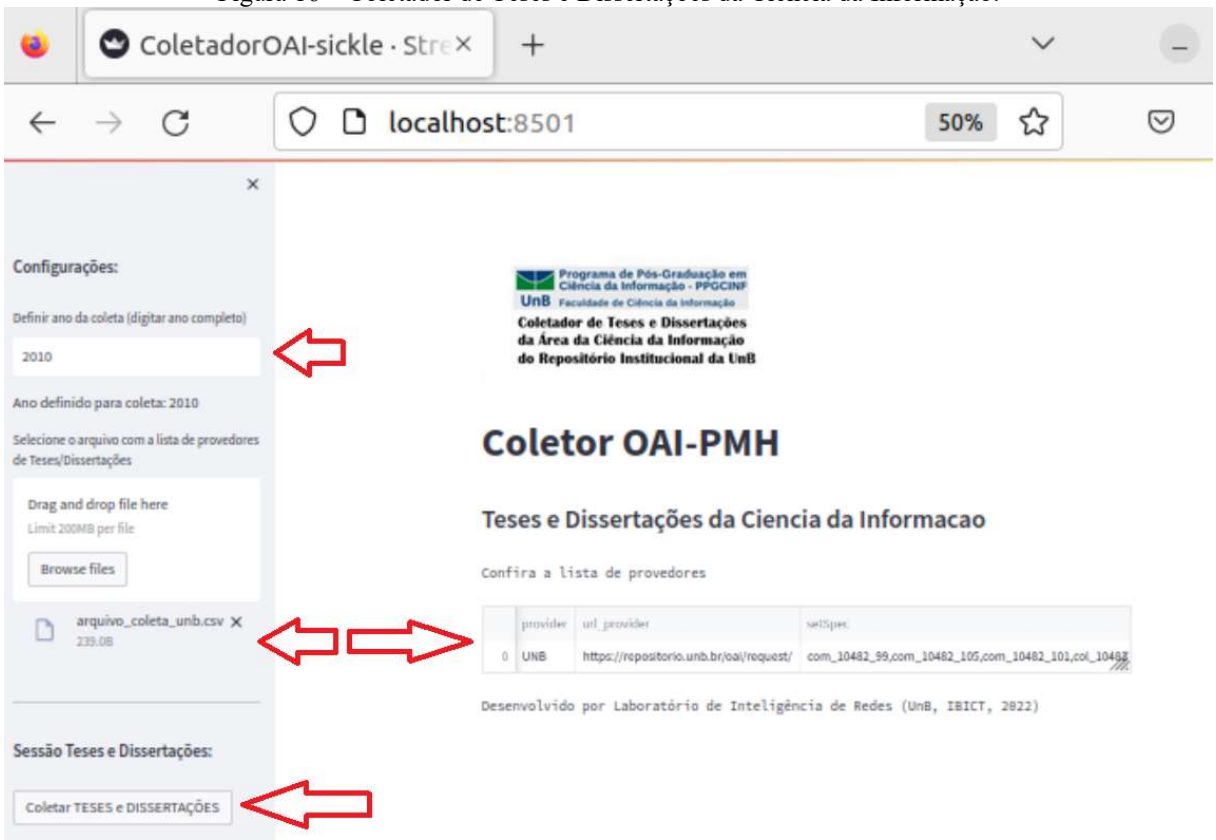
Quadro 5 – Parametrização do arquivo CSV a ser carregado no ColetadorOAI.

provider,url_provider,setSpec
UNB,https://repositorio.unb.br/oai/request/,"com_10482_1522,col_10482_1523,com_10482_4807,col_10482_5361,com_10482_5363,col_10482_5364,col_10482_5365,col_10482_5359,col_10482_5458,col_10482_11745,com_10482_20757,col_10482_20867"

Fonte: Elaborado pelo autor.

Após as configurações descritas, basta iniciar a aplicação "$ *streamlit run ColetadorOAI-sickle.py*" e carregar o arquivo "csv", conforme Figura 16:

Figura 16 – Coletador de Teses e Dissertações da Ciência da Informação.



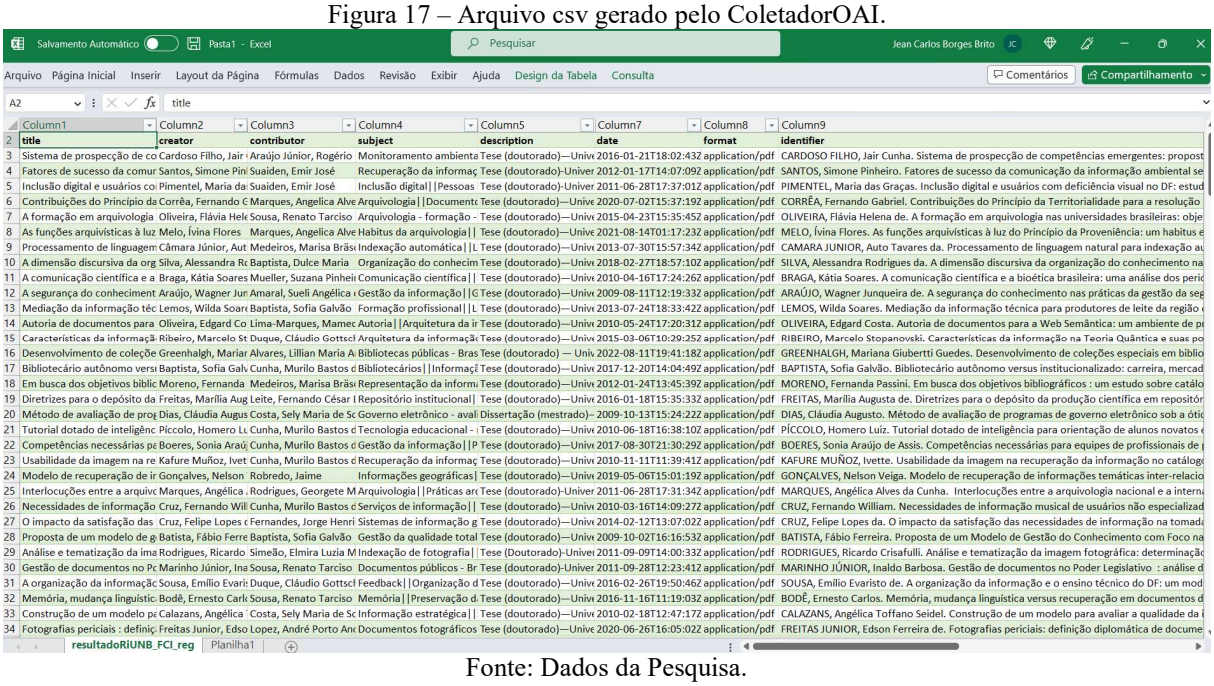
Fonte: Adaptado de Laboratório de Inteligência de Redes (UnB, IBICT, 2022).

A pesquisa realizada pelo ColetadorOAI é executada por ano e gera como saída um arquivo “csv” (*resultado_RiUNB_FCI_reg.csv*) com as seguintes colunas: *title, creator, contributor, subject. description, coverage, date, formate, identifier, language, provider, publisher, relation, rights, source, type e setSpec*. No momento da coleta, o RiUnB tinha armazenada em sua base de dados, o total de 660 pesquisas dos discentes de mestrado e doutorado em Ciência da Informação, além da biblioteconomia e documentação. Com os parâmetros configurados de *SetSpec*, obteve-se 258 registros (dados e metadados) de teses e dissertações, sendo que desse valor, apenas 85 possuíam o acesso aberto.

3.4.5. Tratar os dados

Os dados gerados pelo ColetadorOAI possuem como delimitador (separação das colunas, o sinal de vírgula) e o formato de codificação é o *Unicode* UTF-8. Nesse contexto, ao importar os dados, seja para o *Excel* ou *Jupiter Notebook* para realizar o tratamento e análise dos dados, deve-se realizar as parametrizações necessárias. Na Figura 17, apresenta-se os registros carregados na aplicação *Excel*.

Figura 17 – Arquivo csv gerado pelo ColetadorOAI.



Fonte: Dados da Pesquisa.

Caso seja encontrada alguma discrepância nos dados, esta é a fase do tratamento, limpeza e análise. Percebe-se que na coluna *subject* (assunto), *date* (data), *identifier* (identificador) e *rights* (direito autoral) são apresentados campos multivalorados (contém

mais de um dado no mesmo campo) e foi configurado para possuir o delimitador "||" entre os dados.

Os campos *publisher* (publicador), *relation* (relação), *coverage* (cobertura) e *source* (origem) apresentaram para todos os registros, o valor "Dado ausente no provedor".

Importante ressaltar que o ColetadorOAI coleta apenas os metadados e não realiza o *download* do arquivo completo da tese ou dissertação. Para executar essa ação, observou-se que no campo "*identifier*" contém o *Uniform Resource Identifier* **(**URI) para o registro na RiUNB, conforme exemplo do Quadro 6:

Exemplo: campo da coluna *identifier* multivalorado, contendo o endereço da URI após o delimitador "||":

Quadro 6 – Campo "*identifier"* multivalorado, contendo URI.

'SANTOS, Simone Pinheiro. Fatores de sucesso da comunicação da informação ambiental segundo especialistas. 2011. 160 p. Tese (Doutorado em Ciência da Informação) - Universidade de Brasília, Departamento de Ciência da Informação e Documentação, 2011.|| http://repositorio.unb.br/handle/10482/9853'

Fonte: Dados da pesquisa.

Elaborou-se um código em *python* para remover o (URI) do objeto "*identifier*" que direcionava ao recurso de informação da RiUnB, conforme o Quadro 7:

Quadro 7 – Obtendo a URI do objeto *identifier.*

```
# Importando a biblioteca pandas
import pandas as pd
# Importando a base de Dados obtida via Coletador OAI-PMH
RiUNB = pd.read_csv('resultadoRiUNB_FCI_reg.csv', encoding='utf8')
# Verificando as informações da base de dados
RiUNB.info()
# informar as colunas da base importada
RiUNB.columns
# Utilizando a biblioteca de expressões regulares para remover apenas a URI do objeto identifier
import re
for n in RiUNB.identifier:
    uri = [re.findall(r'http://repositorio.unb.br/handle/\d+/\d+', n)]
```

Fonte: Elaborado pelo autor.

Após a execução do código em *python*, será retornada uma lista com todas as URIs das dissertações e teses, sendo armazenado no arquivo "*uri.csv*".

Próximo passo é realizar o *download* dos arquivos e armazenar em diretório local, sendo que essa ação pode ser desenvolvida através de script via *bash* do Linux Ubuntu ou por código no próprio *Jupyter Notebook*, emulando o *bash* e executando o comando abaixo:

"$ *wget -nd -r -A pdf -i uri.csv -P artigosRiUnB/*"

Onde os parâmetros correspondem a:

- **-nd** = (*no directory*) não recriar no computador cliente a estrutura de diretórios, baixar apenas os arquivos no diretório local;

- **-r** = *download* recursivo;

- **-A pdf** = tipo de extensão do arquivo, nesse caso as teses e dissertações estão em "pdf";

- **-i uri.csv** = arquivo com as URIs para baixar, neste caso o arquivo "*uri.csv*";

- **-P *artigosRiUnB/*** = diretório onde os documentos serão salvos localmente.

Finalizado os *downloads* dos documentos completos de teses e dissertações do RiUNB, possuímos as informações dos dados, metadados e os arquivos em formato "pdf". Dos 85 registros em formato aberto, obteve-se *download* apenas de 26 documentos, ou seja 30%. Na execução do código para baixar os documentos, apresentou-se o erro "acesso não autorizado" para 59 documentos da RiUNB, considerada a primeira limitação encontrada. Apesar de estar demonstrado no portal do repositório que as publicações são abertas, nas configurações de pasta e de documentos exige credenciais de autenticação para baixá-las.

3.4.6. Configurar o vocabulário controlado

Na seção 2.2.5.4, abordou-se sobre o Tesauro Brasileiro da Ciência da Informação (TBCI) e na seção 3.4.2 (etapa de design do estudo de caso) executou-se sua seleção para servir de base para adequação de um vocabulário controlado. O TBCI conta com 2.058 termos, 1.828 relacionamentos entre termos, 744 termos não preferidos e 336 notas de escopo. O TBCI online está hospedado na Universidade Estadual de Londrina (UEL), através da URL <http://www.uel.br/revistas/informacao/tbci/vocab/>. Entretanto, o TBCI disponibilizado pelo IBICT está em formato pdf e o online pela UEL não disponibiliza a base em formato aberto, seja *SKOS*, *Turtle* ou *XML*.

Nesse contexto, após estudar o formato de vocabulário admitido pela ferramenta ANNIF, elaborou-se manualmente a adequação de um vocabulário controlado, baseado na TBCI com 439 termos da Ciência da Informação.

O arquivo "*tesauro-tbci.txt*" que contém o vocabulário controlado adaptado possui a estrutura "*URI_TBCI_ONLINE<tab>termo_tbci*", conforme Quadro 8:

Quadro 8 – Vocabulário controlado adaptado do TBCI com 439 termos da CI

http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=1323&/marc	MARC
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=1025&/inovacao	Inovação
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=1474&/redes-neurais	Redes Neurais
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=192&/ontologias	Ontologias
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=217&/semantica	Semântica
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=792&/gatekeepers	Gatekeepers
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=135&/aldeia-global	Aldeia Global
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=177&/bibliometria	Bibliometria
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=1118&/questionarios	Questionários
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=867&/entrevistas	Entrevistas
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=744&/educacao	Educação
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=805&/e-commerce	E-Commerce
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=1192&/encadernacao	Encadernação
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=1243&/hermeneutica	Hermenêutica
http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=1244&/historicismo	Historicismo

Fonte: Elaborado pelo autor.

Após a conclusão da elaboração do vocabulário controlado, gerou-se outro arquivo denominado "*taxonomia-ci.csv*", contendo apenas os 439 termos da CI (2ª coluna). Esse arquivo será carregado posteriormente na solução selecionada para implementar o repositório digital.

3.4.7. Instalar o ANNIF

A máquina virtual (VM) com a imagem do tutorial ANNIF, possui o Sistema Operacional *Ubuntu* (64-bits), Memória RAM de 6GB e 2 (dois) processadores. Por ser ambiente Linux, a configuração do ANNIF e a execução dos testes são facilitados para quem tem mais familiaridade com sistemas Unix/Linux.

O *download* da *VirtualBox* pode ser realizado através do *link* <https://annif.org/download/>. Essa máquina é para execução de testes e treinamento, não sendo adequada seu uso em servidores para ambiente de produção.

Ao subir a máquina virtual e iniciar o Terminal, será demonstrado o diretório padrão:

(annif-venv) annif@annif-tutorial:~/Annif-tutorial$

3.4.7.1. Configuração do Projeto

O ANNIF requer a configuração de um ou mais projetos para sua utilização. O projeto consiste em um conjunto de definições tais como: sua identificação (ID), descrição, linguagem/idioma, *backend*/algoritmo, vocabulário controlado, analisador. Nesta etapa serão realizadas as parametrizações de configuração no arquivo "*projects.cfg*" que podem ser visualizadas no exemplo a seguir:

[brapci]
name=Projeto Tesauro Brasileiro da Ciencia da Informacao
language=pt
backend=mllm
vocab=tesauro-tbci
analyzer=snowball(portuguese)

O vocabulário controlado de assuntos foi o arquivo "*tesauro-tbci.txt*", adequando os termos do Tesauro Brasileiro da Ciência da Informação (TBCI) com as URIs do TBCI online, devendo ser carregado ao projeto. É importante ressaltar que a maioria dos *backends*/algoritmos do ANNIF requerem alguns dados de treinamento. Os vocabulários, modelos e outros arquivos de dados são armazenados no diretório "data" por padrão, mas o caminho pode ser alterado através de variável de ambiente.

3.4.7.2. Definir *corpus* de conhecimento

Para o desenvolvimento de um projeto de corpus, Aluísio e Almeida (2006) citam os seguintes passos:

- **Seleção dos textos**: selecionou-se a Base de Dados de Pesquisa em Ciência da Informação (BRAPCI) como repositório, onde serão extraídos os metadados (título, resumo e palavras-chave) dos artigos publicados;

- **Compilação e manipulação do corpus**: selecionou-se apenas os artigos que tem relação com os termos gerados para elaboração do vocabulário controlado na seção 6.6.3. Utilizou-se o *software* Excel para compreensão dos dados e o *Jupyter Notebook* para manipulação dos dados;

- **Nomeação de arquivos e geração de cabeçalhos**: O arquivo final nomeou-se como "*corpus-brapci.xls*", contendo os seguintes dados de cabeçalho extraídos como metadados da Brapci: *author*, *title*, *source*, *year*, *keywords*, *abstract* e *link*. Adicionou-se duas colunas ao arquivo, sendo uma termotbci (termo do vocabulário controlado) e uri (link para termo da TBCI online);

- **Proteção de identidade ou pedido de direitos de uso dos textos**: Foi solicitado autorização de acesso aos dados para a Prof. Leilah Santiago Bufrem e o Prof. Rene Faustino Gabriel Junior (ambos da UFRGS), mantenedores da BRAPCI;

- **Anotação**: cabeçalho explicativo do metadado, tamanho do arquivo 2.150KB (anotação estrutural). Não houve a obtenção de textos completos dos artigos; as anotações linguísticas foram realizadas através de comparação do termo do vocabulário controlado com o título, resumo e palavras-chaves de cada artigo, referenciando o termo da TBCI em coluna específica na mesma tupla do registro analisado.

Uma limitação na atividade de desenvolvimento do corpus foi que o arquivo disponibilizado em 06/08/2022 pelo mantenedor da BRAPCI tinha o total de 46.686 registros. Entretanto, ao analisar os dados havia muitas colunas com valores de campo multivalorados e sem delimitador. Exemplo: registros da coluna *abstract*, que continham os dados referentes aos resumos dos artigos, tinham textos em português, inglês e espanhol, sem delimitador; registros da coluna *keyword* estavam sem delimitadores. Outro fator verificado foi que alguns metadados omitiam alguns termos de *keywords* ou esses não correspondiam aos descritos no arquivo do artigo científico. Neste caso, optou-se por realizar a construção manual do corpus com auxílio do *Excel* e *Jupyter Notebook*, culminando em arquivo com o total de 438 registros de dados de artigos da BRAPCI, referenciadas aos termos de assuntos do vocabulário controlado elaborado em passos anteriores.

Após a finalizar o corpus "*metadadosBRAPCI_to_corpus.xls"*, construiu-se arquivo em formato "tsv" com os campos: título. resumo. palavras-chave.<tab>uri_tbci.

No Quadro 9, demonstra-se a representação do registro do corpus para o artigo: SANTOS, J. C. Gestão documental e gestão da Informação: abordagens, modelos e etapas. Informação@Profissões; v. 10, n º1, p. 99-120, 2021.

Quadro 9 – Exemplo de registro de corpus

Gestão documental e gestão da informação abordagens, modelos e etapas. Os processos de Gestão Documental e Gestão da Informação possuem propósitos que visam a organicidade dos fluxos formais de informação visando a competitividade organizacional Objetivo: apresentar as abordagens, modelos e etapas mais citados na literatura de CI, buscando evidenciar a relevância desses processos de gestão para área e para as organizações empresariais. Metodologia: se trata de um estudo teórico, de natureza qualitativa, tipologicamente descritivo e exploratório. Resultados: a partir dar abordagens, modelos e etapas dos processos de Gestão Documental e Gestão da Informação, evidencia-se que ambos os processos são complementares e fundamentais para competitividade organizacional. Conclusões: considera-se que os estudos contribuem e sistematizam as abordagens, modelos e etapas mais citados pela literatura da área de Ciência da Informação no que tange os processos de Gestão Documental e Gestão da Informação. Ciência da Informação. Gestão Documental. Gestão de Documento. Gestão da Informação. Modelo de Gestão. <http://www.uel.br/revistas/informacao/tbci/vocab/index.php?tema=3&/3-gestao-da-informacao>

Fonte: Dados da pesquisa.

O artigo representado no Quadro 9 foi classificado com o termo de assunto “Gestão da Informação” contido no vocabulário controlado elaborado, além da URI direcionando para o termo na TBCI online.

3.4.7.3. Treinar o modelo com os *backends* (algoritmos léxicos e associativos) e *ensemble*

Finalizada a elaboração do vocabulário controlado do arquivo com o corpus e carregado na ferramenta, a atividade posterior é realizar o treinamento do modelo com os *backends*/algoritmos suportados.

O ANNIF suporta diversos algoritmos léxicos e associativos, conforme descrito na seção 2.4 desta tese. Selecionou-se os algoritmos MLLM e STWFSA (léxicos); TFIDF (associativo), além de configurar um projeto *ensemble* (pacote com os três algoritmos trabalhando em conjunto) e extraindo o melhor de cada um deles. No Quadro 10, demonstra a configuração do arquivo “*projects.cfg*” com os projetos:

Quadro 10 – Arquivo “*projects.cfg*” do ANNIF

```
[projeto-mllm]
name=Modelo MLLM
language=pt
backend=mllm
vocab=tesauro-tbci
analyzer=snowball(portuguese)
[projeto-stwfsa]
name=Modelo STWFSA
language=pt
backend=stwfsa
vocab=tesauro-tbci
analyzer=snowball(portuguese)
[projeto-tfidf]
name=Modelo TFIDF
language=pt
backend=tfidf
```

```
vocab=tesauro-tbci
analyzer=snowball(portuguese)
[projeto-ensemble]
name=Modelo Ensemble
language=pt
backend=ensemble
vocab=tesauro-tbci
sources=modelo-mllm, modelo-stwfsa:2, modelo-tfidf
```

Fonte: Dados da pesquisa.

Em seguida, carregou-se o vocabulário controlado para o projeto e executou-se o treinamento dos modelos configurados, exceto o modelo Ensemble que não possui analisador configurado. Esse modelo encapsula os modelos dos outros *backends* treinados, realizando a junção do processamento de aprendizagem de máquina. Quando for realizar a sugestão de assuntos para uma tese ou dissertação, basta utilizar o projeto ensemble que ele vai realizar a ação baseado nos treinamentos processados pelos *backends* TFIDF, STWFSA e MLLM.

No Quadro 11 é apresentado os comandos no ANNIF para fazer a carga do vocabulário controlado ao projeto e o treinamento de cada *backend*/algoritmo:

Quadro 11 – Comandos para carga do VC e treinamento do *backend*

```
#Realizando a carga do vocabulário controlado "tesauro-tbci.txt" aos projetos
$annif loadvoc projeto-tfidf data-sets/projeto-tfidf/tesauro-tbci.txt
$annif loadvoc projeto-stwfsa data-sets/projeto-stwfsa/tesauro-tbci.txt
$annif loadvoc projeto-mllm data-sets/projeto-mllm/tesauro-tbci.txt
$annif loadvoc projeto-ensemble data-sets/projeto-ensemble/tesauro-tbci.txt

#Executando o treinamento dos backends TFIDF, STWFSA e MLLM utilizando o
#arquivo do corpus-brapci.tsv
$annif train projeto-tfidf data-sets/projeto-tfidf/corpus-brapci.tsv
$annif train projeto-stwfsa data-sets/projeto-stwfsa/corpus-brapci.tsv
$annif train projeto-mllm data-sets/projeto-mllm/corpus-brapci.tsv
```

Fonte: Dados da pesquisa.

3.4.7.4. Gerar assuntos automaticamente

Depois que todos os modelos foram treinados, passa-se a atividade de realizar a geração de assuntos. Optou-se por gerar 10 assuntos para cada *backend* isolado (TFIDF, MLLM e STWFSA), assim como gerar 10 assuntos para o modelo ensemble. Essa escolha se deu com o objetivo de se ampliar as possibilidades de busca facetada no repositório digital. O usuário terá a opção de 40 termos de busca para recuperar um documento.

Nesse contexto, como as teses e dissertações coletadas na seção 3.4.4 desta pesquisa estão em formato "pdf", deve-se realizar a transformação deste formato para texto com a

extensão “txt”. Em seguida realizar a sugestão de assuntos para tese ou dissertação, conforme Quadro 12:

Quadro 12 – Conversão de arquivos e sugestão de assuntos

#Conversão de pdf para txt
$pdftotext 1997_LillianAlvares.pdf 1997_LillianAlvares.txt

#Sugestão de assuntos de uma Dissertação de Mestrado nos três modelos de treinamento e no modelo ensemble.
$annif suggest projeto-tfidf <datasets/RiUNB/1997_LillianAlvares.txt
$annif suggest projeto-stwfsa <datasets/RiUNB/1997_LillianAlvares.txt
$annif suggest projeto-mllm <datasets/RiUNB/1997_LillianAlvares.txt
$annif suggest projeto-ensemble <datasets/RiUNB/1997_LillianAlvares.txt

Fonte: Dados da pesquisa.

3.4.7.5. Exportar os dados em formato aberto CSV

Finalizada a etapa anterior de geração automática de assuntos, realizou-se a exportação dos dados para o formato aberto “csv”. Esse arquivo contém todas as informações necessárias para ser importado no repositório digital a ser utilizado e foi nomeado como “*RiUNB_sugestao_assuntos_ANNIF.csv*”.

Esse arquivo contém os seguintes metadados: Título, Autor, Orientador, Descritores, Resumo, Abstract, Data de Publicação, Data de Defesa, Citação, Tipo de Pesquisa, URI, URL, Sugestão de Assunto – Annif MLLM, Sugestão de Assunto – Annif STWFSA, Sugestão de Assunto – Annif TFIDF e Sugestão de Assunto – Annif Ensemble.

O campo “descritores” e todos aqueles relativos à sugestão de assuntos do ANNIF são multivalorados, sendo nesse contexto inserido o delimitador “||”, conforme Figura 18:

Figura 18 – Recorte de colunas do arquivo “*RiUNB_sugestao_assuntos_ANNIF.csv*”

	A	B	D	Q	R	S	T
1	Título	Autor	Descritores	Sugestão de Assunto - Annif Ensemble	Sugestão de Assunto - Annif MLLM	Sugestão de Assunto - Annif STWFSA	Sugestão de Assunto - Annif TFIDF
2	Educação continuada e a distância de profissionais da Ciência da informação no Brasil via internet	Naves, Carlos Henrique Tomé	Ensino a distância||Ciência da informação||estudo e ensino||Ciência da informação||Brasil	Correio Eletrônico||Educação Continuada||Aldeia Global||Superposição||Geografia||Listas de Discussão||Engenharia Química||Tutoriais||Estágios||Correspondência	Listas de Discussão||Correio Eletrônico||Educação Continuada||Aldeia Global||Instituições de Ensino e Pesquisa||Incerteza||Salas Virtuais||Conferências na Web||Transferência de Arquivos||Superposição	Educação Continuada||Superposição||Estágios||Correspondência||Geografia||Correio Eletrônico||Serviços de Biblioteca||Ensino Técnico||Aldeia Global||Listas de Discussão	Institutos de Pesquisa||Ambiguidade||Sociologia do Conhecimento||Sistemas de Informação Geográficas||Ensino à Distância||Conteúdo da Informação||Análise Qualitativa||Confiabilidade||Ciência Política||Redes de Comunicação e Informação, Internet, Web
3	O contexto do desenvolvimento de coleções em coleções digitais jurídicas	Oliveira, Anastácia Freitas de	Desenvolvimento de coleções||Informação jurídica||Biblioteca digital||Livro eletrônico||Coleções digitais||Bases de dados	E-Book||Download||OCR||Sinopse||Museologia||Audiolivro||Objetos Digitais||Ergonomia||Correio Eletrônico||Tutoriais	E-Book||Download||Buscas em Texto Completo||Domínio Público||Transferência de Arquivos||Correio Eletrônico||OCR||Encadernação||Sinopse||Credibilidade	E-Book||Objetos Digitais||Museologia||OCR||Institutos de Pesquisa||Arranjo||Estágios||Suportes de Informação||Obras de Referência||Sinopse	Institutos de Pesquisa||Redes de Comunicação e Informação, Internet, Web||Bibliotecas de Pesquisa||Ambiguidade||Superposição||E-Book||Documentos e Informação como Componente||MARC||Usuários e Usos da Informação||Confiabilidade
	Fotografias periciais : definição diplomática de documentos imagéticos forenses	Freitas Junior, Edson Ferreira de	Documentos fotográficos||Diplomática||Prova documental||Arquivos||Proteção	Credibilidade||Organograma||Câmeras Digitais||Arquivamento||Correspondência||Fotografia||Química||Obras de Referência||Dicionário||Indexação por Assuntos	Documento Iconográfico||Conservação de Documentos||Credibilidade||Câmeras Digitais||Transferência de Arquivos||Publicação Oficial||Empirismo||Arquivos Públicos||Artes	Credibilidade||Arquivamento||Estágios||Correspondência||Câmeras Digitais||Química||Dicionário||Organograma||Obras de Referência||Indexação por Assuntos	Documentos e Informação como Componente||AACR2||Institutos de Pesquisa||Documento Audiovisual||Categorização Automatizada de Textos||Engenharia de Transportes||Ambiguidade||Arquivamento||Documento Iconográfico||Confiabilidade

Fonte: Dados da pesquisa

3.4.8. Instalar o repositório digital (Tainacan)

O Tainacan é um plugin do *Wordpress*[17], um Sistema de Gestão de Conteúdo (CMS) para internet, baseado em PHP e banco de dados MySQL, que utiliza o servidor web apache.

Nesse contexto, se faz necessário instalar inicialmente o *Wordpress* para depois ativar o plugin do Tainacan nesse CMS.

Na internet, há sites específicos das ferramentas com os seus manuais de instalação:

- *Wordpress* – https://br.wordpress.org/support/forum/instalacao-e-migracao/;
- Tainacan - https://tainacan.github.io/tainacan-wiki/#/pt-br/instalacao.

Os manuais contemplam todas as atividades necessárias para: instalar o servidor web apache, instalar os pacotes do PHP, instalar o MySQL e configurar o banco de dados do Wordpress, iniciar os serviços, configurar o Wordpress e acessá-lo via web, ativar o plugin do Tainacan, além de como parametrizar esse repositório digital.

3.4.8.1. Configurar coleção do repositório digital e importar os dados

Ao iniciar o Tainacan, a primeira tarefa é configurar um conjunto de itens agrupados com metadados das Teses e Dissertações, além de seu arquivo digital em "pdf". A esses itens agrupados e organizados dá-se o nome de coleção.

Para realizar a criação da coleção, o Tainacan fornece a funcionalidade de realizar a carga de dados, a partir da importação de arquivo "csv". Na seção 3.4.7.5, foi realizada a exportação do arquivo final "*RiUNB_sugestao_assuntos_ANNIF.csv"* com todos os metadados de Teses e Dissertações da Ciência da Informação. Nesse sentido, esse mesmo arquivo será importado nessa etapa de configuração da coleção no Tainacan.

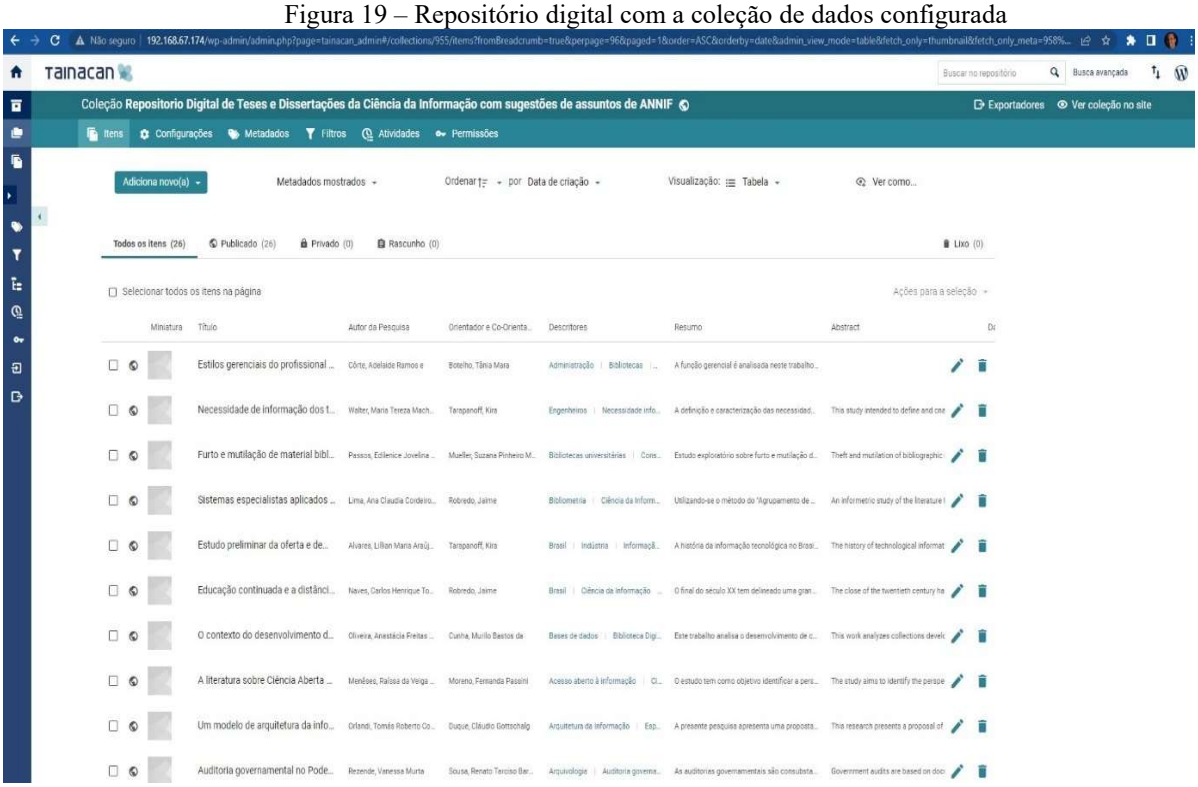
Importante salientar que ao criar uma coleção a partir da importação de arquivo, é necessário observar os seguintes parâmetros: delimitador do CSV (caractere usado para separar a coluna do arquivo), delimitador de metadados multivalorados (caractere utilizado "||" para separar cada valor dentro de uma célula com múltiplos valores, sendo necessário atentar para que o metadado de destino aceite múltiplos valores); delimitador de textos

[17] Disponível em <https://br.wordpress.org/>, acesso em 23/02/2023

(caractere que delimita o campo de cada célula) e a codificação do arquivo (utilizamos desde a geração dos arquivos o formato UTF-8).

A Coleção para essa pesquisa foi criada com o nome de "**Repositório Digital de Teses e Dissertações da Ciência da Informação com sugestão de assuntos de ANNIF**", conforme pode ser observado através da Figura 19:

Figura 19 – Repositório digital com a coleção de dados configurada



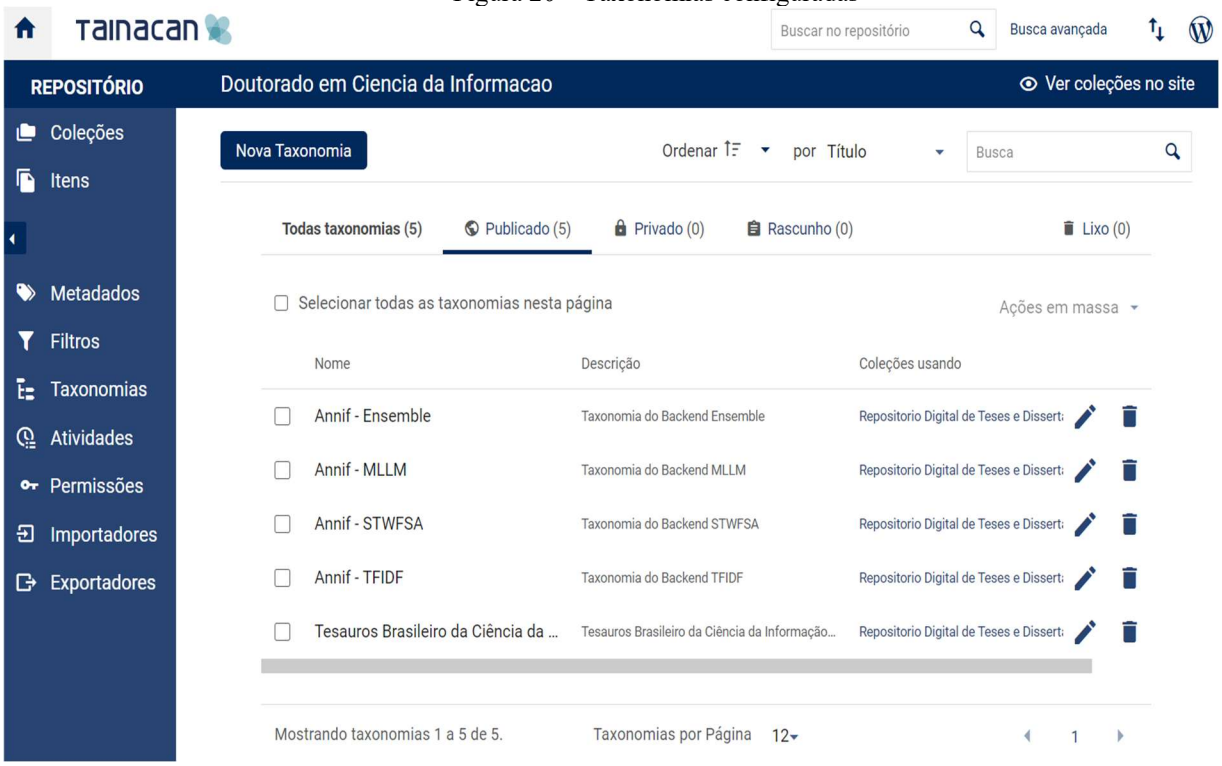
Fonte: Dados da pesquisa

3.4.8.2.Configurar taxonomias e filtros de busca

Uma taxonomia pode ser compreendida como um termo de um vocabulário controlado com intuito de descrever um item. Esses termos podem ser uma lista em ordem alfabética ou estruturada de forma hierárquica.

Para a taxonomia "Tesauros Brasileiro da Ciência da Informação", importou-se os dados multivalorados do campo "Descritor" (Palavras-chave das Teses e dissertações), conforme Figura 20:

Figura 20 – Taxonomias configuradas



Fonte: Dados da pesquisa

As demais taxonomias referentes aos *backends* do ANNIF foram extraídas dos registros multivalorados das colunas correspondentes as sugestões de assuntos do ANNIF dento do arquivo "*RiUNB_sugestao_assuntos_ANNIF.csv*", exportado na seção 3.4.7.5.

Em seguida, configurou-se os filtros de busca com os nomes: *Descritor, Annif-Ensemble, Annif-TFIDF, Annif-MLLM e Annif-STWFSA*.

3.4.8.3.Realizar buscas facetadas

Essa é a principal seção de todo esse estudo de caso, pois nota-se que o esforço executado propiciou a organização do acervo de teses e dissertações no repositório digital, além de fornecer mais opções de metadados para auxiliar na recuperação da informação.

O Tainacan executa formas diferentes de classificação dentro da coleção, funcionando como facetas de busca, que melhoram sua utilização e facilita a construção do repositório pelo gestor do acervo digital. As categorias criadas permitem a distribuição do conteúdo, conforme as necessidades da coleção.

No presente estudo, há possibilidade de realizar a busca facetada através dos filtros de busca, selecionando-se o termo em cada filtro. Criou-se um filtro denominado "Descritor",

que busca as palavras-chaves para cada registro de Tese e Dissertação armazenada, com a facilidade de o usuário criar termos de busca ao filtro aplicado, denominado de etiquetas, conforme Figura 21. Nesse sentido, a ferramenta possibilita interação com o usuário, propiciando a personalização dos termos de busca.

Figura 21 – Filtros de busca aplicados

Fonte: Dados da pesquisa

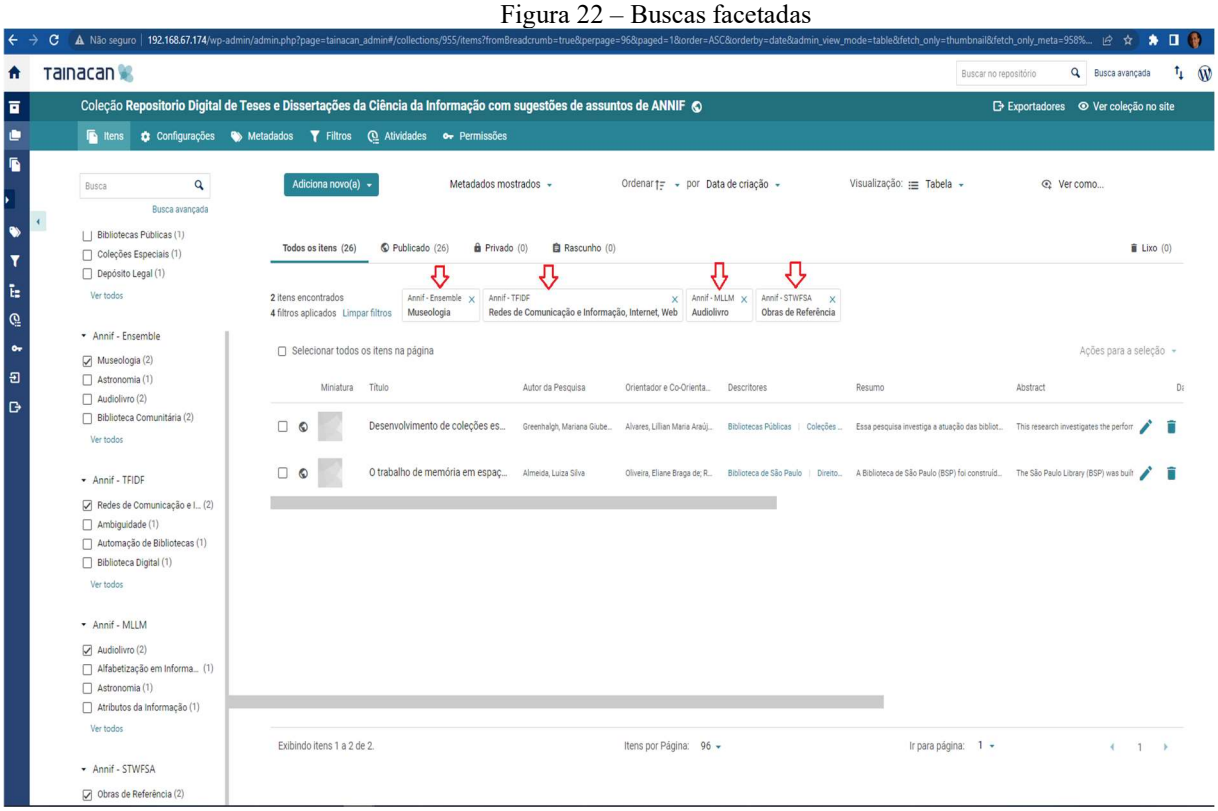
No exemplo demonstrado na Figura 21, realizou-se a busca facetada aplicando-se quatro filtros, tendo como resultado 2 itens no repositório digital. Os filtros aplicados foram:

- *Annif-Ensemble*: Museologia;
- *Annif-TFIDF*: Redes de Comunicação e Informação, Internet, Web;
- *Annif-MLLM*: Áudiolivro;
- *Annif-STWFSA*: Obras de referência.

É importante observar que para cada Tese e Dissertação armazenada há um conjunto de 10 sugestões de assuntos para cada *backend* do Annif, ou seja, 40 termos que são adicionados ao número de 5 descritores inicialmente coletados e relacionados às palavras-chave dos trabalhos de pesquisa.

Para cada documento registrado na base de dados do Tainacan, apresenta 45 termos relacionados que auxiliam na aplicação dos filtros, melhorando sua indexação e recuperação do recurso de informação, conforme Figura 22:

Figura 22 – Buscas facetadas



Fonte: Dados da pesquisa

O menu à esquerda, contém os filtros de busca com todas as sugestões de assuntos do ANNIF, além das palavras-chaves de cada documento original. A ferramenta Tainacan propiciou uma melhor organização, além de fornecer mecanismos para recuperar de forma eficaz os recursos de informação.

3.4.9. Análise

A atividade de análise e resultados da pesquisa será realizada no Capítulo 4.

3.4.10. Compartilhamento

Com o objetivo de dar a visibilidade aos dados deste estudo de caso, todos os arquivos gerados estão listados no Quadro 13 associado ao seu *link* para consulta:

Quadro 13 – Arquivos gerados na pesquisa

Rótulo	Descrição	URL
Revisão Sistemática da Literatura – Parte 1	Arquivo contendo as tabulações das extrações de dados nas revistas científicas.	https://cutt.ly/pWiqjrN
Revisão Sistemática da Literatura – Parte 2	Critérios de qualidade e categorias de análise aplicadas para seleção dos artigos para a síntese qualitativa.	https://cutt.ly/tWiDg3k
arquivo_coleta_unb.csv	Arquivo contendo as coleções e *datasets* da CI na RiUNB	https://cutt.ly/p8UO0B9
tesauro-tbci.txt	Tesauro adequado a partir da TBCI para carga no ANNIF.	https://cutt.ly/L8UPZ0Q
metadadosBRAPCI_to_corpus.xls	Arquivo contendo os metadados de artigos da BRAPCI para composição do corpus.	https://cutt.ly/r8UA27W
BRITO, J. C. B; MARTINS, D. L. **Geração automática e semiautomática de metadados: revisão bibliográfica**. XXI Enancib, Rio de Janeiro, 25-29 outubro, 2021.	Artigo publicado no XXI Enancib, relacionado à Revisão Sistemática da Literatura – RSL para a compreensão das ferramentas de geração de metadados.	https://brapci.inf.br/index.php/res/download/216427
BRITO, J. C. B; MARTINS, D. L. **Geração automática de metadados: estudo de caso utilizando a técnica de Indexação automática estatística com a ferramenta ANNIF**. XXII Enancib, Porto Alegre, 07-11 novembro, 2022	Artigo Publicado no XXII Enancib, relacionado com a seleção da ferramenta ANNIF para geração automática de assuntos de teses e dissertações da RiUNB.	https://enancib.ancib.org/index.php/enancib/xxiienancib/paper/viewFile/777/719
BRITO, J. C. B; MARTINS, D. L. **Revisão sistemática da literatura na Ciência da Informação: uma descrição detalhada dos passos metodológicos.** InCID: R. Ci. Inf. e Doc., Ribeirão Preto, v. 14, n. 2, p. 24 – 47, set. 2023/fev. 2024.	Artigo Publicado na Revista Incid (USP). Qualis A3	https://www.revistas.usp.br/incid/article/view/209021/200645
BRITO, J. C. B; MARTINS, D. L. **Geração automática de metadados: estudo de caso utilizando a técnica de indexação automática estatística com a ferramenta ANNIF**. Revista Caribeña de Las Ciências Sociales, Miami, v.12, n.4, p. 1980-1996. 2023. ISSN 2254-7630.	Artigo Publicado na Revista Caribenã de Las Ciências Sociales. Qualis B1.	https://ojs.southfloridapublishing.com/ojs/index.php/rccs/article/view/2995
BRITO, J. C. B; MARTINS, D. **Framework Genérico para Geração Automática de Assuntos e Indexação em Repositório Digital.** Perspectivas em Ciência da Informação, Belo Horizonte, v. 28, Fluxo Contínuo, 2023: e-46629.	Artigo Publicado na Revista Perspectivas em Ciência da Informações (UFMG). Qualis A2.	https://periodicos.ufmg.br/index.php/pci/article/view/46629/39251

Fonte: Dados da Pesquisa

4. ANÁLISE E DISCUSSÃO DOS RESULTADOS

Este capítulo tem por objetivo apresentar a análise e discussão dos resultados relativos aos estudos desenvolvidos nessa pesquisa, avaliando:

- A revisão sistemática realizada, a compreensão de estudos correlatos e suas contribuições;
- O modelo de pesquisa proposto;
- O desenvolvimento de procedimentos e atividades com o auxílio de soluções tecnológicas para implementar a coleta de metadados e arquivos de um repositório de dados institucional para geração automática de assuntos, indexação em repositório digital e implementação de busca facetada.

4.1. DA REVISÃO SISTEMÁTICA

A revisão sistemática corroborou para a compreensão metodológica dos passos necessários para definição da questão de pesquisa, aplicando questões de *background* e *foreground* para delimitação e escopo da investigação.

Em seguida, passou-se a seleção de base de dados científicas, construção dos termos de busca e as sentenças, utilizando os operadores lógicos para recuperar as informações relacionadas com as ferramentas de geração automática e semiautomáticas de metadados, além da definição dos critérios de seleção/exclusão e atividades relacionadas com a qualidade.

Esse processo, definido através de um protocolo rígido para revisão, forneceu a fundamentação necessária e a compreensão do contexto do uso, aplicação e as limitações das ferramentas analisadas.

Para a síntese qualitativa, selecionou-se 12 artigos que apresentaram estudos correlatos, descrevendo as técnicas de extração de metadados e aplicações de diversas ferramentas automática e semiautomáticas de metadados.

Em relação as ferramentas automáticas identificadas a partir dos artigos da RSL, depreendeu-se que: quatro utilizavam a técnica de extração de conteúdo; duas, a mineração de textos e dados; uma, relacionada com *folksonomia* ou marcação social; uma, através de colheita de *metatags*; duas, com a indexação ou classificação automática; e apenas uma, para geração automática de dados extrínsecos. Nesse contexto, as ferramentas analisadas abordaram todas as técnicas elencadas por Polfreman *et al.* (2008).

As ferramentas semiautomáticas identificadas foram em número de 39, mas 6 foram descontinuadas, não disponibilizando serviços de atualizações ou suporte. Diversas ferramentas implementavam mais de uma técnica para extração de metadados. Outra característica foi que as ferramentas GAM e GSAM foram utilizadas para resolver problemas específicos e não foi identificado um modelo geral para aplicação, o que motivou a proposição do *framework* genérico e sua validação através do modelo de pesquisa aplicadas através do estudo de caso.

Após a revisão sistemática, percebeu-se o quantitativo de estudos correlatos em diversos países do mundo, exceto a ausência de pesquisas sobre a temática pesquisada no Brasil e na América Latina.

Realizou-se a identificação na seção 2.1.4.3 desta tese, as possibilidades de uso das ferramentas GAM e GSAM, além da análise de suas limitações apresentadas na seção 2.1.4.4.

Identificou-se que as ferramentas automáticas e semiautomáticas analisadas, possuíam várias limitações que inviabilizaram sua seleção para uma das atividades desta investigação que era a geração automática de assuntos. Nesse contexto, continuou-se a executar pesquisas para encontrar soluções que mitigassem as limitações levantadas das ferramentas identificadas na RSL.

Após a submissão de artigo da revisão sistemática para o XXI Enancib em 2021, um dos revisores fez uma crítica, afirmando que os termos utilizados para a RSL não contemplavam estudos no Brasil, por não englobar o termo “indexação automática” que é o mais comum utilizado em pesquisas brasileiras. Foi realizada nova busca com o termo proposto, identificando 29 publicações no Brasil. Além disso, realizou-se busca exploratória em sites de bibliotecas que contemplassem pesquisas sobre geração automática de metadados, identificando mais uma ferramenta GAM.

Nesse contexto, foram identificadas algumas soluções à posteriori da RSL, sendo 01 ferramenta de GSAM e 11 ferramentas de GAM, conforme Quadro 14:

Quadro 14 – As ferramentas de GAM e GSAM identificadas à posteriori

Ferramenta	**GSAM**	**GAM**
Coletador	X	
Bib/diálogo		X
Sisa		X
Precis		X
Ogma		X
Sintagmed		X

Zstation		X
Sriac		X
Spirit		X
Mistral		X
Sirilico		X
Annif		X

Fonte: Dados da pesquisa

A pesquisa analisou o total de

- 23 (vinte e três) estudos correlatos no Brasil e exterior;
- 5 (cinco) técnicas de extração de metadados implementadas, tanto em ferramentas de GSAM, quanto de GAM;
- 40 (quarenta) ferramentas GSAM, sendo 6 (seis) descontinuadas; e 22 (vinte e duas) ferramentas que implementam a GAM.

Após a verificação das ferramentas, optou-se por selecionar as soluções Coletador (GSAM) e ANNIF (GAM) para serem as soluções para implementar, respectivamente, a coleta de metadados de repositório institucional e a geração automática de assuntos.

Foram realizados testes e verificou-se o funcionamento à contento das duas soluções e a mitigação das limitações das ferramentas apresentadas na RSL. Ambas as ferramentas são desenvolvidas em *python* e de código aberto, o que facilita a compreensão e sua adequação para a pesquisa.

Nesse contexto a RSL contribuiu para identificar e compreender o uso das ferramentas de geração automática e semiautomática de metadados, suas técnicas, características, funções e limitações. A partir das fundamentações observadas, propiciou selecionar à posteriori as ferramentas que melhor se adequassem às necessidades desta investigação.

4.2. DA VALIDAÇÃO DO *FRAMEWORK* GENÉRICO

Foram executados todos os passos do *framework* genérico conceitual proposto na seção 2.5 dessa tese, o que comprova que a seleção de cada elemento (ferramenta, fonte de dados) desdobrou-se em ações que foram executadas à contendo e contribuiram para o alcance dos resultados esperados.

Passo 1 – Coletar Teses, dissertações e metadados: utilizou-se a ferramenta de GSAM, denominada coletador, que através do protocolo OAI-PMH conseguiu efetuar a conexão com o repositório institucional da UnB para coleta de metadados. A limitação ocorreu apenas com

o *download* do arquivo completo em "pdf" da tese e dissertação. A atividade de baixar esses documentos foi realizada via *python* e linha de comando do Linux. Entretanto, a limitação ficou nas restrições da RiUNB, que exigia autenticação para obter vários documentos, diminuindo a quantidade de publicações baixadas. Esse processo não limitou os passos seguintes da investigação realizada, apenas reduziu a quantidade de registros que havia sido identificado no repositório.

Passo 1.2 – Tratar os dados: foi realizado à contento, identificando os dados faltantes ("*missing data*") que não ocasionaram impacto na análise; e os dados multivalorados, sendo observado que foram coletados com delimitadores, organizando melhor o registro com os termos ou sentenças.

Passo 2 – Configurar o vocabulário controlado: utilizou-se o Tesauro Brasileiro da Ciência da Informação como fonte original para adequação de um VC. Executou-se a elaboração do artefato, contendo 438 termos da TBCI, visto que essa fonte de dados está em formato "pdf" ou no formato online mantida pela UEL, mas que não possui os dados em formato aberto, seja SKOS, Turtle ou XML.

Passo 3 – Instalar o ANNIF: utilizou-se máquina pré-instalada através da plataforma virtualBox. Não foi identificado dificuldades nesse passo, visto que a ferramenta possui site explicativo com manuais de utilização detalhados, assim como comunidade atuante com diversos utilizadores que compartilham experiências e informações na internet (https://annif.org/ e annif-users@googlegroups.com).

Passo 3.1– Configurar o projeto: procedimento simples, desde que planejado, pois é necessário definir quais os analisadores, o vocabulário controlado, os *backends*/algoritmos a serem utilizados.

Passo 3.2 – Definir o corpus de conhecimento: selecionou-se a BRAPCI pelo fato de ser um dos repositórios com um grande acervo de metadados de artigos científicos na área da Ciência da Informação. O corpus é um dos elementos essenciais do processo de aprendizagem de máquina, pois a partir dele é que são analisados os termos utilizados, sua frequência, cotejamento com o vocabulário controlado, fornecendo o aprendizado necessário para que os diversos algoritmos consigam sugerir assuntos posteriormente sobre as teses e dissertações. Para elaboração do corpus não foi definido a necessidade de textos completos, apenas o uso de metadados associados ao título, resumo e palavras-chaves dos artigos da BRAPCI. O impacto nessa etapa se deu na análise da base do repositório que continha muitos campos multivalorados sem delimitadores, ausência de dados (*missing data*) ou discrepantes (*outliers*). Verificou-se que a entrada para alguns artigos deve ser manual, pois identificou

que alguns metadados cadastrados não correspondiam aos descritores no documento original. O trabalho para sanitizar uma base com mais de 46 mil registros iria ser dispendiosa e dessa forma optou-se por realizar a adequação do corpus de forma manual. Ao final da atividade, construiu-se um corpus com 438 conjuntos de registros de metadados de artigos da BRAPCI.

Passo 3.3 – Treinar o modelo com os *backends* (algoritmos léxicos e associativos) e *ensemble*: Uma vez definido o corpus e o vocabulário controlado, a atividade de treino do modelo é simples, pois o próprio ANNIF realiza a tarefa automaticamente a partir de comandos. Selecionou-se os algoritmos léxicos MLLM e STWFSA; e do associativo TFIDF; além do modelo *ensemble*, que encapsula os três algoritmos anteriores e extrai deles o melhor resultado, trabalhando em conjunto. Percebeu-se que os assuntos sugeridos apresentavam similaridade satisfatória com os termos e assuntos dispostos nos documentos analisados. Entretanto, para que a sugestão de assuntos possa melhorar o seu nível de precisão, é necessário ampliar os termos do vocabulário controlado e principalmente o número de registros do corpus de conhecimento. Para atividades de aprendizagem de máquina é necessário um volume com milhares de registros para conduzir de forma eficaz o treinamento do modelo.

Passo 3.4 – Após o modelo treinado, basta utilizar os *backends*/algoritmos para sugerir assuntos com a temática pela qual eles foram treinados.

Passo 3.5 – Finalizada a sugestão de assuntos e gerado o arquivo final para ser exportado em formato aberto e servir como insumo/entrada para a próxima atividade do processo.

Passo 4 – Instalar o Tainacan: atividade bem fácil, pois existem manuais e documentações, além de vídeos explicativos. Esse repositório digital possui dependências tais como a instalação do WordPress, PHP, banco de dados MySQL, servidor web apache e algumas configurações de rede (*firewall*). Esse repositório é intuitivo e fornece acessibilidade e manuseio facilitados. Muitos repositórios na área de artes e museus utilizam essa solução para catalogação e indexação de seus acervos.

Passo 4.1 – Configurar coleção do repositório digital e importar os dados: atividade trivial e necessária, pois aqui foi testado a integração entre o que foi gerado pelo ANNIF e sua indexação no Tainacan. A área de administração do repositório é intuitiva para execução das tarefas de importação/exportação.

Passo 4.2 – Configurar taxonomia e filtros de busca: propiciou utilizar o vocabulário controlado utilizado para o ANNIF para a criação de taxonomias no Tainacan. Essa parametrização propiciou a configuração dos filtros de busca baseado nos metadados

sugeridos por ANNIF, o que aumenta a capacidade dos usuários do repositório digital em recuperar de forma facilitada o recurso de informação.

Passo 4.3 – Realizar buscas facetadas: foi utilizado os descritores da própria tese e dissertação, além de 10 (dez) termos de assuntos sugeridos pelo ANNIF para cada um dos quatro *backends*/algoritmos utilizados. Ampliou-se a assim, a indexação que antes era executada apenas por 5 palavras-chaves, para 45 descritores, sendo 40 sugeridos de forma automática pelo ANNIF, através da técnica de aprendizagem de máquina.

Dessa forma, o modelo de pesquisa proposto foi executado por completo, corroborando na afirmação de que: o processo; as atividades; a seleção da RiUNB, TBCI e BRAPCI; seleção das soluções tecnológicas Coletador, ANNIF e Tainacan; atingiram os resultados esperados.

4.3. DO ESTUDO DE CASO

O estudo de caso seguiu o protocolo adequado de YIN (2015) e aplicou as soluções e elementos necessários para realização inicial do caso-piloto, que serviu de referência para tomada de decisão da continuidade da pesquisa com as definições iniciais. Após a verificação da consistência e resultados obtidos, passou-se ao desenvolvimento de atividades com o escopo maior de análise na elaboração do vocabulário controlado e das taxonomias, o corpus de conhecimento, o treinamento do modelo com os algoritmos léxicos e associativos, além da sugestão de assuntos e indexação e recuperação da informação facetada no repositório digital.

O *software* coletador demostrou-se uma ferramenta importante para colher metadados e arquivos digitais que utilizam interoperabilidade baseada no protocolo OAI-PMH, sendo recomendado seu uso. Entretanto, é necessário a implementação via script ou uma linguagem de programação, como o *phyton*, para realizar os *downloads* dos documentos.

O Annif comprovou ser uma ferramenta robusta e que pode auxiliar de forma eficiente e eficaz no processo de geração automática de assuntos. A solução utilizou algoritmos distintos, trabalhando em conjunto, auxiliando no processo de indexação/classificação. Nessa pesquisa, utilizou-se algoritmos léxicos, associativo e junção de *backends*, com intuito de extrair o melhor que cada um deles poderia oferecer. Em situações em que não há um vocabulário controlado disponível, recomenda-se utilizar os algoritmos associativos para realizar a indexação de documentos com textos completos. Muitas ferramentas utilizam apenas um algoritmo de classificação e o Annif possui o diferencial para trabalhar com “n” *backends* possíveis, basta configurar um projeto para isso.

O Tainacan também demonstrou ser uma ferramenta versátil e fácil de operar e administrar. Sua interface é bem intuitiva e a capacidade de realizar importações em massa, facilita o trabalho de indexação automatizada dos dados. Seu diferencial consiste na capacidade de criação das coleções, filtros de busca, taxonomias e implementação de buscas facetadas.

A metodologia aplicada e os resultados obtidos no estudo de caso forneceram informações suficientes para ratificar que o *framework* genérico proposto é válido e passível de ser aplicado em outras temáticas e abordagens. Nesse sentido, colaborou-se nesse estudo para a mitigação de limitação encontrada, ao se investigar sobre as ferramentas automáticas e semiautomáticas de metadados, quanto ao uso dessas tecnologias para a solução de um problema específico. A proposta do *framework* genérico é para que ele possa ser adequado para diversas áreas, com a seleção de ferramentas e atividades diversas, sugeridas por esse trabalho.

5. CONSIDERAÇÕES FINAIS

Retomamos a questão de pesquisa inicial: "**Como a técnica de geração automática de metadados pode apoiar usuários e gestores de repositórios digitais na melhoria dos recursos de organização de informação visando facilitar a busca e recuperação da informação dos seus acervos?**"

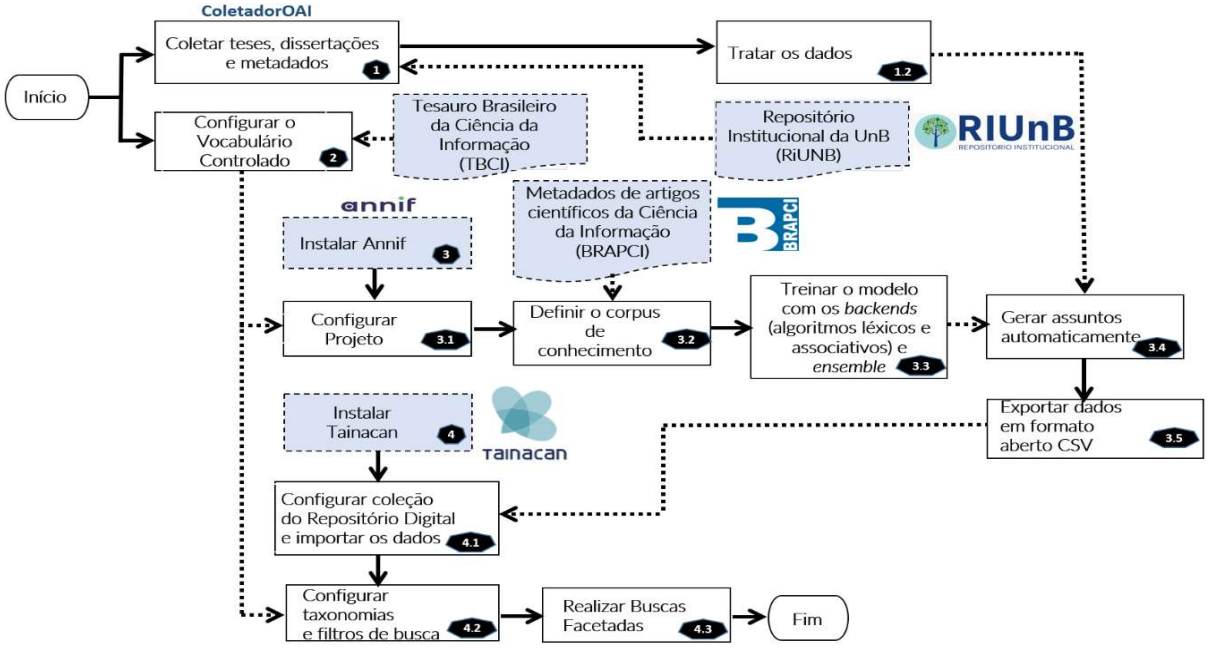
Conclui-se que as técnicas de geração automática de metadados auxiliam na sugestão de assuntos para documentos robustos como uma Tese e Dissertação, ampliando o quantitativo de descritores, de modo a facilitar a configuração de taxonomias, filtros e facetas. Esse trabalho propôs o *Framework* Genérico para ser aplicado em qualquer área do conhecimento, com intuito de melhorar e facilitar a busca e a recuperação da informação nos acervos digitais pelos usuários e a organização da informação pelos gestores responsáveis.

5.1. CONCLUSÕES

Inicialmente, essa pesquisa focou em uma investigação geral sobre as ferramentas de geração automática e semiautomática de metadados. Ampliou-se o conhecimento sobre as diferentes técnicas para extração de metadados, assim como as diversas soluções tecnológicas que fornecem suporte à organização e recuperação da informação.

Para atendimento ao objetivo geral, apresentou-se um *framework* genérico no Capítulo 2.5 desta tese, validado pelo estudo de caso executado e demonstrado na Figura 23:

Figura 23 – *Framework* Genérico Conceitual validado através do estudo de caso



Fonte: o Autor.

Esse modelo de pesquisa utilizou diversos elementos caracterizados na cor azul na Figura 23 e que podem ser substituídos por outras abordagens, seja a seleção de um tesauro, base de dados ou documentos para compor um corpus, ferramentas de apoio para aplicar técnicas de extração de metadados e sugestão de assuntos, além de outros repositórios digitais.

Para atender os objetivos específicos delimitados para essa investigação, realizou-se:

- A identificação de técnicas, tecnologias e metodologias de geração automática/semiautomática de metadados, apresentadas através do Capítulos 2;

- A identificação de um repositório digital para o estudo de caso, e caracterizar seus recursos atuais de organização da informação que possam ser melhorados para busca e recuperação pela aplicação de técnicas de geração automática/semiautomática de metadados. Na seção 1.4 sobre a "justificativa da pesquisa", evidenciou-se falhas na recuperação da informação no RiUNB, selecionando-se esse repositório para desenvolver a coleta de metadados, dados e arquivos para executar a aprendizagem de máquina e posterior carga e indexação. Na seção 2.2.4, abordou-se a temática de repositórios digitais, além de realização de estudos comparativas entre Omeka, Dspace e Tainacan, sendo selecionado este último para essa investigação, apresentando resultados satisfatórios para indexação, criação de taxonomias, organização da informação e buscas facetadas;

- A identificação dos instrumentos de tratamento descritivo e/ou tratamento temático da informação que possam ser utilizados para melhorar a busca e recuperação da informação do repositório digital escolhido como estudo de caso. Na seção 2.2.5, explanou-se sobre os tipos de vocabulário controlado, sendo selecionado na fase de design do estudo de caso (seção 3.4.2) o Tesauro Brasileiro da Ciência da Informação e a base de dados da BRAPCI para fornecer os metadados para compor o corpus de conhecimento;

- A análise e identificação de um conjunto de técnicas específicas de geração automática/semiautomática de metadados que fossem aplicadas a dados reais de um repositório digital. Foram realizadas técnicas de extração de metadados da RiUNB,

através do *software* Coletador. Para a geração automática de assuntos das dissertações e teses da RiUNB, utilizou-se a aplicação ANNIF, compreendendo o uso de *backends*/algoritmos com técnicas de aprendizagem de máquina.

- Aplicar as técnicas escolhidas a dados reais, comparar resultados e analisar criticamente as contribuições e desafios em relação às possibilidades de melhoria dos recursos de busca e recuperação da informação. Na seção 5.4 deste capítulo é sugerido as pesquisas futuras, onde o pesquisador poderá aplicar o *framework* genérico proposto em outras abordagens;

- Identificar recomendações a gestores de repositórios digitais para a implementação de técnicas de geração automática/semiautomática de metadados. Foi possível demonstrar neste estudo diversas técnicas e ferramentas, além de processos, atividades e aplicação prática através de estudo de caso, através do modelo de pesquisa proposto.

5.2. CONTRIBUIÇÕES DA PESQUISA

Essa tese propiciou contribuições teóricas e práticas no âmbito da organização e recuperação da informação. Foi proposto um *framework* genérico, aplicável a qualquer área de conhecimento e que foi validado pelo modelo de pesquisa para a área da Ciência da Informação, através de: processos, atividades e soluções tecnológicas para coleta de dados, aplicação de inteligência artificial, com técnicas de aprendizagem de máquina para sugestão de assuntos e a parametrização de buscas facetadas em um repositório digital.

O estudo demostra a importância do uso de soluções tecnológicas de geração automática e semiautomáticas de metadados, uma complementando a outra, e dos repositórios digitais, demonstrando o impacto significativo na organização e recuperação da informação.

Outro fator importante de contribuição dessa investigação é o compartilhamento dos dados, análises e modelo validado através do estudo de caso, servindo de apoio à comunidade científica e outros pesquisadores que estudam a organização e recuperação da informação, assim como as soluções tecnológicas que fornece o suporte necessário para implementação de funcionalidades que facilitam as atividades dos usuários dos acervos digitais.

5.3. LIMITAÇÕES DA PESQUISA

A primeira limitação desta pesquisa foi não considerar inicialmente o termo "indexação automática" como termo de busca para a RSL. A seleção dos termos de busca por ferramentas de geração automática e semiautomática de metadados tinha por objetivo realizar uma pesquisa mais ampla sob o aspecto das soluções tecnológicas, suas aplicabilidades e limitações. Nesse contexto, não se focou inicialmente na técnica específica de indexação automática para extração ou geração de metadados. Entretanto, ampliou-se a investigação pelo termo à posteriori, que culminou na seleção das soluções utilizadas nessa pesquisa.

A segunda limitação refere-se à ausência de um tesauro *web* de repositórios institucionais brasileiros com interoperabilidade e acesso em formato aberto, tais como: RDF, XML, TURTLE ou JSON. Nesse estudo, utilizou-se o TBCI em *pdf* e da UEL online para adequar um arquivo com 438 termos manualmente. Exemplos a serem seguidos, tais como o tesauro da/de:

- Unesco[18], com diversos termos em inglês, espanhol, russo e árabe; com possibilidade de navegação e *download* em formato aberto;

- Finto.Fi[19], sendo interoperável, assim como as ontologias e esquemas de classificação por diferentes áreas temáticas. Propicia ao usuário procurar vocabulários controlados ou integrar vocabulários de sistemas próprios, utilizando APIs abertas;

- Arte e Arquitetura (TA&A)[20], composto de mais de 100.000 termos relacionados à arte, arquitetura; da antiguidade até o presente.

A terceira limitação ficou relacionada à RiUNB, pois realizou-se a seleção das coleções da área da pós-graduação da Ciência da Informação, que possuía número superior a 600 registros de teses e dissertações. Entretanto, ao realizar o refinamento para a coleta de dados, obteve-se apenas 16 documentos completos, devido ao repositório exigir autenticação. No momento de executar o *download* da tese ou dissertação, apresentava erro informando

[18] Disponível em <https://vocabularies.unesco.org/browser/thesaurus/en/>, acesso em 25/02/2023.
[19] Disponível em <http://finto.fi/en/>, acesso em 25/02/2023.
[20] Disponível em <https://www.aatespanol.cl/>, acesso em 25/02/2023.

“acesso negado”. Nesse sentido, parametrizações de autenticação para acesso aos registros que estão publicizados como “aberto”, devem ser reavaliados.

Uma última limitação, é relacionada à BRAPCI, decorre da dificuldade de sanitização da base de dados, pois havia diversos registros com campos multivalorados e sem delimitadores. Além disso, haviam dados e metadados faltantes (*missing data*) e discrepantes (*outliers*) que impactavam o tratamento dos dados. Optou-se por adequar um corpus de conhecimento manualmente com 438 registros a partir daquela base de dados. Esse fato impactou o uso das técnicas de indexação automática para geração de assuntos. Para execução de atividades que executam aprendizagem de máquina, o ideal é possuir um corpus com dezenas ou centenas de milhares de registros. Quanto maior o corpus sobre uma temática e o uso de um vocabulário com um conjunto de termos abrangente, maior a probabilidade do modelo treinado ser mais acurado e assertivo na sugestão de assuntos. Os mantenedores do ANNIF, por exemplo, disponibilizam base de dados com milhões de registros, seja com textos completos de documentos ou parciais para realização de testes e aprendizado. A Biblioteca Nacional da Finlândia realiza um trabalho que merece destaque pela qualidade dos repositórios digitais, vocabulários controlados e pela disponibilização de *software* com código aberto para indexação automática estatística para geração de assuntos, sendo um exemplo a ser seguido e implementado.

5.4. PESQUISAS FUTURAS

São indicadas quatro oportunidades de pesquisas futuras em relação à abordagem proposta nessa tese. A primeira é aplicar o *framework* genérico proposto, utilizando outras abordagens, tais como:

- Solução tecnológica para coletar dados (extração de metadados) e obtenção de documentos completos. Foram estudadas nessa pesquisa, diversas soluções que podem ser objeto de implementação futura;

- Vocabulários controlados e tesauros de outras áreas de conhecimento e base de dados para o desenvolvimento do corpus de conhecimento; geração automática de assuntos; além de outros repositórios digitais. Essa sugestão de trabalho futuro, forneceria maior robustez ao *framework* genérico proposto nessa tese,

no sentido de ampliar a validação de sua efetividade, além de auxiliar outras áreas de conhecimento na compreensão do processo de melhorar a organização e recuperação da informação dos acervos digitais sob sua tutela.

A segunda, seria ampliar o uso de outros *backends*/algoritmos suportados pelo ANNIF para sugestão de assuntos, analisando sua performance, qualidade e precisão. Esse tipo de pesquisa, amplia a quantidade de metadados dos documentos e a indexação do acervo digital, facilitando sua organização e busca.

A terceira, realizar pesquisa para a esfera pública de governo com a sugestão de algumas possibilidades:

- No Brasil havia uma iniciativa de elaboração de um Vocabulário Controlado de Governo Eletrônico[21] (VCGE), mas desde 2019 não há atualização, sendo poucos termos e áreas de conhecimento abordados. Sugere-se pesquisas futuras para criação de grupos interdisciplinares para elaboração de um tesauro de governo mais amplo, abrangendo diversas áreas de conhecimento, com informações de termos e assuntos de forma centralizada, aberta e interoperável;

- Uso de ferramenta de geração automática de metadados aplicada em base de dados de vários sistemas de entidades governamentais, pois muitos problemas de organização da informação acontecem, devido os usuários não realizarem a entrada correta, completa dos dados e metadados, dificultando a recuperação da informação. É salutar para que a resolução desse tipo de incidente seja realizada por uma solução tecnológica e de forma automática. A implementação dessa abordagem diminuiria o tempo de recuperação da informação, impactando diretamente nos atendimentos ao cidadão ou implementação de políticas públicas.

E, por última, a oportunidade de desenvolvimento e adequação de uma solução tecnológica que integrasse através de módulos: a coleta de dados, metadados e documentos de repositório digital; adequação ou elaboração de um vocabulário controlado, ontologias e taxonomias; o auxílio para criação de um corpus de conhecimento, utilizando diversas bases

[21] Disponível em <https://www.gov.br/governodigital/pt-br/governanca-de-dados/vocabulario-controlado-do-governo-eletronico>, acesso em 27/02/2023.

de dados abertas; execução automática de sugestão de termos e assuntos; e organização, filtros de busca e recuperação facetada em repositório digital.

REFERÊNCIAS

ALUISIO, S. M; ALMEIDA, G, M. de B. **O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística**. Calidoscópio, Vol. 4, nº 3, p. 1156-178, set/dez 2006.

"***ANNIF – Tool for Automated Subject Indexing***". n.d. Disponível em http://annif.org/, acessado em 25 de janeiro de 2023.

ARAÚJO, A. K. S; MAIA, F. H; VECHIATO, F. L. **Encontrabilidade da informação em repositórios digitais: um estudo de caso na Biblioteca Digital de Monografias da UFNR**. Rev. Inf. na Soc. Contemp., Natal, RN, v.2, n1, jan./jun., 2018.

AUDICHYA, M, K; SAINI J, R. ***Computational linguistic prosody rule-based unified technique for automatic metadata generation for Hindi poetry***. 1st International Conference on Advances in Information Technology, 2019.

BANDIM, M. A. S; CORREA, R. F. **Indexação automática por atribuição de artigos científicos em português da área de Ciência da Informação**. Transinformação,v.31, e180004, 2019. http://dx.doi.org/10.1590/2318-0889201931e180004

BELKIN, N. J. ***Information concepts for information science***. *Journal of Documentation,* v. 34, n. 1, p. 55-85, Mar. 1978.

BELKIN, N; ROBERTSON, S. ***Information Science and the Phenomenon of Information***. *Journal of the American Society for Information Science* Jul/aug, v. 34, n.4, p 197-204, 1976.

BRASCHER, M; CAFÉ, L. **Organização da Informação ou Organização do Conhecimento?** *In*: IX Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação (IX ENANCIB), GT-02 - Organização e Representação do Conhecimento. São Paulo, 2008.

BORKO, H. ***Information Science: What is it?*** American Documentation, v.19, n.1, p.3-5, Jan. 1968.

BROOKES, Bertram. ***The foundations of information science: part I: philosophical aspects***. *Journal of Information Science*, v. 2, p. 125-133, 1980.

BUCKLAND, M. K. ***Information as thing***. *Journal of the American Society for Information Science*, v. 42, n. 05, p. 351-360, 1991. Disponível em: <http://people.ischool.berkeley.edu/~buckland/thing.html>. Acesso em: 22 ago. 2022.

BUCKLEY, W. ***Signals, meaning, and control in social systems***. In: MACHLUP, F.; MAUSTIELD, U. (Orgs.) *The study of information interdisciplinary menages* USA: John Wile & Jons Inc., 1983.

BUSH, V. ***As we may think***. Atlantic Monthly, v.176, 1, p.101-108, 1945. Disponível em <https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>, Acesso em: 14 out. 2022.

CAFÉ, L. C.; MUÑOZ, I. K. **Avaliação de usabilidade no repositório institucional da Universidade de Brasília**. Informação & Tecnologia, v. 3, n. 2, p. 39-61, 2016. Disponível em: http://hdl.handle.net/20.500.11959/brapci/40954. Acesso em: 22 fev. 2022.

CAPURRO, R; HJORLAND, B. ***The concept of information***. *Annual Review of Information Science and Technology*, v.37, n.8, p.343-411, 2003.

CERRAO, N. G; CASTRO, F. F: **Repositórios institucionais das Universidades Federais brasileiras: análise da representação da informação**. Informação & Tecnologia (ITEC), Marília/João Pessoa, v.5, n.1, jan./jun. 2018.

CONSELHO NACIONAL DE ARQUIVOS (CONARQ). **e-ARQ Brasil: Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos**. [recurso eletrônico] /Câmara Técnica de Documentos Eletrônicos. 2ª versão. Dados eletrônicos (1 arquivo: 1MB). Rio de Janeiro: Arquivo Nacional, 2022. Disponível em https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/EARQV205MAI2022.pdf, acesso em 15/02/2023.

CONSELHO NACIONAL DE ARQUIVOS (CONARQ). **Glossário: Documentos arquivísticos digitais**. [recurso eletrônico] /Câmara Técnica de Documentos Eletrônicos. 8ª versão. – Dados eletrônicos (1 arquivo: 1 MB). Rio de Janeiro: Arquivo Nacional, 2020. Disponível em <https://www.gov.br/conarq/pt-br/assuntos/camaras-tecnicas-setoriais-inativas/camara-tecnica-de-documentos-eletronicos-ctde/glosctde_2020_08_07.pdf>, acesso em 15/02/2023.

CORRÊA, R. F; LAPA, R. C. **Panorama de estudos sobre indexação automática no âmbito da ciência da informação no Brasil (1973-2012)**. Ciência da Informação, Brasília, DF, v. 42, p.255-273, mai/ago., 2013.

COSTA, A; FIRDAUSY, T, P; INNEREBNER, M; MONSORNO, R. ***EURAC SDI: A Near Real Time and Offline Automatic Metadata Generation Processing Chain***. GI Forum, Conference Proceedings, volume 1, 2013.

COX, A. ***How artificial intelligence might change academic library work: Applying the competencies literature and the theory of the professions***. Journal of the Association for Information Science and Technology, 74(3), 367–380. https://doi.org/10.1002/asi.24635

CRYSTAL, A; LAND, P. ***Metadata and Search: Global Corporate Circle DCMI 2003 Workshop***. 2003. Disponível em <http://www.dublincore.org/groups/corporate/Seattle/> (acessado em 24 de janeiro de 2023).

EGAN, M, E; SHERA, J, H. ***Foundations of a theory of bibliography***. The Library Quarterly, v. 22, n. 2, p. 125-137, 1952.

FERENHOF, H. A; FERNANDES, R. F. ***Desmistificando A Revisão de Literatura como Base para Redação Científica: Método SSF***. Revista Acb: Biblioteconomia em Santa Catarina, Florianópolis, v. 21, n. 3, p. 550-563, nov. 2016

FLORIDI, L. ***Open problems in the philosophy of information*** **Metaphilosophy**. Volume 35, n. Number 4, p. 554-582. Blackwell Publishing. 2004

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

GALVAO, M. C. B.; PLUYE, P.; RICARTE, I. L. M. Métodos de pesquisa mistos e revisões de literatura mistas: conceitos, construção e critérios de avaliação. **InCID: Revista de Ciência da Informação e Documentação**, *[S.l.]*, v. 8, n. 2, p. 4-24, 2017. DOI: 10.11606/issn.2178-2075.v8i2p4-24.

GALVÃO, M. C. B; RICARTE, I. L. M. **Revisão sistemática da literatura: conceituação, produção e publicação.** Logeion: Filosofia da Informação, v. 6, n. 1, p. 57-73, 2019.

GIDDENS, A. **Sociologia.** 6ª Edição, Tradução de Alexandra Figueiredo, Ana Patrícia Duarte Baltazar, Catarina Lorga da Silva, Patrícia Matos e Vasco Gil. Fundação Calouste GulBenkian, 725 p., 2007. Original Sociology 4ª Edição, Polity Press, 2001.

GIL LEIVA, I. ***Manual de indización: teoría y práctica***. Gijón: Trea, 2009

GOFFMAN, W. ***Information science: discipline or disappearance***? Aslib Proceedings, 22, 1970, 589-95.

GONZALO, P, R; MATT, H; GUNTHER, H, W; COLIN, O; KATIE, A; LAVANYA, R. ***ScienceSearch: Enabling Search through Automatic Metadata Generation*.** 2018 IEEE 14th International Conference on e-Science.

GREENBERG, J. ***Metadata Extraction an Harvesting: a comparison of two automatic metadata generation applications***. Journal of Internet Cataloging, vol. 6, (4), 2003.

GUSMÃO, F. C. M; SILVA, M. P. B; PEREIRA, G. M; LIMA, I. F; OLIVEIRA, H. P. C. **Elementos de arquitetura da informação no Repositório Eletrônico Institucional da UFPB**. Rev. Inf. na Soc. Contemp., Natal, RN, Número Especial, 2017.

HARPRING, P. ***Introduction to controlled vocabularies: terminology for art, architecture, and other cultural works***. Los Angeles: Getty Publications, 2010. Disponível em: <http://d2aohiyo3d3idm.cloudfront.net/publications/virtuallibrary/160606018X.pdf>, Acesso em: 15 fev. 2023.

HAYNES, D. ***Metadata for Information Management and Retrieval: understanding metadata and its use***. 2ª Edição, London: Facet Publishing, 2018.

HEDDEN, H. ***Taxonomies and controlled vocabularies best practices for metadata***. *In*: *Journal of Digital Asset Management*, vol. 6, no. 5, out, 2010.

ISO/IEC 17788:2014. ***Information Technology – Cloud Computing – Overview and Vocabulary***. ISO, 2014. Disponível em < https://www.iso.org/obp/ui/#iso:std:iso-iec:17788:ed-1:v1:en>, acesso em 5 fev 2023.

KHANDAGALE, J; XIAO, H; BABBAR, R. ***Bonsai: Diverse and Shallow Trees for Extreme Multi-label Classification***. *Mach Learn* 109, p. 2099–2119, 2020. https://doi.org/10.1007/s10994-020-05888-2

KITCHENHAM, B. ***Procedures for performing systematic reviews***. NICTA Technical Report 040001 IT.1, Keele University Technical Report TR/SE-0401, ISSN 1353-7776, 2004.

KIVUNJA, C. ***Distinguishing between Theory, Theoretical Framework, and Conceptual Framework: A Systematic Review of Lessons from the Field***. International Journal of Higher Education, Vol. 7, No. 6; 2018. https://doi.org/10.5430/ijhe.v7n6p44.

KLEPPE, M; VELDHOEN, S; WAAL-GENTENAAR, M. V. D; OUDSTEN, B. D; HAAGSMA, D. ***Exploration possibilities Automated Generation of Metadata***. 2019. Disponível em <https://doi.org/10.5281/zenodo.3375192> acesso em 02/02/2023.

KOVACEVIC, A; IVANOVIC, D; MILOSAVLJEVIC, B; KONJOVIC, Z. ***Automatic extraction of metadata from scientific publications for CRIS systems.*** Program: Electronic Library and Information Systems Vol. 45 No. 4, pp. 376-396, 2011.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. Brasília: Briquet de Lemos Livros. 452p., 2004.

LAPPALAINEN, M; HULKKONEN, J; INKINEN, J; KALLIO, A; LEHTINEN, M; KOSKELA, M; SJÖBERG, M; SUOMINEN, O; YETUKURI, L. ***Automaattisen sisällönkuvailun ohjelmiston rakentaminen – case Annif***. Signum, vol. 53, nº 4, 14–20, 2021.

LUHN, H, P. ***Key word-in-context index for technical literature (kwic index)***. American Documentation, v. 11, n. 4, p. 288-295, 1960.

MARATEA A; PETROSINO A; MANZO, M. ***Automatic Generation of SCORM Compliant Metadata for Portable Document Format Files***. International Conference on Computer Systems and Technologies – CompSysTech, 2012.

MARTINS, D. L; LEMOS, D; L; da S; ANDRADE, M. C. **Tainacan e Omeka: Proposta de análise comparativa de softwares para gestão de coleções digitais a partir do esforço tecnológico para uso e implantação**. Inf. Inf., Londrina, v. 26, n. 2, p. 569 – 595, abr./jun. 2021. Disponível em <http://repositorio.ufes.br/bitstream/10/11774/1/41208-221479-1-PB.pdf>, acesso em 15/02/2023.

MEDELYAN, O. ***Human-competitive automatic topic indexing***. Tese de Doutorado, Universidade de Waikato, Hamilton, Nova Zelandia, 2009. Disponível em https://hdl.handle.net/10289/3513

MIKHAILOV, A. I; CHERNYI, A. I; GILYAREVSKY, R. S. **Estrutura e principais propriedades da informação científica**. In: GOMES, H. E. (Org.). Ciência da Informação ou Informática? Rio de Janeiro: Calunga, 1980.

MIKSA, F. L. ***La bibliotecología y la ciência de la información: dos paradigmas***. Traductor: Rúben Urbizagástegui Alvarado. Revista Interamericana de Bibliotecología, Medéllin, Vol. 22, nº 2, Julio-Diciembre de 1999.

MOHER, D; LIBERATI, A; TETZLAFF, J; ALTMAN, D, G.***Preferred Reporting Items for Systematic Reviews and MetaAnalyses: The PRISMA Statement***.The PRISMA Group 2009.

PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097. Disponível em http://www.prisma-statement.org/, acesso em 12/02/2023.

MOOERS, C. ***Zatocoding applied to mechanical organization of knowledge***. American Documentation, v.2, n.1, 1951, p.20-32.

MORRIS, V. ***Automated Language Identification of Bibliographic Resources***. Cataloging & Classification Quarterly, 58:1, 1-27, 2020.

PARK, J; BRENZA, A. ***Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art***. Information Technology and Libraries, Volume 34, Ed. 3, p. 22-42, Chicago, USA, 2015.

PAVÃO, C. G; COSTA, J. S. B; FERREIRA, M. K; HOROWITZ, Z. **Metados e repositórios institucionais: uma relação indissociável para a qualidade da recuperação e visibilidade da informação**. PontodeAcesso, Salvador, v.9, n.2, p.103-116, dez. 2015.

ROBINSON, L; KARAMUFTUOGLU, M. ***The nature of information science: changing models***. Information Research, v. 15, n. 4, 2010.

PETTIGREW, K, E; MCKECHNIE, L. ***The Use of Theory in Information Science Research***. Jornal da Sociedade Americana para Ciência e Tecnologia da Informação, 52(1):62–73, 2001.

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

POLFREMAN, M; BROUGHTON, V; WILSON, A. ***Metadata Generation for Resource Discovery***. JISC, 2008. Disponível em <http://www.jisc.ac.uk/whatwedo/programmes/resourcediscovery/autometgen.aspx>, acesso em 20 jul 2021.

POMERANTZ, J. ***Metadata***. Cambridge, MA: The MIT Press, 2015.

PONTES, F. V; LIMA, G. A. B. de O. **A organização do conhecimento em ambientes digitais: aplicação da teoria da classificação facetada**. Perspectivas em Ciência da Informação. 17, (4), Dez 2012. https://doi.org/10.1590/S1413-99362012000400003

RAFFERTY, J; NUGENT, C; LIU, J. ***Automatic Metadata Generation Through Analysis of Narration Within Instructional Videos*****.** Transaction Processing Systems**,** *J Med Syst* nº 39**,** 94, 2015.

REINSEL, D; GANTZ, J; RYDNING, J. ***Data Age 2025: The Digitization of the world: from edge to core***. International Data Corp – IDC, Seagate, November 2018, Data refreshed May 2020. Disponível em: https://seagate.com/files/www-content/our-story/trends/files/dataage-idc-report-final.pdf . Acesso em: 12 fev. 2023.

SAH, M; WADE, V. ***Automatic metadata mining from multilingual enterprise content***. Web semantics: Science, services and agents on the world wide web, Vol 11, p. 41-62, 2012.

SALES, L. F; ROCHA, L. de L; CAVALCANTI, M. T. **Desenvolvimento de um vocabulário controlado para o repositório institucional CarpedIEN**. RECIIS - Revista Eletrônica de Comunicação, Informação e Inovação em Saúde, Rio de Janeiro, v. 11, p. 1-5, nov. 2017.

SANTARÉM SEGUNDO, J. E; MARCEL, F. S; MARTINS, D. L. **Revisitando a Interoperabilidade no contexto dos acervos digitais**. Inf. & Soc.:Est., João Pessoa, v.29, n.2, p. 61-84, abr./jun. 2019.

SANTOS, J. C. F dos; CERVANTES, B. M. N; FUJITA, M. S. L. **Tesauro Eletrônico: importação no Tematres e disponibilização na *web***. In: XIX Encontro Nacional de Pesquisa em Ciência da Informação, 22-26 de Outubro, Londrina, Enancib, 2018.

SANTOS, R. F. D. **Indexação em repositórios digitais: uma abordagem sobre o metadado assunto da biblioteca digital de monografias da UFRN**. Revista Informação na Sociedade Contemporânea, n. Especial, p. 1-22, 2017. Disponível em: http://hdl.handle.net/20.500.11959/brapci/106607. Acesso em: 22 jan. 2023.

SERACEVIC, T. ***Evaluation of Evaluation in information retrieval***. SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. p. 138–146. july, 1995a. https://doi.org/10.1145/215206.215351.

SARACEVIC, T. ***Interdisciplinary nature of information science***. *Ciência da Informação*, Brasília, v. 24, n. 1, 1995b.

SHANNON, C. E; WAEVER, W. ***The mathematical theory of communication***. University of Illinois Press, 1949.

SHINTAKU, M; GOMES, R. F; BRITO, R. F de; RODRIGUES, L; PEREIRA, V. C; SCHIMIDT, K. **Guia do Usuário do Omeka**. Brasília: Ibict, 2018.

SHINTAKU, M; VIDOTTI, S. A. B. G. **Bibliotecas e repositórios no processo de publicação digital**. Biblos: Revista do Instituto de Ciências Humanas e da Informação, v. 30, n.1, 2016.

SILVA, S. R. de BRITO; CORREA, R. F. **Sistemas de Indexação Automática por atribuição: uma análise comparativa**. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 25, p. 01-15, 2020. Universidade Federal de Santa Catarina. ISSN 1518-2924. DOI: https://doi.org/10.5007/1518-2924.2020.e70740

SILVA, S. R. de B; CORREA, R. F; GIL-LEIVA, I. **Avaliação Direta e Conjunta de Sistemas de Indexação Automática por Atribuição**. Inf. & Soc.:Est., João Pessoa, v.30, n.4, p. 1-27, out./dez. 2020.

SILVEIRA, L. A da; MACEDO, D. J; VECHIATO, F. L; SCHIESSL, I. T; SHINTAKU, M. SILVA, N . B. P; BRITO, R. F. de. **VuFind: uma ferramenta para recuperação da informação**. Brasília: Ibict, p. 110, 2019.

SINCLAIR, J. ***Developing Linguistic Corpora: a guide to good practice. AHDS, literature, languages and linguistics***. *Edited by* Martin Wynne, ISBN 1463-5194. 2004. Disponível em <https://users.ox.ac.uk/~martinw/dlc/>, acesso em 22/02/2023.

SUOMINEN, O. ***Annif: Feeding your subject indexing robot with bibliographic metadata***. Liber's 47th Annual Conference in Lille, France, Data Enhancements in the Service of Research Libraries, session 10, 2018.

SUOMINEN, O. ***Annif, l'indexation automatique à la Bibliothèque nationale de Finlande***. Ar(abes)ques, Bibliothèques de recherche en Europe, n°94 Juillet, août, septembre, 2019.

SUOMINEN, O. ***Annif: DIY Automated Subject Indexing Using Multiple Algorithms***. Liber Quarterly, vol. 29, 2019.

SUOMINEN, O; INKINEN, J; LEHTINEN, M. ***Annif and Finto AI: Developing and Implementing Automated Subject Indexing***. JLIS.it, vol. 13, nº 1, january, 2022. Disponível em <https://www.jlis.it/index.php/jlis/article/view/437/430>, acesso em 25 abr 2022.

TARAPANOFF, K; SUAIDEN, E; OLIVEIRA, C, L. **Funções Sociais e Oportunidades para Profissionais da Informação**. DataGramaZero - Revista de Ciência da Informação - v.3 n.5, out, 2002.

TAUBE, M; GULL, C. D; WACHTEL, I, S. ***Unit terms in coordinate indexing***. American Documentation (pre-1986), v. 3, n. 4, p. 213, 1952.

UNB; IBICT. Programa ColetadorOAI-sickle.py. 2022. Código fonte disponível em <https://github.com/tainacan/data_science/blob/master/FUNARTE/BIBLIOTECA_DIGITAL/ColetadorOAI-sickle.py>

VEGA-ALMEIDA, R. L; FERNÁNDEZ-MOLINA, J. C; LINARES, R. ***Coordenadas paradigmáticas, históricas y epistemológicas de la Ciencia de la Información: una sistematización***. Information Research, v. 14, n. 2, jun. 2009. Disponível em <https://informationr.net/ir/14-2/paper399.html>, Acesso em: 15 jan. 2023.

VERBORGH, R; VAN DEURSEN, D; MANNENS, E; POPPE, C; WALLE, R, V. ***Enabling context-aware multimedia annotation by a novel generic semantic problem-solving platform***. *Multimed Tools Appl* 61, 105–129, 2012.

VLACHIDIS, A; BINDING, C; MAY, K; TUDHOPE, D. ***Automatic Metadata Generation in an Archaeological Digital Library: Semantic Annotation of Grey Literature***. In: Przepiórkowski A., Piasecki M., Jassem K., Fuglewicz P. (eds) Computational Linguistics. Studies in Computational Intelligence, vol 458. Springer, Berlin, Heidelberg, 2013.

WANG, W. T; PONTES, J. ***Science, metascience, and information science***. In: VALENTIS, M. Techknowledgies: new imaginaries in the humanities, arts and technosciences. Newcastle: Cambridge Scholars Publishing, 2007. p. 256, 2007.

WERSIG, G; NEVELING, U. ***The phenomena of interest to Information Science***. The Information Scientist, v. 9, n. 4, Dec. 1975. Disponível em <https://sigir.org/files/museum/pub-13/18.pdf>, acesso em janeiro de 2023.

WIENER, N. ***Cibernética e sociedade*****: o uso humano de seres humanos**. Trad. José Paulo Paes. 4a. ed. São Paulo: Cultrix, 1954.

YANG, G; PARK, J. ***Automatic Extraction of Metadata Information for Library Collections*****.** International Journal of Advanced Culture Technology, Vol.6, nº 2, p. 117-122, 2018.

YIN, R. K. **Estudo de caso: planejamento e métodos**.[recurso eletrônico] / Robert K. Yin; [tradução: Cristhian Matheus Herrera]. – 5.ed – Porto Alegre: Bookman, 2015.

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