DOI http://dx.doi.org/10.1590/1981-5344/46629

FRAMEWORK GENÉRICO PARA GERAÇÃO AUTOMÁTICA DE ASSUNTOS E INDEXAÇÃO EM REPOSITÓRIO DIGITAL

Jean Carlos Borges Brito
http://lattes.cnpq.br/1005167212309269 – https://orcid.org/0000-0001-7421-1642
mailto:zuluaer@gmail.com
Universidade de Brasília (UnB)
Brasília, DF, Brasil

Dalton Lopes Martins
http://lattes.cnpq.br/3774617443225038 – http://orcid.org/0000-0002-6244-6791
mailto:dmartins@gmail.com
Universidade de Brasília (UnB)
Brasília, DF, Brasil

RESUMO
Este estudo tem por objetivo apresentar um *framework* genérico para geração automática de assuntos, utilizando técnicas de aprendizagem de máquina na ferramenta Annif. Posteriormente, executar a indexação de dados e metadados em repositório digital, propiciando a recuperação de registros através de busca facetada. Para alcance desse objetivo, aplicou-se o *framework* na área da Ciência da Informação, construindo um *corpus* de conhecimento, baseado em metadados de 438 artigos da Base Brasileira de Ciência da Informação (BRAPCI). Utilizou-se o Tesauro Brasileiro de Ciência da Informação (TBCI) como vocabulário controlado. Empregou-se a aplicação "coletor" desenvolvida em *phyton* para baixar metadados e arquivos completos de Dissertações e Teses de coleções existentes no Repositório Institucional da Universidade de Brasília (RiUnB). Após o processo de treinamento do modelo com Annif, foram executadas geração automática de assuntos e indexados em repositório digital Tainacan. Nesse repositório, foram criadas taxonomias baseadas no vocabulário controlado elaborado. Ao final, foi possível parametrizar buscas facetadas com possibilidade de o usuário inserir etiquetagem e ao mesmo tempo realizar navegação *web*, selecionando os termos da taxonomia facetada. Conclui-se que o *framework* genérico proposto pode ser aplicado em qualquer área de conhecimento, auxiliando na geração automática de assuntos, indexação em repositório digital e parametrização de taxonomias facetadas para recuperação da informação.

Palavras-Chave: Geração Automática de Assuntos. Indexação. Coleções. Repositório Digital. Busca Facetada.

GENERIC FRAMEWORK FOR AUTOMATIC SUBJECT GENERATION AND INDEXING IN A DIGITAL REPOSITORY

ABSTRACT
This study aims to present a generic framework for automatic subject generation, using machine learning techniques in the Annif tool. Subsequently, perform the indexing of data and metadata in a digital repository, providing the recovery of records through faceted search. To achieve this objective, the framework was applied in the area of Information Science, building a corpus of knowledge, based on metadata of 438 articles from the Base Brasileira de Ciência da Informação (BRAPCI). The Tesauro Brasileiro de Ciência da Informação (TBCI) was used as controlled vocabulary. The "collector" application developed in Phyton was used to download metadata and complete files of Dissertations and Theses from existing collections in the Institutional Repositório Institucional da Universidade de Brasília (RiUnB). After the model training process with Annif, subjects were automatically generated and indexed in the Tainacan digital repository. In this repository, taxonomies were created based on the elaborated controlled vocabulary. In the end, it was possible to parameterize faceted searches with the possibility for the user to insert labeling and at the same time perform web browsing, selecting the terms of the faceted taxonomy. It is concluded that the proposed generic framework can be applied in any area of knowledge, helping in the automatic generation of subjects, indexing in a digital repository and parameterization of faceted taxonomies for information retrieval.

Keywords: Automatic Subject Generation. Indexing. Collections. Digital Repository. Faceted Search.

DOI http://dx.doi.org/10.1590/1981-5344/46629

Recebido em: 25/06/2023.
Aceito em: 16/11/2023.

1 INTRODUÇÃO

Até 2025 o volume de dados aumentará em quatro vezes, se comparado a 2019, ampliando de 45 *zettabytes* para 175 *zettabytes*, conforme Reinsel, Gantz e Rydning (2018). Esses autores enfatizam que grande parte da economia atual depende de dados e a confiança só aumentará no futuro à medida que as entidades capturam, classificam, gerenciam e analisam os seus dados a partir de processos e tecnologias que aumentam a sua qualidade e permitam melhor exploração de seu valor agregado.

Nesse contexto, a busca e a recuperação de objetos de pesquisa nos repositórios digitais têm sido impactadas devido a diversos problemas relacionados aos metadados e a representação da informação nesses acervos. Greenberg (2003) discorre que o incremento de metadados com qualidade fornece valor agregado ao conjunto de dados, além de melhorar sua classificação e busca.

Polfreman, Broughton e Wilson (2008) enfatizam que sem os metadados apropriados, os recursos permanecem ocultos e sem uso, gerando desperdício de investimento. O autor também discorre que a baixa qualidade ou metadados inexistentes são igualmente eficazes para tornar os recursos inutilizáveis, tornando-se praticamente invisível dentro de um repositório e, portanto, fica desconhecido e inacessível.

Greenberg (2003) afirma que a geração automática de metadados (GAM), 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.

De acordo com Maratea, Petrosino e Manzo (2012), a 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.

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, 2019a). 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 pesquisa.

DOI http://dx.doi.org/10.1590/1981-5344/46629

Para criar metadados para um milhão de documentos deveriam ser alocados 60 empregados/ano para realizar essa tarefa (Crystal; Land, 2003), sendo considerado um trabalho ineficiente, oneroso e humanamente impossível de se executar, diante do aumento exponencial do volume de dados.

A problemática identificada neste estudo foi delimitada com a seguinte questão: "**Como a técnica de geração automática de metadados pode apoiar os usuários e os gestores de repositórios digitais na melhoria dos recursos de organização de informação visando facilitar a busca e recuperação da informação dos seus acervos?**".

Diante do cenário apresentado, esse estudo tem por objetivo propor um *framework* genérico a ser validado por um estudo de caso, aplicando um modelo de pesquisa na área da Ciência da Informação.

O referencial teórico e os estudos sobre as ferramentas de geração automática e semiautomática de metadados, assim como a seleção e descrição da solução para uso, foram descritos através de uma revisão sistemática da literatura demonstrada nos trabalhos apresentados no XXI Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB)[1] e XXII ENANCIB[2].

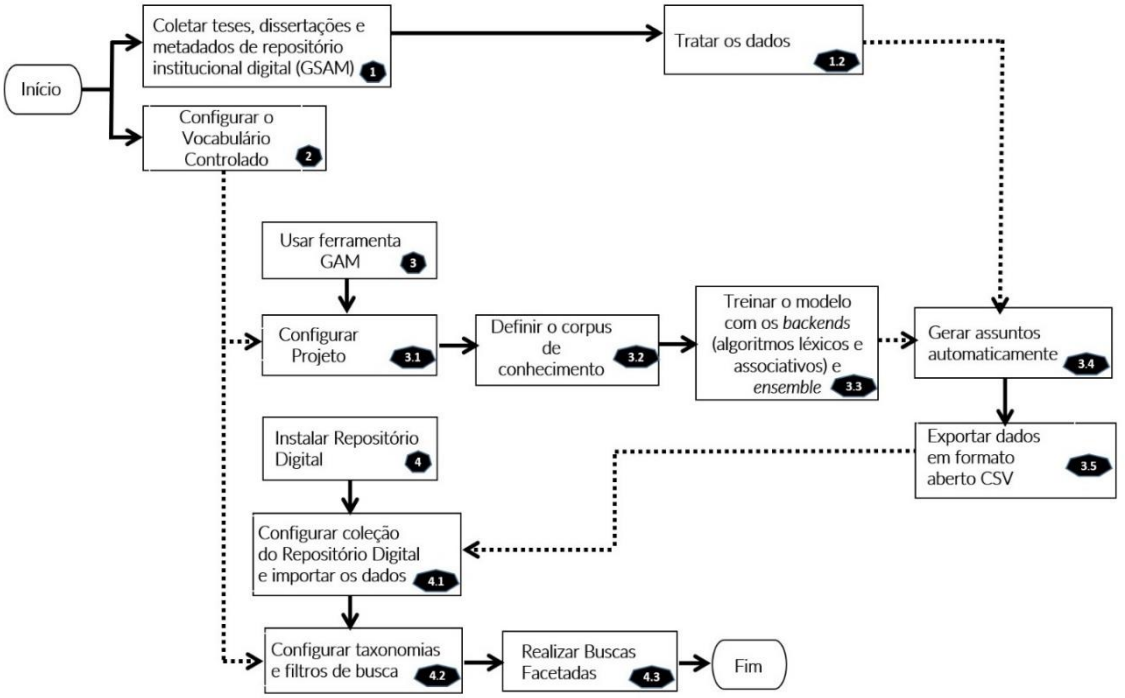
2 MÉTODO

A presente investigação é de natureza qualitativa com a tipologia descritiva-bibliográfica com o objetivo de demonstrar um *framework* genérico referencial para geração automática de assuntos, ampliando os componentes de representação de informação e sua indexação em um repositório digital real. Em seguida, será realizado um estudo de caso aplicando um modelo de pesquisa para validação do *framework* proposto.

O *framework* genérico referencial proposto na Figura 1, descreve o processo com as seguintes atividades:

[1] "**Geração automática e semiautomática de metadados: uma revisão sistemática de literatura**" (Brito; Martins, 2021).
[2] "**Geração automática de metadados: estudo de caso utilizando a técnica de indexação automática estatística com a ferramenta ANNIF"** (Brito; Martins, 2022).

Figura 1 – *Framework* genérico referencial proposto



Fonte: elaborado pelo autor (2023).

- **Passo 1** – Selecionar um repositório digital real para coleta de documentos e metadados através de uma ferramenta semiautomática para extração de metadados (GSAM);
- **Passo 1.2** – Tratar os dados, metadados e documentos obtidos no "Passo 1";
- **Passo 2** – Elaborar ou adequar um Vocabulário Controlado (VC) com lista de termos e assuntos relacionados com a temática em análise;
- **Passo 3** – Selecionar e instalar uma ferramenta de GAM;
- **Passo 3.1** – Configurar ou parametrizar o projeto na ferramenta GAM;
- **Passo 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);
- **Passo 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 (Passo 3.2) e do VC desenvolvido (Passo 2);

DOI http://dx.doi.org/10.1590/1981-5344/46629

- **Passo 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;
- **Passo 3.5** – Exportar os dados gerados no "Passo 3.4" em padrão aberto, preferencialmente com a extensão CSV;
- **Passo 4 –** Selecionar e instalar a ferramenta tecnológica para implantação do repositório digital;
- **Passo 4.1** – Configurar coleção no repositório digital e importar os dados gerados no "Passo 3.5 para essa coleção;
- **Passo 4.2** – Configurar no repositório digital as taxonomias e filtros de busca, a partir do VC executado no "Passo 2";
- **Passo 4.3** – Testar o repositório digital criado, executando buscas facetadas através dos filtros configurados.

2.1 Síntese da abordagem conceitual para o estudo de caso proposto.

Identificou-se problemas no **Repositório Institucional da Universidade de Brasília (RiUNB)**, explicitados por Café e Kafure 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)[3] para ser o objeto de análise.

Para coletar os dados, metadados e documentos completos de dissertações e teses da RiUnB, propõe-se a utilização do **ColetadorOAI**[4] que implementa o protocolo *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH). Essa solução foi desenvolvida pelo Laboratório de Inteligência de Redes da Universidade de Brasília (UnB), através de uma parceria entre o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e a Fundação Nacional de Artes (FUNARTE) em atendimento a meta 2 do Termo de Execução Descentralizada (TED) 01/2020 referente a implementação do repositório digital e da ferramenta de coleta, busca e recuperação da informação científica. É uma ferramenta de código aberto,

[3] Para mais informações acesse: https://repositorio.unb.br/handle/10482/5363.
[4] Para mais informações acesse: https://github.com/tainacan/data_science/blob/master/FUNARTE/BIBLIOTECA_DIGITAL/ColetadorOAI-sickle.py.

desenvolvida em *Python* e de fácil assimilação e implementação, sendo uma solução simples e leve, além de ser uma ferramenta semiautomática de metadados, que implementa a técnica de extração de conteúdo/colheita de *metatags.* Após a definição dos critérios de seleção das fontes de informação e o mapeamento dessas fontes, o ColetadorOAI executa a coleta dos produtos científicos a partir das fontes de informação disponíveis, sendo posteriormente realizado o tratamento dos dados coletados e a análise dos produtos científicos obtidos (IBICT; FUNARTE, 2022).

Para adequação de um VC sob a temática da Ciência da Informação, escolheu-se o **Tesauro Brasileiro de Ciência da Informação (TBCI)** desenvolvido por Pinheiro e Ferrez (2014), contendo os termos e assuntos comumente utilizados na área. Esse tesauro, também foi utilizado para a construção das taxonomias para configuração dos filtros no repositório digital para propiciar as buscar facetadas, possibilitando aos utilizadores do repositório digital a executar buscas com a utilização de filtros e atributos.

O *corpus* de conhecimento para o treino do modelo na temática da Ciência da Informação foi derivado de artigos publicados na **Base de Dados em Ciência da Informação (BRAPCI)**[5].

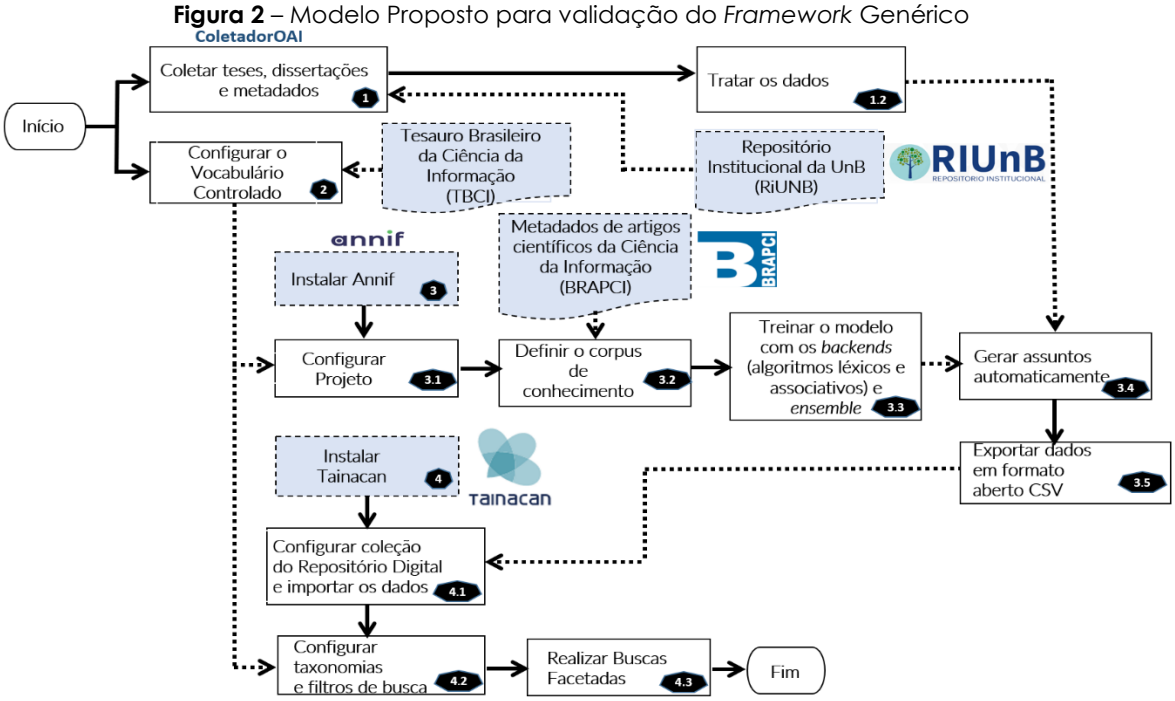
Por último, a partir dos estudos comparativos entre os repositórios digitais *Omeka*, *Dspace* e *Tainacan* executados por Shintaku *et al.* (2018) e Martins, Lemos e Andrade (2021), selecionou-se como repositório digital a ser aplicado neste estudo de caso, a solução **Tainacan**[6]. Essa ferramenta 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. O *software* é de ampla investigação e disseminação pelo Laboratório de Inteligência de Redes da UnB, possui um acervo documentário de tutoriais expressivos, além de vídeos explicativos e facilidade de uso.

Após a seleção de cada solução tecnológica para execução das atividades propostas, elaborou-se o modelo de pesquisa a seguir, que foi aplicado no presente estudo de caso para validação do *framework* genérico proposto:

[5] Para mais informações acesse: https://brapci.inf.br/.
[6] Para mais informações acesse: https://tainacan.org/.

DOI http://dx.doi.org/10.1590/1981-5344/46629

Figura 2 – Modelo Proposto para validação do *Framework* Genérico



Fonte: elaborado pelo autor (2023).

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), 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**.** O Quadro 1 a seguir, resume os elementos que foram utilizados neste estudo de caso:

Quadro 1 – 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 de 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*, *Structured Query Language* (MySQL), *Hypertext Preprocesso* (PHP), servidor apache. Posteriormente, importar dados para a coleção, configurar as taxonomias e os filtros para disponibilização das buscas facetadas.

Fonte: elaborado pelo autor (2023).

3 ANÁLISE E DISCUSSÃO DOS RESULTADOS

O *framework* genérico proposto nessa pesquisa possui o potencial de ser aplicado não somente na Ciência da Informação, mas em qualquer área de conhecimento. Na literatura acadêmica discorre que diversas ferramentas são desenvolvidas localmente para atender necessidades específicas.

Nesse contexto, os itens com a coloração azul na Figura 2 podem ser substituídos por qualquer ferramenta de geração automática/semiautomática de assuntos; coletor de teses e dissertações/pesquisa científica; repositório digital; e outras fontes de dados, tais como periódicos e revistas científicas para construção das taxonomias e tesauros, além do *corpus* de conhecimento.

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.

Uma limitação observada foi a ausência de um tesauro com interoperabilidade e acesso através de formatos abertos, tais como: *Resource Description Framework* (RDF), *eXtensible Markup Language* (XML), *Terse RDF Triple Language* (Turtle) ou *JavaScript Object Notation* (JSON). Tentou-se utilizar

DOI http://dx.doi.org/10.1590/1981-5344/46629

o TBCI online[7], mas não foi possível sua utilização devido a falta de conhecimento de sua estrutura. Nesse estudo utilizou-se o TBCI em *Portable Document Format* (PDF) para adequar um arquivo com 438 termos manualmente.

O Annif se demonstrou uma solução robusta e que pode auxiliar 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 pode fornecer. Em situações onde não há um VC 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 essa finalidade.

O Tainacan propiciou um ambiente fácil de instalação e administração, além de disponibilizar mecanismos para melhorar a organização do repositório digital, tais como: criação de taxonomias, filtros, metadados, itens e coleções, além de implementar funcionalidades de exportação e importação em massa, facilitando a manipulação de dados. A principal característica sob a ótica da gestão da informação da perspectiva do usuário e do gestor do repositório digital, encontra-se na funcionalidade de recuperação da informação através das buscas facetadas. Ao integrar as sugestões de assuntos de cada *backend* do Annif como um conjunto de termos dentro de uma taxonomia configurada e aplicá-las aos filtros de busca, considera-se aqui um achado importante.

3.1 Coletar teses, dissertações e metadados

A busca e a recuperação de objetos de pesquisa nos repositórios digitais têm sido impactadas devido a diversos problemas relacionados aos metadados e a representação da informação nesses acervos.

O RiUnB foi selecionado e foi utilizado para coleta de metadados e arquivos completos, com acesso aberto, das dissertações e teses na área da Ciência da Informação. Verificou-se que esse repositório oferece um modelo de

[7] TBCI online, para mais informações acesse: http://www.uel.br/revistas/informacao/tbci/vocab/index.php.

interoperabilidade, compartilhando o formato de coleta em um padrão de metadados que pode ser acessado via protocolo OAI-PMH.

De acordo com Oliveira e Carvalho (2009), o protocolo OAI-PMH, possibilita a comunicação entre sistemas que adotem o mesmo modelo de compartilhamento de dados, seja no formato de arquivo XML, seja no padrão de metadados *Dublin Core* (DC).

O código "ColetadorOAI-sickle.py"[8] foi adaptado para essa pesquisa. A adequação desse código propiciou a coleta de metadados de teses e dissertações armazenadas no repositório institucional da UnB. Para utilização do código foi necessário instalar as bibliotecas *Streamlit* e *Sickle*, além de configurar o arquivo "csv" com as definições do 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".

Quadro 2 – Parametrização do arquivo CSV a ser carregado no Coletador

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"

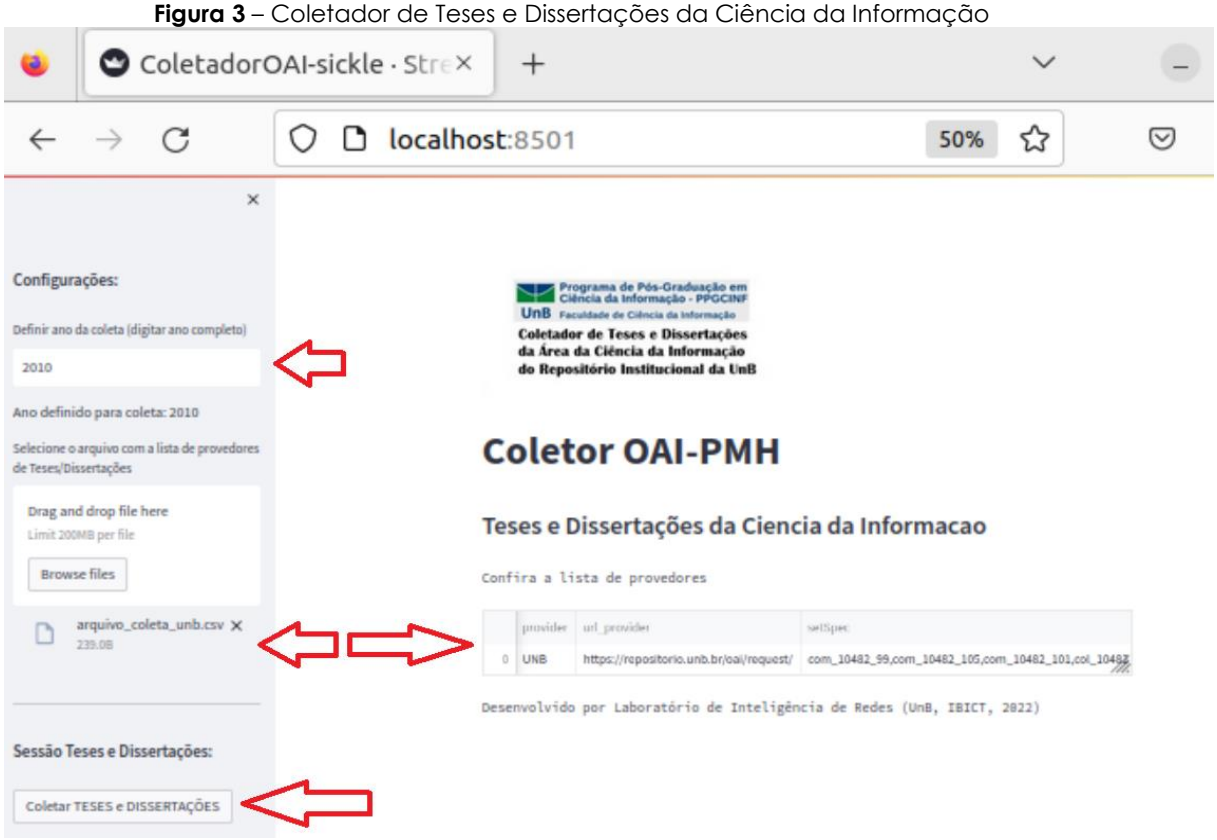
Fonte: elaborado pelo autor (2023).

Após as configurações descritas, iniciou-se a aplicação através do comando "$ *streamlit run ColetadorOAI-sickle.py*" e efetuou-se a carga do arquivo "csv", conforme ilustração a seguir:

[8] Código do coletor disponível em: https://github.com/tainacan/data_science/blob/master/FUNARTE/BIBLIOTECA_DIGITAL/ColetadorOAI-sickle.py.

DOI http://dx.doi.org/10.1590/1981-5344/46629

Figura 3 – Coletador de Teses e Dissertações da Ciência da Informação



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

A pesquisa é realizada por ano e gera como saída um arquivo "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.

3.2 Tratar os dados

Para a compreensão inicial dos dados, utilizou-se o *software Microsoft Excel®*, executando a importação dos dados do arquivo "csv" gerado pelo coletador. Parametrizou-se a codificação Unicode (UTF-8) e como delimitador o símbolo de vírgula. Dessa forma, foi possível verificar se os dados, metadados e arquivos eram relacionados com a área de conhecimento delimitada.

Para realizar análise e tratamento dos dados, criou-se um projeto no *software Jupyter Notebook*, utilizando a biblioteca *pandas* e importou-se o arquivo "csv".

Elaborou-se um código em *python* para remover o *Uniform Resource Identifier* (URI) do objeto "*identifier*" que direcionava ao recurso de informação do RiUnB, conforme o quadro a seguir:

Quadro 3 – Parametrização do arquivo CSV a ser carregado no Coletador

```
# Importando a biblioteca pandas
import pandas as pd
# Importando a base de Dados obtida via Coletador OAI-PMH
RiUNB = pd.read_csv('resultadoRiUNB_FCI.csv')
# 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 (2023).

Após a execução do código acima, gerou-se as URIs dos recursos, conforme o seguinte exemplo: https://repositorio.unb.br/handle/10482/44528. Em seguida, realizou-se o *download* dos arquivos "pdf" utilizando o comando: "$ *wget -nd -r -A pdf -i uri.csv -P artigosRiUnB/*"

Verificou-se que 66% dos trabalhos estavam com classificação restrita ou solicitavam credenciais de acesso para *download*. Dos 34% de registros com acesso aberto, definiu-se uma amostra de 26 registros para as ações ulteriores.

3.3 Configurar o Vocabulário Controlado

Lancaster (2004) discorre que um VC é uma lista de termos autorizados com uma forma de estrutura semântica (significado), que controlam sinônimos, distinguem homógrafos e agrupam termos afins. Como exemplo podemos citar o Tesauros, sendo esse tipo de VC definido como:

> "instrumento de organização do conhecimento, ou melhor, como **linguagens documentárias utilizadas no processo de indexação**, são listas estruturadas de termos e suas relações, onde cada um deve representar um único conceito ou ideia, de forma a **orientar indexadores** e usuários, levando-os de uma ideia ao termo que melhor a expresse. Desta forma, tesauros de diversos campos do saber vêm sendo publicados para facilitar a recuperação da informação" (Pinheiro; Ferrez, 2014, p. 9, grifo nosso).

DOI http://dx.doi.org/10.1590/1981-5344/46629

Utilizou-se o TBCI, extraindo termos para estruturação de uma base de dados que servirá para adequação de um VC para uso da ferramenta Annif na geração automática de assuntos e para o repositório Tainacan, na configuração da taxonomia.

Enquanto a estrutura do VC para Annif utiliza a estrutura "URI<tab>termo", a estrutura do VC para criação de taxonomias no Tainacan, utiliza apenas a descrição do "termo". Foi realizada uma adequação de um VC com 438 termos da área da Ciência da Informação, baseada no TBCI.

3.4 Instalar Annif

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[9] é uma solução mantida pela Biblioteca Nacional da Finlândia de código aberto e baseada em microserviço. A ferramenta foi apresentada pelo Sr. Osma Suominen (2018), 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*".

Confome Suominen (2018), o protótipo inicial foi desenvolvido em 2017 e vem sofrendo atualizações constantes e que estão sendo disponibilizadas no repositório do *GitHub*, 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 denominado *Representational State Transfer* –

[9] Suominen *et al*. (2022). Annif, Zenodo. Para mais informações acesse: https://doi.org/10.5281/zenodo.2578948.

Application Programming Interface (REST-API). Essa solução combina o uso de ferramentas de processamento de linguagem natural e aprendizagem de máquina, incluindo os algoritmos *Maui-like Lexical Matching* (MLLM), *STW Thesaurus for Economics – Finite-State-Automaton* (STWFSA) e *Term Frequency – Inverse Document Frequency* (TF-IDF).

A Máquina Virtual (VM) com a imagem do tutorial ANNIF, possui o Sistema Operacional Ubuntu (64-bit), Memória de Acesso Aleatório (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/.

3.5 Configurar Projeto

O ANNIF requer a configuração de um ou mais projetos para sua utilização. 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*[10] que é extraído dos registros de metadados presentes e/ou documentos indexados. 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.

Na ferramenta poderá haver vários *backends* independentes que auxiliam com sugestões de assuntos. Na prática, o projeto consiste em um conjunto de definições tais como: sua identificação (ID), descrição, linguagem/idioma, *backend*/algoritmo, VC, analisador.

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

Nesta etapa, realizou-se as parametrizações no arquivo "projects.cfg", configurando quatro projetos: brapci-mllm, brapci-stwfsa, brapci-tfidf e brapci-

[10] "[...] coleção de trechos de texto linguístico em formato eletrônico, selecionados de acordo com critérios externos para representar, na medida possível, uma língua ou variedade linguística como fonte de dados para pesquisa linguística" (Sinclair, 2004, tradução nossa).

DOI http://dx.doi.org/10.1590/1981-5344/46629

ensemble. Este último projeto, empacota os três algoritmos utilizados nos projetos anteriores, sendo possível a atribuição de pesos para cada algoritmo; é necessário parametrizar a linguagem a ser utilizada, neste caso será o português; o analisador será "*snowball*" e o nome do VC utilizado.

3.6 Definir o *corpus* de conhecimento

O *corpus* de conhecimento e documentos foi construído sobre os dados de título, resumo e palavras-chave de 438 artigos extraídos da BRAPCI. O formato para este arquivo foi: "Título do Artigo". "Resumo do Artigo". "Palavras-chave".tab<uri>.

Depois do *corpus* de conhecimento criado (*corpus*-brapci.tsv), executou-se o treinamento para os projetos: brapci-mllm, brapci-stwfsa e brapci-tfidf.

3.7 Treinar o modelo com os *backends* (léxicos e associativo) e *ensemble*

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 VC. Essa abordagem executa comparação utilizando poucos dados de treinamento. Exemplo: *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: *TF-IDF, fastText e Vowpal Wabbit*.

Nessa investigação definiu-se o uso dos algoritmos *MLLM, STWFSA e TFIDF*, além do *backend ensemble* que faz a fusão dos três anteriores para realização do treinamento do modelo;

Suominen, Inkinen e Lehtinen (2022) 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 VC.

O algoritmo *STWFSA* é um pacote em torno do *STWFSAPY*, sendo desenvolvido como parte do esforço da automatização da indexação de assuntos (AutoSE[11]) 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.

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 (2019b), TF-IDF é implementado 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 Suominen, Inkinen e Lehtinen (2022), 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.

[11] *Leibniz Information Centre for Economics Your partner for research and studies*, para mais informações acesse: https://www.zbw.eu/en/about-us/key-activities/automated-subject-indexing.

DOI http://dx.doi.org/10.1590/1981-5344/46629

3.8 Gerar assuntos automaticamente

De posse dos 26 trabalhos baixados no item 3.2 (Teses e Dissertações), a atividade seguinte foi converter os arquivos do formato "pdf" para formato "txt" através do comando "$ *pdftotext* arquivo_original.pdf arquivo_derivado.txt".

Para cada *backend*/algoritmo configurado, realizou-se a geração automática de 10 assuntos para cada documento, através da ferramenta Annif.

3.9 Exportar dados em formato aberto CSV

Ao final da atividade de geração automática de assuntos, exportou-se arquivo *CSV*, com as seguintes colunas: Título, Autor, Orientador, Descritores, Resumo, *Abstract*, Data de Publicação, Data de Defesa, Citação, Linguagem, Provedor, Tipo de Pesquisa, Formato, URI, URL, *Annif-Ensemble*, *Annif-MLLM*, *Annif-STWFSA* e *Annif-TFIDF* (as colunas com Annif, continham para cada registro multivalorado, o número de 10 sugestão de assuntos delimitados pelo símbolo "||").

3.10 Instalar Tainacan

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.

Nesse contexto, Martins *et al*., (2017) discorrem em seu artigo que havia uma necessidade de um *software* livre capaz de disponibilizar diversas funcionalidades com características de sistemas *Web* 2.0 (recursos para interação do usuário, compartilhamento de conteúdo, revisão de metadados, melhorias na descrição de conteúdo, dentre outros aspectos. Esse fato fez surgir a motivação para o apresentar o *software* livre Tainacan para a construção de repositórios digitais. Essa solução vem sendo utilizado na área de arte e cultura no Brasil para armazenamento de seus acervos.

Segundo Silva e Santarém Segundo (2019), o Tainacan é uma aplicação flexível e propicia o gerenciamento e a publicação de coleções digitais, fornecendo a parametrização de filtros de busca diferentes. O *software* fornece a facilidade de uso de um VC, configurado como taxonomia, auxiliando no controle de ambiguidades e sinônimos.

O Tainacan[12] funciona através de um *plugin* do *software* livre *Wordpress*, sendo necessário para a sua ativação, os seguintes pré-requisitos:

- Servidor *Web Apache*;
- Banco de dados MySQL;
- Interpretador PHP e suas bibliotecas;
- *Software* livre *Wordpress* instalado.

Após a instalação dos pré-requisitos, basta ativar o *plugin* do Tainacan no *Wordpress* para iniciar seu uso.

3.11 Configurar coleção do repositório digital e importar os dados

Ao iniciar o Tainacan, a primeira tarefa é configurar a criação de 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 3.9, foi realizada a exportação do arquivo final com todos os metadados de Teses e Dissertações da Ciência da Informação. Nesse sentido, esse mesmo arquivo será importado nessa etapa de criação da coleção no Tainacan.

Importante salientar que ao criar uma coleção a partir da importação de arquivo, é necessária a configuração dos 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

[12] Cf. Silva, Martins e Silva ([2014?]).

DOI http://dx.doi.org/10.1590/1981-5344/46629

de textos (caractere que delimita o campo de cada célula) e a codificação do arquivo.

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 Figura 4:

Figura 4 – Criação de Coleção e seus itens

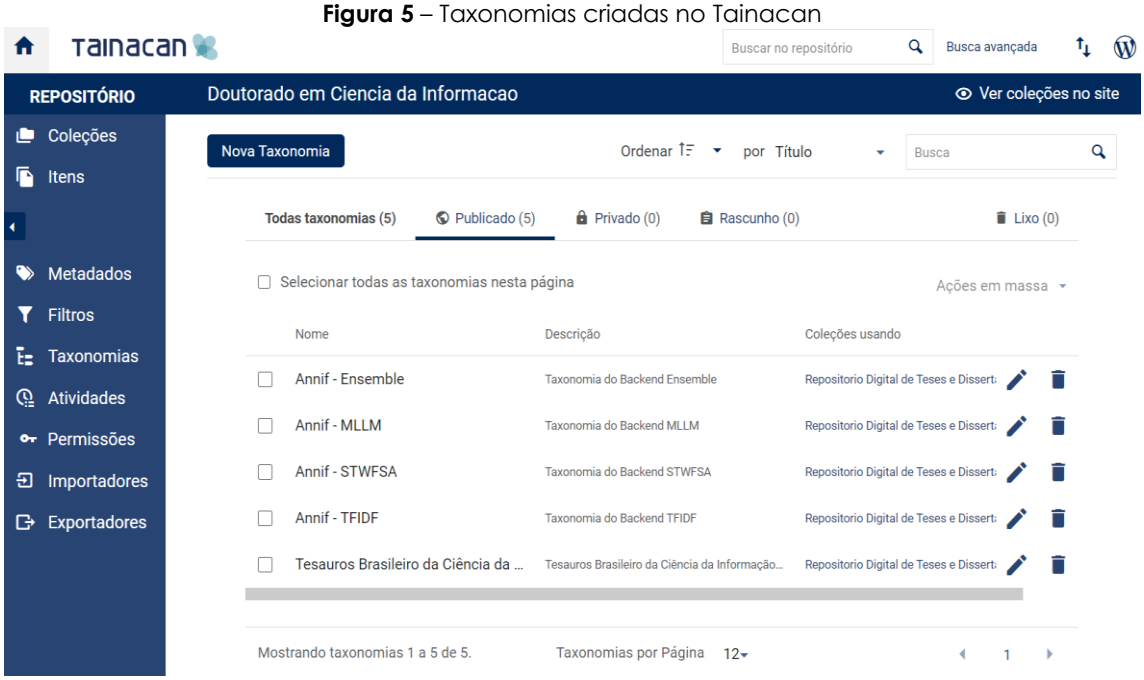
Fonte: elaborado pelo autor (2023).

3.12 Configurar taxonomias e filtros de busca

Uma taxonomia pode ser compreendida como um termo de um VC 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 obtidas, conforme Figura 5:

Figura 5 – Taxonomias criadas no Tainacan



Fonte: elaborado pelo autor (2023).

As demais taxonomias referentes aos *backends* do Annif foram extraídas do registro multivalorado da coluna correspondente ao arquivo exportado na seção 3.9.

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

3.13 Realizar buscas facetadas

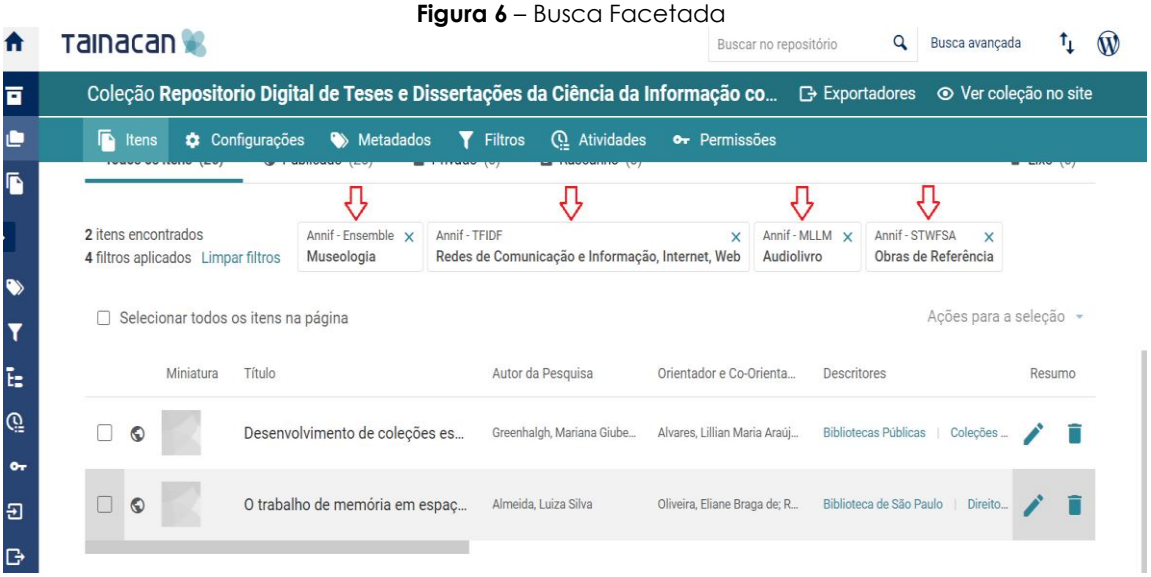
O Tainacan executa formas diferentes de classificação dentro da coleção, funcionando como facetas de busca, que melhoram sua utilização e facilita a construção do repositório pelo gestor do acervo digital.

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 novos termos de busca ao filtro aplicado, denominado de etiquetas. Nesse sentido, a ferramenta possibilita interação com o usuário, propiciando a personalização dos termos de busca.

DOI http://dx.doi.org/10.1590/1981-5344/46629

Figura 6 – Busca Facetada



Fonte: elaborado pelo autor (2023).

No exemplo representado na Figura 6, realizou-se a busca facetada aplicando-se quatro filtros, tendo como resultado dois 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 termos sugeridos de assuntos para cada *backend* do Annif, ou seja, no RiUnB tínhamos inicialmente de dois a sete descritores relacionados às palavras-chave dos trabalhos publicados. Após a aplicação do modelo proposto nessa pesquisa, adicionou-se uma quantidade maior de termos para indexação e uso através da busca facetada, conforme exemplificação demonstrada no Quadro 4, a seguir:

Quadro 4 – Comparação de descritores da RiUNB e Descritores sugeridos por Annif

Pesquisa	Descritores no RiUnB	Assuntos sugeridos por Annif
CÔRTE, Adelaide Ramos e. *Estilos gerenciais do profissional da informação de Biblioteconomia*. 1988. 187 f., Dissertação (Mestrado em Biblioteconomia e Documentação) – Universidade de Brasília, Brasília, 1988.	Bibliotecas, Administração, Gerência da Informação, Competência Gerencial	Educação Continuada, Encadernação, ABNT, Anuário, Obras de Referência, Arquivamento, Educação Básica, Instituições de Ensino e Pesquisa, Sistemas de Informação Gerencial, Publicação Oficial, Ensino Técnico, Profissão e

		Mercado de Trabalho, Documentos e Informação como Componente, Bibliotecas de Pesquisa, Formatos *Machine Readable Cataloging* (MARC).
WALTER, Maria Tereza Machado Teles. *Necessidade de informação dos técnicos de nível superior da Engevix Engenharia S/A. 1988*. 92 f., Dissertação (Mestrado em Biblioteconomia e Documentação) – Universidade de Brasília, Brasília, 1988.	Necessidade informacional, Engenheiros	Credibilidade, Organograma, Engenharia Elétrica, Engenharia Civil, ABNT, Plantas, Usuários e Uso da Informação, Serviços de Informação, Sistemas de Informação Geográficas, Documentos e Informação como Componente, Ambiguidade, Institutos de Pesquisa, Pertinência.
LIMA, Ana Claudia Cordeiro Correia. *Sistemas especialistas aplicados à ciência da informação*: tendências para um futuro próximo baseadas em um estudo infométrico da literatura. 1993. 133 f., Dissertação (Mestrado em Biblioteconomia e Documentação) – Universidade de Brasília, Brasília, 1993.	Ciência da informação, Bibliometria, Sistemas de informação,	Leis Bibliométricas, Sistema Especialista, Termos de Indexação, Publicação Eletrônica, Buscas em Linhas, Modelos de Recuperação da Informação, Redes de Citação, Classificação Automática, Métodos Matemáticos e Estatísticos, Alfabetação, Interação Humano-Computador, Coerência na Indexação, Confiabilidade, Sistemas de Organização do Conhecimento.
CORRÊA, Fernando Gabriel. *Contribuições do Princípio da Territorialidade para a resolução de contenciosos arquivísticos*. 2020. 210 f. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, Brasília, 2020.	Arquivologia, Documento Arquivístico, Organicidade	Arquivo Público, Microfilmagem, Transferência de Arquivos, Domínio Público, Conservação de Documentos, Normas e Protocolos, Credibilidade, Arquivos Privados, Documentos e Informação como Componente, Arquivistas, American *Association for Cancer Research* (AACR2), Documento Iconográfico.

Fonte: elaborado pelo autor (2023).

A sugestão de assuntos pela ferramenta Annif além de ampliar os filtros de busca facetada, auxiliou na recuperação da informação no repositório digital implementado de forma facilitada.

4 CONSIDERAÇÕES FINAIS

Retomamos a questão de pesquisa inicial: "**Como a técnica de geração automática de metadados pode apoiar os usuários e os 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 GAM auxiliam na sugestão de assuntos para documentos robustos como uma Tese ou Dissertação, ampliando o quantitativo de descritores, de modo a facilitar a configuração de taxonomias,

DOI http://dx.doi.org/10.1590/1981-5344/46629

filtros e facetas. Esse trabalho propõe 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.

A pesquisa possui a contribuição de demonstrar os benefícios da GAM, sendo que essa técnica pode:

- ajudar a aumentar a precisão e a consistência das descrições dos recursos;
- processar grandes volumes de informações em um curto período. Isso ajuda a acelerar o processo de organização e descrição dos recursos em um repositório digital. Em vez de depender exclusivamente de esforços manuais, a tecnologia automatizada pode analisar rapidamente o conteúdo e gerar metadados relevantes;
- complementar as descrições existentes dos recursos, adicionando informações adicionais e relevantes, enriquecendo assim as descrições e facilitando a busca e a recuperação posterior;
- suportar a pesquisa e recuperação de informações, através da criação de índices mais abrangentes e precisos, permitindo uma busca mais eficaz pelos usuários;
- permitir a criação de recomendações personalizadas, sugestões de conteúdo relacionado e experiências de navegação mais ricas.

No entanto, é importante notar que a GAM tem suas limitações. Os algoritmos podem cometer erros e algumas informações contextuais sutis podem ser perdidas. É importante que haja um VC e um *corpus* de conhecimento adequado e robusto para treinamento do modelo de aprendizado de máquina para minimizar essas limitações. Portanto, é recomendável que os metadados gerados automaticamente sejam revisados e validados por especialistas para garantir sua qualidade e precisão.

REFERÊNCIAS

BRITO, J. C. B; MARTINS, D. L. Geração automática e semiautomática de metadados: uma revisão sistemática de literatura. *In*: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO. 21., 2021, Rio de Janeiro, **Anais...** Rio de Janeiro: ANCIB, 2021. Disponível em https://brapci.inf.br/index.php/res/download/216427, Acesso em: 20 jan. 2023.

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. *In*: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 22., 2022, Porto Alegre. **Anais...** Porto Alegre: UFRGS: ANCIB, 2022. Disponível em https://enancib.ancib.org/index.php/enancib/xxiienancib/paper/viewFile/777/719, Acesso em: 20 jan. 2023.

CAFÉ, L. C.; KAFURE MUÑOZ, I. Avaliação de usabilidade no repositório institucional da Universidade de Brasília. **Informação & Tecnologia**, [s.*l.*], v. 3, n. 2, p. 39-61, 2016. Disponível em: http://hdl.handle.net/20.500.11959/brapci/40954. Acesso em: 20 jan. 2023.

CRYSTAL, A; LAND, P. **Metadata and Search**: Global Corporate Circle DCMI 2003 Workshop**.** [*s.l.*]: Dublin Core, 2003. Disponível em http://www.dublincore.org/groups/corporate/Seattle/. Acesso em: 20 jan. 2023.

GREENBERG, J. Metadata Extraction an Harvesting: a comparison of two automatic metadata generation applications. **Journal of Internet Cataloging**, [s.*l.*],v. 6, n. 4, 2003.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA [IBICT]; FUNDAÇÃO NACIONAL DE ARTES [FUNARTE]. **Repositório temático com foco na produção científica a respeito das artes no Brasil**: relatório referente à meta 2 do TED 001/2020 (IBICT e FUNARTE) – Implementação do repositório digital da ferramenta de coleta, busca e recuperação da informação da produção científica. Brasília, D.F.: IBICT: FUNART, 2022.

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

LAPPALAINEN, M. *et al.* Automaattisen sisällönkuvailun ohjelmiston rakentaminen – case Annif. **Signum**, Helsinki, v. 53, n. 4, p.14–20, 2021.

MARATEA A; PETROSINO A; MANZO, M. Automatic Generation of SCORM Compliant Metadata for Portable Document Format Files. *In*: INTERNATIONAL CONFERENCE ON COMPUTER SYSTEMS AND TECHNOLOGIES: COMPSYSTECH, 13., 2012, New York. **Proceedings...** New York: Association for Computing Machinery, 2012.

DOI http://dx.doi.org/10.1590/1981-5344/46629

MARTINS, D. L. *et al.* Repositório Digital com o software livre Tainacan: revisão da ferramenta e exemplo de implantação na área cultural com a revista filme cultura. *In*: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília. **Anais**… Marília: ANCIB, 2017.

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: https://brapci.inf.br/index.php/res/download/161686. Acesso em: 15 fev. 2023.

OLIVEIRA, R. R; CARVALHO, C. L de. **Implementação de Interoperabilidade entre Repositórios Digitais por meio do Protocolo OAI-PMH**. Goiás: Universidade Federal de Goiás, 2009. (Technical report, RT-INF_003-09),

PAVÃO, C. G. *et al*. Metadados 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.

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

POLFREMAN, M; BROUGHTON, V; WILSON, A. **Metadata Generation for Resource Discovery**. [*s.l.*]: AHDS, 2008. Disponível em: https://textarchive.ru/c-2308841-pall.html. Acesso em: 20 jan. 2023.

REINSEL, D; GANTZ, J; RYDNING, J. **Data Age 2025**: the digitization of the world, from edge to core. International Data Corp – IDC, [*s.l.*]: Seagate, 2018,. Disponível em: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. Acesso em: 20 jan. 2023.

SHINTAKU, M; *et al*. Guia do Usuário do Omeka. Brasília, D.F.: IBICT, 2018. Disponível em: https://ridi.ibict.br/handle/123456789/1157. Acesso em: 22 fev. 2023.

SILVA, E. A; MARTINS, D. L; SILVA, M. F. **Tainacan**: Manual do Usuário. [Goiás]: L3P: FIC; UFG: Ministério da Cultura, [2014?]. Disponível em: https://tainacan.org/wp-content/uploads/2017/02/Manual-Repositorio.pdf. Acesso em 19 jan. 2023.

SILVA, L. C.; SANTAREM SEGUNDO, J. E. Componentes de representação da informação em ambientes de informação digital: estudo do sistema de organização do *software* Tainacan. *In*: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 18., 2019, Florianópolis. **Anais…** Florianópolis: ANCIB, 2019.

SINCLAIR, J. Developing Linguistic Corpora: a guide to good practice. *In*: WYNNE, M. **AHDS literature, languages and linguistics**. [*s;l.*]: AHDS, 2004. Disponível em: https://users.ox.ac.uk/~martinw/dlc/chapter1.htm. Acesso em: 22 fev. 2023.

SUOMINEN, O. *et al.* **Annif**. Version v1.0.0. [*s.l.*]: Zenodo, 2022. DOI: https://doi.org/10.5281/zenodo.2578948.

SUOMINEN, O. Annif: Feeding your subject indexing robot with bibliographic metadata. Annual Conference in Lille, France, Data Enhancements in the Service of Research Libraries, 47., 2018, Lille. **Proceedings…** Lille: [*s.n.*], 2018. (session 10),

SUOMINEN, O. Annif, l'indexation automatique à la Bibliothèque nationale de Finlande. **Ar(abes)ques**: Bibliothèques de recherche en Europe, [*s.l.*], [*s.n.*], n; 94, juil./sept. 2019a.

SUOMINEN, O. Annif: DIY Automated Subject Indexing Using Multiple Algorithms. **Liber Quarterly**, [*s.l.*], v. 29, [*s.n.*], 2019b.

SUOMINEN, O; INKINEN, J; LEHTINEN, M. Annif and Finto AI: Developing and Implementing Automated Subject Indexing. **JLIS.it**, [*s.l.*], v. 13, n. 1, janv. 2022.

UNIVERSIDADE DE BRASÍLIA [UNB]; INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA [IBICT]. **Software ColetadorOAI-sickle.py**. Brasília, D.F: UNB: IBICT, 2022. 1 recurso digital. Disponível em: https://github.com/tainacan/data_science/blob/master/FUNARTE/BIBLIOTECA_DIGITAL/ColetadorOAI-sickle.py. Acesso em: 20 jan. 2023.

YIN, R. K. **Estudo de caso**: planejamento e métodos. Tradução: Cristhian Matheus Herrera. 5. ed. Porto Alegre: Bookman, 2015.