Artigo Original

Encontros Bibli

AVALIAÇÃO DO SISTEMA DE INDEXAÇÃO AUTOMÁTICA ANNIF EM ARTIGOS DE PERIÓDICOS EM CIÊNCIA DA INFORMAÇÃO

Evaluation of the Annif automatic indexing system in information science journal articles

Remi Correia Lapa
Programa de Pós-Graduação em Ciência da Informação, Universidade Federal de Pernambuco, Recife, PE, Brasil.
remi.lapa@ufpe.br
https://orcid.org/0009-0000-7333-0121

Renato Fernandes Correa
Departamento de Ciência da Informação, Universidade Federal de Pernambuco, Recife, PE, Brasil.
renato.correa@ufpe.br
https://orcid.org/0000-0002-9880-8678

A lista completa com informações dos autores está no final do artigo

RESUMO

Objetivo: Avaliar o desempenho da ferramenta de indexação automática Annif na atribuição de descritores a artigos científicos na área de Ciência da Informação.
Método: Adota uma abordagem empírica e de caráter exploratório. A pesquisa utilizou um *corpus* composto por 60 artigos escritos em português e previamente indexados manualmente com o Tesauro Brasileiro de Ciência da Informação (TBCI). O *corpus* foi dividido em duas subseções: 30 artigos foram destinados ao treinamento e 30 ao teste. O Annif foi configurado com o *backend* MLLM, e a avaliação foi realizada por comparação com os termos da indexação intelectual aplicando as métricas de Precisão, Revocação, F1 e Consistência. Também foi realizada uma análise qualitativa dos termos gerados.
Resultado: O sistema demonstrou um desempenho moderado, alcançando uma Precisão de 49,52%, Revocação de 50,1%, F1 de 49,23% e Consistência de 31,63%.
Conclusões: Conclui-se que o Annif é viável como ferramenta de apoio em fluxos semiautomáticos de indexação, desde que suas sugestões sejam validadas por indexadores humanos.
PALAVRAS-CHAVE: Annif; aprendizado de máquina; indexação automática; repositórios digitais; Tesauro Brasileiro de Ciência da Informação.

ABSTRACT

Objective: This study evaluates the performance of the Annif automatic indexing tool in assigning descriptors to scientific articles within the field of Information Science.
Methods: The research adopted an empirical and exploratory approach. It used a corpus composed of 60 articles written in Portuguese and previously indexed manually with the Brazilian Thesaurus of Information Science (TBCI). The corpus was divided into two subsets: 30 articles were used for training and 30 for testing. Annif was configured with the MLLM backend and compared with the terms assigned through intellectual indexing using the metrics of Precision, Recall, F1-score, and Consistency. A qualitative analysis of the generated terms was also conducted.
Results: The system demonstrated moderate performance, achieving a Precision of 49.52%, Recall of 50.1%, F1-score of 49.23%, and Consistency of 31.63%.
Conclusions: It is concluded that Annif is a viable tool for supporting semi-automatic indexing workflows, provided that its suggestions are validated by human indexers.
KEYWORDS: Annif; machine learning; automatic indexing; digital repositories. Brazilian Thesaurus of Information Science.

1 INTRODUÇÃO

A crescente disponibilidade de informações digitais impõe desafios à organização e à recuperação do conhecimento. As bibliotecas e os repositórios digitais, responsáveis por lidar com grandes volumes de metadados, historicamente têm recorrido à indexação manual baseada em vocabulários controlados. Embora eficiente do ponto de vista da padronização e da consistência semântica, esse processo é oneroso, demanda alta especialização e esforço humano, além de consumir um tempo considerável (Medelyan, 2009; Suominen, Inkinen, Lehtinen, 2022).

Nesse cenário, a indexação automática (IA) se apresenta como uma alternativa tecnológica para ampliar a escalabilidade dos processos de representação temática, reduzindo custos operacionais e tempo de processamento, sem comprometer a qualidade da recuperação da informação.

A literatura recente evidencia avanços no desenvolvimento de ferramentas baseadas em aprendizado de máquina para indexação automática, bem como sua aplicação em diferentes contextos institucionais e linguísticos. Entre as ferramentas desenvolvidas para essa finalidade, encontra-se o Annif, sistema de código aberto desenvolvido em 2017 pela Biblioteca Nacional da Finlândia, que tem sido progressivamente estudado por instituições de diversos países.

No cenário brasileiro, pesquisas iniciais com o Annif utilizando o Tesauro Brasileiro de Ciência da Informação (TBCI) indicam potencial promissor para aplicação em repositórios digitais, mas ainda carecem de avaliações sistemáticas com um *corpus* robusto (Brito; Martins, 2023a, 2023b). Adicionalmente, observa-se uma escassez de estudos que avaliem a qualidade da indexação automática do Annif em língua portuguesa aplicada a artigos científicos, especialmente considerando as particularidades terminológicas e estruturais do domínio da Ciência da Informação.

Diante desse cenário, formula-se a seguinte pergunta de pesquisa: Como o sistema de indexação automática Annif se comporta na atribuição de descritores do Tesauro Brasileiro de Ciência da Informação a artigos científicos em língua portuguesa, considerando métricas quantitativas de desempenho e análise qualitativa dos termos gerados?

Este artigo tem como objetivo geral avaliar o desempenho da ferramenta de indexação automática Annif na atribuição de descritores a artigos científicos na área de

Ciência da Informação em língua portuguesa. Para tanto, como objetivos específicos, busca-se:

a) configurar o Annif com o *backend* MLLM (*Maui-like Lexical Matching*) e o vocabulário TBCI para processamento de textos em português;
b) quantificar o desempenho do sistema mediante métricas de Precisão, Revocação, F1 e Consistência;
c) analisar qualitativamente os termos sugeridos, identificando padrões de acertos, erros e omissões;
d) comparar os resultados obtidos com estudos prévios que utilizaram o algoritmo Maui em *corpus* similar.

A realização deste estudo justifica-se pela necessidade de aprofundar a compreensão sobre o desempenho de sistemas de indexação automática em cenários linguísticos ainda pouco explorados, como a língua portuguesa no domínio da Ciência da Informação. Embora a literatura internacional registre avanços significativos, a aplicação dessas tecnologias nesse momento ainda carece de investigações empíricas sistemáticas.

Ademais, o estudo apresenta um diferencial metodológico ao empregar um padrão-ouro validado no domínio da Ciência da Informação brasileira, composto por indexação manual realizada por especialistas familiarizados com as especificidades terminológicas do TBCI. A combinação entre análise quantitativa, baseada em métricas consagradas na literatura e que permite comparações diretas com estudos que empregaram o algoritmo Maui (Silva; Corrêa; Gil-Leiva, 2020), e análise qualitativa dos termos gerados possibilita uma compreensão mais abrangente do fenômeno investigado.

2 REVISÃO DE LITERATURA

A indexação automática constitui um dos temas centrais da Ciência da Informação, especialmente no que se refere ao tratamento e à recuperação de conteúdos em bibliotecas digitais, repositórios institucionais e bases de dados acadêmicas importantes para a mediação do acesso ao conhecimento registrado. Estudos clássicos acentuam que a função primordial da indexação consiste em representar conteúdos de forma consistente e normalizada, de modo a favorecer a recuperação da informação (Lancaster, 2004; Gil-Leiva; Alonso-Arroyo, 2007).

Nesse processo, os vocabulários controlados, como os tesauros e os sistemas de classificação, são instrumentos empregados para assegurar padronização terminológica e interoperabilidade semântica na indexação de assunto, promovendo conexão entre a linguagem documental e a linguagem de busca.

Na literatura, distinguem-se duas abordagens principais de indexação automática, uma lexical, que relaciona as palavras do texto diretamente com os descritores de um vocabulário controlado, e outra, estatística ou associativa, fundamentada no aprendizado de padrões de correlação entre palavras do *corpus* com os termos de um vocabulário controlado. Essa distinção é destacada em diferentes estudos, como os de Medelyan (2009) e Toepfer e Seifert (2020). Enquanto os indexadores humanos desenvolvem e aperfeiçoam a habilidade de indexação por meio do treinamento e da experiência profissional, nos algoritmos este processo pode ser simulado mediante técnicas de aprendizado de máquina (Medelyan, 2009), o que justifica a centralidade dessas abordagens na literatura especializada.

Entre as ferramentas que têm se destacado no campo da indexação automática, encontra-se o Annif. Trata-se de um *software* de código aberto e arquitetura modular que integra diferentes algoritmos de aprendizado de máquina (*backends)*, possibilitando o treinamento supervisionado com vocabulários controlados e a integração em sistemas institucionais por meio de Interface de Programação de Aplicações (APIs). O sistema ainda permite combinações híbridas (*ensembles)*, o que amplia sua versatilidade em coleções de diferentes tipos de documentos (Suominen; Inkinen; Lehtinen, 2022).

No sistema Annif, o *backend* corresponde ao algoritmo de aprendizado de máquina responsável pela análise e classificação automática dos documentos. Entre os de abordagem lexical destacam-se o STWFSA, baseado em autômatos finitos ponderados, e o MLLM, inspirado no sistema Maui. Já entre os de abordagem associativa, figuram o TF-IDF, que utiliza vetores de palavras ponderadas por frequência e raridade; o fastText, que combina *embeddings* de palavras e modelo de rede neural; e o Omikuji/Bonsai, que utiliza árvores de decisão para classificação em larga escala.

Além disso, o Annif suporta o uso de *ensembles*, técnica que combina os resultados de múltiplos modelos de *backends* para gerar predições mais robustas do que as obtidas por modelos isolados. Outro recurso relevante é a disponibilização de uma API REST, que possibilita a integração com sistemas bibliotecários e repositórios digitais por meio de requisições HTTP padronizadas, facilitando a inserção do Annif no fluxo de trabalho em ambientes institucionais de indexação.

O Annif incorpora ambas as vertentes, lexical e associativa, permitindo a configuração de modelos isolados ou de forma híbrida via *ensembles*, mitigando as limitações de cada abordagem individual. O Quadro 1 apresenta um panorama sintético dos principais *backends* suportados pelo Annif, destacando seus tipos, requisitos e características operacionais. Esse quadro contribui para compreender as possibilidades de configuração da ferramenta e os diferentes níveis de complexidade associados a cada algoritmo.

Quadro 1 – *Backends* suportados pelo Annif

Nome	Tipo	Dependências adicionais	Descrição	Docs. de treinamento recomendados	Tamanho ideal do texto	Suporta otimização (*hyperopt*)
TF-IDF	Associativo	Não	Algoritmo de base que utiliza pesos TF-IDF; útil para testes e configurações iniciais	-	Curto / longo	Não
fastText	Associativo	Sim	Modelo baseado em *embeddings* de palavras e subpalavras (*n-grams*) com redes neurais	10.000+	Curto / longo	Não
Omikuji	Associativo	Sim	Classificação multilabel em larga escala por árvore de decisão	10.000+	Curto / longo	Não
SVC	Associativo	Não	Classificação linear do tipo *Support Vector Classification*; não *multilabel*	-	-	Não
MLLM	Lexical	Não	Implementação similar ao sistema Maui, baseada em correspondência lexical	100 - 10.000	Longo	Sim
Ensemble	Ensemble	Não	Combina resultados de múltiplos *backends* por média ponderada	N/A	N/A	Sim

Nome	Tipo	Dependências adicionais	Descrição	Docs. de treinamento recomendados	Tamanho ideal do texto	Suporta otimização (*hyperopt*)
NN-nsemble	Ensemble	Sim	Combina resultados por rede neural, otimizando pesos entre modelos	1.000 - 10.000	Longo	Não
PAV	Ensemble	Não	*Ensemble* dinâmico e treinável que combina resultados de múltiplos projetos	-	-	Não
YAKE	Lexical	Sim	Extração não supervisionada de palavras-chave; não requer treinamento	N/A	Longo	Não
STWFSA	Lexical	Sim	Tradutor estatístico baseado em autômatos de estados finitos ponderados	-	Curto / longo	Não
HTTP	Especial	Não	Permite comunicação via API REST com outra instância do Annif ou serviço externo	N/A	N/A	Não
Dummy	Especial	Não	Retorna resultados fixos, usado para testes internos e validações	N/A	N/A	N/A
X-Transformer	Associativo	Sim	É um algoritmo de rankeamento *extreme multilabel text classification* (XMTC) baseado em modelos *Transformer* tipo BERT que faz parte do *framework* PECOS. Sua integração com o Annif é experimental.	10.000+	Longo	Sim

Fonte: Adaptado de Inkinen (2025) e Suominen, Inkinen e Lehtinen (2025).

Essa diversidade de arquitetura confere ao Annif uma flexibilidade que permite sua adaptação a diferentes documentos e idiomas. Em ambientes com vocabulários extensos e bem estruturados, os modelos lexicais, como o MLLM, tendem a oferecer maior precisão conceitual; por outro lado, em cenários de dados ruidosos ou terminologias não normalizadas, os modelos associativos (como o fastText e o Omikuji) demonstram melhor desempenho em termos de revocação e generalização.

A adoção de *ensembles* representa um avanço na busca por equilíbrio entre precisão e cobertura semântica, ao combinar a força dos modelos lexicais com a adaptabilidade dos modelos associativos. Já os *backends* especiais, como o HTTP e o Dummy, exercem funções complementares de integração, teste e validação técnica, reforçando o caráter experimental e interoperável do Annif.

Aplicações do Annif em diferentes domínios de aplicação têm sido relatadas em variados estudos internacionais. Golub *et al*. (2024), descrevem sua aplicação em tarefas de geração automática de descritores em bibliotecas suecas usando sistema de classificação baseados na *Dewey Decimal Classification* (CDD). Hahn (2021) investigou métodos de catalogação semiautomática de assuntos em BIBFRAME em projetos de *linked data*. Lappalainen *et al.* (2021) destacam o uso do Annif em múltiplos idiomas, incluindo finlandês, inglês e sueco em repositórios acadêmicos. Mitra e Mukhopadhyay (2023) abordaram seu uso em domínios especializados de baixo recurso. No Brasil, Brito e Martins (2023a; 2023b) investigaram aplicações iniciais do Annif com o Tesauro Brasileiro de Ciência da Informação (TBCI), indicando potencial promissor, embora ainda careçam de avaliações sistemáticas com um *corpus* robusto.

Apesar dos avanços, a literatura ressalta limitações relacionadas à influência da língua, à qualidade do *corpus* de treinamento e ao vocabulário adotado (Kerketta; Mukhopadhyay, 2025; Dermentzi *et al.*, 2025).

A avaliação da indexação automática baseia-se, em geral, na comparação com padrões-ouro da indexação intelectual, utilizando métricas clássicas de recuperação da informação, como Precisão (P), Revocação (R) e a medida F1, além de indicadores de ranqueamento, como o *Normalized Discounted Cumulative Gain* (NDCG). Complementarmente, análises qualitativas de termos e estudos de impacto no processo da indexação permitem avaliar a utilidade prática das sugestões em situações reais de aplicação (Suominen, 2019; Golub *et al.*, 2016).

Estudos internacionais têm examinado o desempenho do Annif em múltiplos domínios e idiomas, variando de *corpus* bibliográficos a registros multimídia, inclusive em condições de baixo recurso, ou seja, cenários com vocabulários restritos, escassez de dados de treinamento ou ausência de ferramentas linguísticas maduras para determinado idioma ou domínio especializado (Lappalainen *et al.*, 2021; Mitra; Mukhopadhyay, 2023; Golub *et al.*, 2024). Esses trabalhos convergem em dois aspectos: a eficácia do Annif na identificação de conceitos centrais e a necessidade de sua aplicação em processos

semiautomáticos de indexação, em que a curadoria humana assegura a relevância e a coerência semântica (Suominen; Inkinen; Lehtinen, 2022).

O Quadro 2 sintetiza os estudos que avaliaram ou aplicaram o Annif em diferentes contextos, revelando três tendências principais.

Quadro 2 – Estudos sobre a ferramenta Annif

Autor(es) / Ano	**Idioma e tipologia do *corpus***	**Vocabulário controlado**	**Descrição**
Suominen (2019)	Inglês; artigos e registros de repositórios universitários	YSO – Yleinen Suomalainen Ontologia (Ontologia Geral Finlandesa)	Demonstra o potencial do Annif em repositórios universitários, mostrando que parte significativa das sugestões foi adotada por bibliotecários. Introduz a arquitetura modular e a integração via serviço web, ressaltando ganhos práticos no fluxo de trabalho e a necessidade de um *corpus* mais representativo e de refinamento contínuo.
Lehtonen e Piukkula (2020)	Finlandês; registros de programas de rádio e TV da base Ritva	YSO – Yleinen Suomalainen Ontologia (Ontologia Geral Finlandesa)	Avaliaram a aplicação do Annif em indexação de programas audiovisuais em grande escala. Mostram que, treinado com um *corpus* substancial de descrições, o sistema atribuiu termos relevantes de forma consistente. Destacam o potencial do sistema para ampliar a acessibilidade e a recuperação temática de acervos audiovisuais, ressaltando a necessidade de ajustes contínuos de treinamento e de curadoria do vocabulário.
Hahn (2021)	Inglês; registros bibliográficos no formato MARC para entidades de trabalho em bibliotecas	LCSH (Library of Congress Subject Headings)	Investigou métodos semiautomáticos em BIBFRAME. Utilizou o Annif para sugerir descritores do LCSH, articulando-o a *linked data*. Destacou a viabilidade da abordagem *librarian-in-the-loop*. Privilegiou análises de integração e usabilidade, sem foco em métricas clássicas.
Lappalainen *et al.* (2021)	Multilíngue (finlandês, inglês, sueco); registros de bibliotecas, arquivos e museus	YSO – Yleinen Suomalainen Ontologia (Ontologia Geral Finlandesa) e vocabulários específicos de domínio	Apresentam os desafios e soluções no desenvolvimento do Annif em bibliotecas, arquivos e museus. Documentam aspectos técnicos da arquitetura, o uso de múltiplos algoritmos e *ensembles*. Enfatizam a importância da avaliação contínua por métricas (precisão, revocação, F1, NDCG) e da participação humana em processos semiautomáticos.

Autor(es) / Ano	Idioma e tipologia do *corpus*	Vocabulário controlado	Descrição
Suominen, Inkinen e Lehtinen (2022)	Multilíngue (finlandês, sueco, inglês); registros institucionais e catálogos	YSO – Yleinen Suomalainen Ontologia (Ontologia Geral Finlandesa) e vocabulários nacionais integrados ao Finto AI	Descrevem o desenvolvimento do Annif e do serviço Finto AI, destacando sua integração em fluxos automatizados (ex. DSpace). Relatam desafios técnicos e operacionais de escalabilidade e manutenção em ambientes multilíngues.
Suominen e Koskenniemi (2022)	Multilíngue (finlandês, sueco e inglês); *corpus* de teses (University of Jyväskylä), descrições de livros (Kirjavälitys Oy) e registros de descoberta (Finna.fi)	YSO (incluindo o YSO Places) e YKL - Classificação das Bibliotecas Públicas da Finlândia	Compararam nove métodos de normalização lexical (*stemming* e lematização) no Annif. Testaram *backends* MLLM, Omikuji Parabel e Bonsai. Concluíram que lematização avançados (ex. stanza) superam abordagens básicas em finlandês e sueco, enquanto em inglês métodos simples (*Snowball*) se mostraram eficazes. O estudo orientou a integração do Simplemma como módulo do Annif.
Brito e Martins (2023a)	Português; 438 títulos, resumos e palavras-chave de artigos (BRAPCI) e texto completo de teses e dissertações do Repositório da Universidade de Brasília (RiUnB)	Termos do Tesauro Brasileiro de Ciência da Informação (TBCI)	Propõem um *framework* genérico para a geração automática de assuntos em repositórios digitais. Destacam a importância de um vocabulário controlado e um *corpus* adequado e robusto para treinamento e recomendam que os metadados gerados automaticamente sejam revisados e validados por especialistas para garantir a qualidade e precisão.

Autor(es) / Ano	Idioma e tipologia do *corpus*	Vocabulário controlado	Descrição
Brito e Martins (2023b)	Português; 52 títulos, resumos e palavras-chave de artigos (BRAPCI) e texto completo de tese do Repositório Institucional da UnB (RiUnB)	Termos do Tesauro Brasileiro de Ciência da Informação (TBCI)	Estudo de caso que aplicou o Annif (*backend* MLLM) em dois níveis: (i) treinamento com 52 artigos da BRAPCI; (ii) teste preliminar com texto completo de tese do RiUnB para validar um *framework* de indexação. Os resultados mostraram boa similaridade entre termos sugeridos e palavras-chave do autor, mas sem comparações sistemáticas entre diferentes *backends*.
Ahmed, Mukhopadhyay e Mukhopadhyay (2023)	Inglês e hindi; registros bibliográficos MARC	LCSH em *Linked Open Data* (LOD)	Aplicaram o Annif em um ambiente experimental de bibliotecas, testando diferentes *bakcends* (TF-IDF, Omikuji-Parabel, Omikuji-Bonsai e NN-Ensemble). O estudo evidenciou a viabilidade da indexação semiautomatizada baseada em IA/ML, mas ressaltou a dependência da qualidade do *corpus* de treinamento e a necessidade de supervisão humana.
Mitra e Mukhopadhyay (2023)	Inglês; corpus de treinamento composto por mais de 137 mil registros obtidos de bases abertas (CoRE, CrossRef, OpenAlex, Semantic Scholar) relacionados ao domínio LGBTQIA+	Homosaurus (versão 3.3, SKOS)	Desenvolveram um sistema semiautomático de indexação para um domínio de baixo recurso em bibliotecas indianas. Avaliaram múltiplos *backends* no Annif (TF-IDF, Omikuji Bonsai/Parabel e *Ensemble* Neural Network), destacando que tanto o Omikuji quanto o ensemble baseado em redes neurais obtiveram desempenho superior ao TF-IDF, alcançando métricas. Ressalta ainda o potencial do Annif quando aliado a vocabulários especializados.
Ahmed (2023)	Inglês; registros bibliográficos e metadados do AGRIS (artigos, livros, relatórios técnicos, teses e anais de conferências)	AGROVOC – Tesauro Multilíngue	Desenvolve um quadro experimental de indexação automática para o domínio da agricultura, integrando os sistemas AGRIS, AGROVOC e Annif. O estudo treina o Annif com um conjunto de dados de cerca de 99,99% do corpus de 0,82 milhões de registros e testa o Annif com cerca de 0,01% dos 0,82 milhões de registros. O sistema utiliza *backends* como TF-IDF, fastText, Omikuji e redes neurais para gerar descritores automáticos com base em títulos e resumos.

Autor(es) / Ano	Idioma e tipologia do *corpus*	Vocabulário controlado	Descrição
Golub *et al.* (2024)	Sueco e inglês; registros bibliográficos do catálogo nacional Libris	CDD (Classificação Decimal de Dewey)	Aplicaram o Annif à CDD em um catálogo nacional. Compararam múltiplos *backends*: lexicais (TF-IDF e BM25), além de SVC, fastText, Omikuji Bonsai e combinações via *ensemble*. O melhor desempenho foi obtido com o *ensemble*, complementado por análise qualitativa sobre o valor das classes automáticas.
Dermentzi *et al.* (2025)	Multilíngue (17 idiomas, com predominância de inglês, alemão, francês, hebraico e russo) em descrições arquivísticas do EHRI Portal (metadados de coleções, registros em níveis de coleção a item, baseados no padrão ISAD(G))	EHRI Terms, tesauro específico do Holocausto, estruturado em SKOS e expandido para 913 termos em 12 línguas	Estudo comparativo que avalia abordagens estatísticas, lexicais, *ensembles* (incluindo o NN Ensemble do Annif) e modelos de LLMs (*zero-shot* e *fine-tuning*, como BERT multilíngue). Os resultados mostraram que o *ensemble* neural do Annif obteve desempenho elevado em métricas de F1, mas os LLMs refinados também se destacaram, revelando potencial para uso em fluxos semi-automáticos de indexação em larga escala. O estudo conclui que, embora promissoras, tais ferramentas são mais adequadas em *workflows* híbridos (humano-máquina), em linha com os princípios FAIR e a necessidade de maior interoperabilidade entre acervos dispersos.
Kerketta e Mukhopadhyay (2025)	Inglês; 213.879 registros bibliográficos em formato MARC-21 (títulos, resumos e descritores)	CDD (Classificação Decimal de Dewey)	Exploraram o uso do Annif para predição automática de números de classificação na CDD, a partir de registros bibliográficos em formato MARC-21. Treinaram o sistema com um grande corpus e testaram múltiplos *backends* (fastText, Omikuji, SVC e um *ensemble* (NN-Ensemble). O modelo *ensemble* apresentou desempenho superior, avaliado por métricas como F1@5 e NDCG.
Kivimaki *et al.* (2025)	Inglês e Finlandês; descrições de cursos universitários (títulos, objetivos de aprendizagem e conteúdos curriculares)	YSO – Yleinen Suomalainen Ontologia (Ontologia Geral Finlandesa)	Estudo experimental que aplicou o Annif para geração automática de palavras-chave em descrições de cursos, resultando em 36.124 termos (3.504 únicos). Obteve acurácia de até 64% no *corpus* em inglês e cerca de 53% em finlandês. Concluiu que o Annif pode apoiar a indexação semiautomática de currículos e catálogos acadêmicos, mas a validação humana é indispensável para garantir precisão e pertinência semântica.

Autor(es) / Ano	Idioma e tipologia do *corpus*	Vocabulário controlado	Descrição
Poley *et al.* (2025)	Alemão e Inglês; publicações digitais e impressas da coleção da Biblioteca Nacional da Alemanha (livros, teses, periódicos, e publicações eletrônicas em PDF/EPUB)	DDC (Dewey Decimal Classification) e GND (*Gemeinsame Normdatei*)	O estudo descreve a evolução e a implementação da catalogação automática de assuntos (*Automatic Subject Cataloguing*) na Deutsche Nationalbibliothek (DNB), detalhando os processos de classificação e indexação baseados em *machine learning* e abordagens léxicas integradas ao Annif. Apresenta o sistema EMa (Erschließungsmaschine), uma arquitetura modular e de código aberto que combina modelos supervisionados (Omikuji, SVC) e léxicos (MLLM) para atribuir categorias DDC e descritores GND.
Suominen, Inkinen e Lehtinen (2025)	Corpus bilíngue (alemão e inglês) composto por registros bibliográficos do banco de dados TIBKAT, abrangendo diferentes tipos documentais (artigos, livros, teses)	GND (*Gemeinsame Normdatei*)	Apresenta a participação do *Annif* na tarefa SemEval-2025 Task 5 (LLMs4Subjects), voltada à indexação automática de assuntos em ambiente bilíngue. O estudo combina métodos tradicionais de aprendizado de máquina (Omikuji/Bonsai, MLLM e XTransformer) com técnicas baseadas em LLMs, utilizadas para tradução e geração sintética de dados de treinamento. O artigo também enfatiza a vantagem do baixo custo computacional dos métodos tradicionais e a importância de ampliar o corpus de treinamento e a avaliação qualitativa em futuras pesquisas.

Fonte: Elaborado pelo autor (2025).

Primeiramente, verifica-se que o desempenho da ferramenta está fortemente condicionado pela escolha do *backend* e pela qualidade do *corpus* e do vocabulário controlado. Estudos que utilizam vocabulários amplos e bem estruturados, como o YSO (Suominen, 2019; Lehtonen; Piukkula, 2020; Lappalainen *et al.*, 2021), relatam resultados consistentes e aplicabilidade prática em processos institucionais de indexação. Por outro lado, pesquisas com vocabulários mais restritos ou em domínios com vocabulários pouco desenvolvidos, como o Homosaurus (Mitra; Mukhopadhyay, 2023), evidenciam tanto o potencial quanto às limitações do sistema, reforçando a necessidade de ajustes de *corpus* e curadoria contínua.

Em segundo lugar, a literatura mostra que a combinação de *backends* via *ensemble* tende a superar modelos individuais. Essa constatação é recorrente em investigações recentes que compararam métodos lexicais (MLLM) com abordagens associativas e híbridas (TF-IDF, fastText, Omikuji, SVC, *ensembles* neurais), como em Golub *et al.* (2024), Kerketta e Mukhopadhyay (2025) e Dermentzi *et al.* (2025). Os resultados indicam que o *ensemble* amplia a robustez da indexação, especialmente em cenários multilíngues e de grande escala, embora ainda demande validação humana para garantir pertinência semântica.

Por fim, observa-se que a adoção mais realista do Annif ocorre em processos semiautomáticos de indexação, nos quais autores e bibliotecários validam ou refinam os termos sugeridos. Essa abordagem é documentada tanto em experimentos com o catálogo nacional sueco Libris, que reúne registros bibliográficos de bibliotecas universitárias e públicas e testou o Annif para atribuição automática de classes da CDD em larga escala (Golub *et al.*, 2024), quanto em investigações aplicadas ao cenário brasileiro, em que o sistema utilizou parte do Tesauro Brasileiro de Ciência da Informação (TBCI), e foi treinado com resumos de artigos científicos e testado em textos completos de teses e dissertações (Brito; Martins, 2023a). Também em sistemas de *linked data*, como o BIBFRAME, no qual o Annif foi empregado para sugerir descritores interoperáveis em registros MARC, explorando o potencial de integração semântica entre dados bibliográficos (Hahn, 2021).

Em todos os casos, a supervisão humana permanece indispensável para assegurar precisão, coerência e relevância terminológica. A revisão da literatura, contudo, revela lacunas importantes no campo brasileiro: os estudos existentes ainda se restringem a amostras reduzidas, sem comparações sistemáticas entre diferentes *backends* e sem avaliações baseadas em padrões-ouro, o que limita o entendimento do desempenho do Annif em cenários de produção científica nacional.

3 METODOLOGIA

Este estudo adota uma abordagem empírica de caráter exploratório voltado à avaliação direta do desempenho do Annif na indexação automática em língua portuguesa. Para tanto, utilizou-se um *corpus* composto pelo texto completo de 60 artigos científicos da área de Ciência da Informação, previamente organizados e indexados intelectualmente por Silva e Corrêa (2019) e compilados como padrão-ouro em Silva e Corrêa (2020).

O vocabulário controlado adotado foi o Tesauro Brasileiro de Ciência da Informação (TBCI), convertido para o formato SKOS, possibilitando sua utilização no Annif. O *corpus* foi dividido em dois subconjuntos: 30 documentos destinados ao treinamento do modelo e os outros 30 empregados na fase de teste para a avaliação empírica.

O Annif foi configurado com o *backend* MLLM, uma implementação baseada no sistema Maui, selecionado por sua relevância na literatura em tarefas de extração de termos com vocabulários controlados e por possibilitar comparações diretas com resultados prévios do Maui adaptado para o português como mostra o trabalho de Silva, Corrêa e Gil-Leiva, 2020. O pré-processamento inclui a remoção de *stopwords* e a radicalização morfológica (*stemming*) por meio do algoritmo *snowball* para português, configurado no Annif como parte do *pipeline* linguístico.

A configuração do experimento foi registrada no arquivo *projects.cfg*, que define os parâmetros de cada projeto no Annif. Cada instância relaciona um vocabulário controlado, um *backend*, e os critérios específicos da tarefa de indexação. O Quadro 3 apresenta os parâmetros utilizados neste estudo.

Quadro 3 – Explicação dos parâmetros

Parâmetro	**Valor**	**Descrição**
name	crci-mllm-pt	Identificador único do projeto.
language	pt	Código de idioma (ISO 639-1) para o português.
backend	mllm	Especifica o uso do *backend* MLLM.
vocab	tbci-skos	Referência no projeto que carrega o vocabulário controlado (TBCI) no formato SKOS.
analyzer	snowball (portuguese)	Define o *pipeline* de pré-processamento: tokenização, remoção de *stopwords* e aplicação do algoritmo *snowball stemmer* para o português.
limit	10	Número máximo de descritores sugeridos para cada documento.

Fonte: Baseado em Suominen (2019) e Suominen, Inkinen e Lehtinen (2022).

O analisador linguístico foi incorporado para melhorar a performance do modelo. O *stemming* normaliza variantes morfológicas, mapeando palavras como ‘bibliotecas e biblioteca’ para o radical ‘bibliotec’, ampliando a capacidade de correspondência com os

termos padronizados do TBCI. Já a remoção de *stopwords* elimina palavras frequentes e pouco informativas (*de, para, o*), contribuindo para maior precisão.

A avaliação seguiu o protocolo de comparação direta proposto por Silva, Corrêa e Gil-Leiva (2020). Para cada documento do conjunto de teste, os descritores sugeridos pelo Annif foram comparados com o padrão-ouro contendo os termos da indexação intelectual. Foram calculadas as métricas tradicionais de: Precisão (P), percentual de descritores sugeridos que estão corretos (presentes no padrão-ouro); Revocação (R), percentual de descritores do padrão-ouro sugeridos pelo sistema; Medida F1, média harmônica entre Precisão e Revocação; Consistência, percentual de sobreposição entre os descritores automáticos e do padrão-ouro em relação ao total de termos únicos da união de descritores automáticos e intelectuais.

Além da análise quantitativa, realizou-se uma avaliação qualitativa, voltada a identificar os padrões de desempenho do sistema. Foram analisados os acertos (casamentos dos termos), quando há correspondências diretas entre o Annif e o padrão-ouro; os erros (falsos positivos), onde os termos são atribuídos pelo MLLM mas ausentes do padrão-ouro; e as omissões (falsos negativos), em que os termos presentes no padrão-ouro não são recuperados pelo sistema.

Conforme as diretrizes éticas do periódico, declara-se que o ChatGPT (versão GPT-5 Thinking mini) foi utilizado de forma pontual e supervisionada exclusivamente para revisão gramatical e clareza textual do conteúdo deste artigo, sem interferência no conteúdo científico.

4 RESULTADOS

Os resultados médios para o conjunto de 30 documentos de teste (D31 a D60) são apresentados na Tabela 1. O valor de Precisão (49%) indica que aproximadamente metade dos termos sugeridos pelo Annif coincidiu com os descritores atribuídos no padrão-ouro. A Revocação (50,10%) mostra que o sistema recuperou metade dos termos relevantes presentes no padrão-ouro, enquanto a Medida F1 (49,23%) expressa um equilíbrio moderado entre ambas as métricas. Já a Consistência (31,63%) evidencia que aproximadamente um terço dos termos totais foi compartilhado entre a indexação automática e a intelectual, revelando divergências significativas na representação

semântica. As métricas individuais de cada documento indexado automaticamente podem ser consultadas no Apêndice A, que apresenta detalhadamente os resultados por registro.

Tabela 1 – Resultados médios de desempenho do Annif (*backend* MLLM)

Métrica	Valor Médio (%)
Precisão (P)	49,00
Revocação (R)	50,10
Medida F (F1)	49,23
Consistência	31,63

Fonte: Elaborado pelos autores (2025).

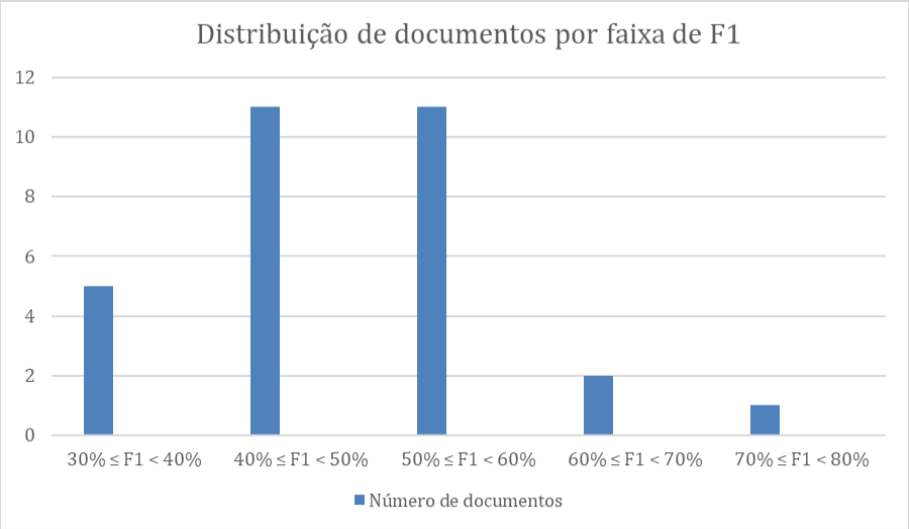
A comparação com o estudo de Silva, Corrêa e Gil-Leiva (2020), que avaliou a qualidade da indexação automática com o mesmo *corpus* utilizando o algoritmo Maui em sua implementação adaptada para o português, indica que o Annif configurado com o *backend* MLLM apresentou um desempenho similar (F1 49% vs 50%). Essa pequena diferença observada pode estar relacionada às etapas adicionais incorporadas pelo *framework* Annif, como o pré-processamento textual, o mapeamento e a normalização de termos no vocabulário controlado, corroborando as observações de Suominen (2019) sobre a influência da arquitetura modular na variação de desempenho do sistema.

A análise da distribuição dos valores de F1 por documento (Figura 1) revela que dos 30 documentos avaliados, 5 situam-se entre 30% e 40%, indicando razoável correspondência com a indexação humana; 11 ficaram entre 40% e 50% e outros 11 entre 50% e 60%, demonstrando que a maior parte dos casos se concentrou em faixas intermediárias de média harmônica. Apenas dois documentos alcançaram entre 60% e 70%, e um único caso superou os 70%, revelando que resultados de alto alinhamento com a indexação intelectual foram raros.

Essa distribuição evidencia que o Annif tem boa capacidade para recuperar conceitos centrais dos textos, mas ainda apresenta lacunas na cobertura semântica e na precisão, o que limita seu uso como substituto da indexação humana. Em termos aplicados, esses resultados sugerem que a ferramenta pode atuar como suporte eficaz à indexação semiautomática, fornecendo listas iniciais de termos que precisam ser validados ou refinados por indexadores humanos.

A partir desse ponto, a análise qualitativa é organizada em três eixos, acertos, erros e omissões, a fim de detalhar a natureza do desempenho observado. Os resultados da indexação automática obtidos com o Annif (*backend* MLLM), em contraste com o padrão-ouro de indexação humana, encontram-se apresentados no Apêndice B.

Figura 2 – Distribuição de documentos por faixa de F1



Fonte: Elaborado pelos autores (2025).

4.1 ACERTOS

Considera-se acerto toda correspondência direta entre os descritores sugeridos automaticamente pelo Annif e os termos presentes no padrão-ouro, isto é, a indexação intelectual. Nesse grupo, o sistema apresentou boa capacidade de identificar conceitos centrais e estruturantes da área da Ciência da Informação, como ‘ciência da informação’, ‘gestão da informação’, ‘sociedade da informação’, ‘bases de dados’, ‘bibliometria’ e ‘comunicação científica’, que foram frequentemente sugeridos e coincidiram com a indexação humana.

Em diversos casos, o sistema demonstrou a capacidade em apresentar os conceitos relevantes, como em D32 por exemplo, em que foram atribuídos automaticamente os descritores ‘bases de dados’, ‘recuperação da informação’, ‘controle de vocabulário’ e ‘estratégias de busca’, coincidindo com o padrão-ouro e revelando profundidade semântica e precisão terminológica.

Outro aspecto positivo foi a habilidade em reconhecer variações terminológicas e sinônimos contextualmente adequados. Em D35, o Annif identificou *HTML* e *XML*, que

estão diretamente associados ao domínio de ‘linguagens de marcação’. Outro exemplo ocorreu em D36, onde o sistema sugeriu corretamente ‘bibliometria’, ‘lei de Lotka’ e ‘produtividade de autor’, alinhando-se à indexação humana.

De forma geral, os acertos demonstram que o Annif, configurado com o *backend* MLLM, apresenta um bom desempenho na identificação de conceitos centrais e frequentes da área da Ciência da Informação, especialmente aqueles que aparecem no TBCI e de forma recorrente no *corpus* de treinamento.

4.2 ERROS (RUÍDOS)

Nesta categoria, foram incluídos os casos em que os descritores sugeridos pelo Annif não coincidem com o padrão-ouro. Contudo, é importante observar que nem todo termo ausente do padrão-ouro está necessariamente incorreto, podendo refletir interpretações alternativas válidas ou variações semânticas aceitáveis, as quais merecem análise qualitativa.

No conjunto analisado, foi possível identificar e categorizar dois padrões principais de erros que comprometem a precisão semântica da indexação. O primeiro diz respeito à geração de **termos genéricos ou ambíguos**, de baixa especificidade semântica. Entre os exemplos mais recorrentes estão ‘pesquisa’ (D33, D41, D52) e ‘paradigmas’ (D43, D58, D60), ambos de natureza abstrata e que, sem uma delimitação semântica mais precisa, adiciona pouco valor discriminatório. Esses termos, embora semanticamente amplos, não contribuem para a representação específica do conteúdo documental, configurando ruído lexical.

Outro exemplo emblemático é o termo ‘levantamentos’, sugerido em D48 e D55. Apesar de gramaticalmente adequado, trata-se de um termo genérico, que pode referir-se a diferentes procedimentos de pesquisa (‘levantamento bibliográfico’, ‘levantamento de dados’...), sem indicar o recorte temático do documento. Além de não constar no padrão-ouro, apresenta baixa discriminação semântica e fraca contribuição informacional, sendo classificado como falso positivo.

O segundo tipo de erro envolve **termos deslocados ou pouco relevantes** para o conteúdo do documento. Em D45, por exemplo, o termo ‘história’ foi sugerido de forma genérica, para corresponder ao aspecto histórico dos conceitos específicos do texto, como ‘webmetria’ e ‘cientometria’. Já em D51, o termo ‘educação’, embora semanticamente

amplo, foi considerado impreciso, pois não reflete o foco conceitual do documento, comprometendo a precisão da indexação.

Por outro lado, alguns erros aparentes revelaram-se interpretações semanticamente válidas, que, embora não coincidam com o padrão-ouro, mantêm coerência com o conteúdo textual, podendo ser considerado um descritor com potencial na representação temática. Um caso positivo ocorreu em D36, onde o termo ‘arquivos privados’ foi sugerido automaticamente, apesar de não constar no padrão-ouro. O artigo, entretanto, discute a aplicação de métodos bibliométricos a diversos conjuntos de dados, incluindo o arquivo privado de Getúlio Vargas, o que torna o termo conceitualmente pertinente. Casos semelhantes foram observados em D33, com ‘periódicos científicos’ pelo Annif e ‘artigos de periódicos’ e ‘produtividade de periódicos’ no padrão-ouro, e D53, com ‘estudos de usuários’ pelo Annif e ‘comportamento do usuário’ pelo padrão-ouro. Logo, mesmo não constando no padrão-ouro, esses termos apresentam correspondência semântica e relevância temática.

Em D48, o sistema sugeriu ‘informação estratégica’, termo convergente com ‘informação para negócios’ do padrão-ouro. Além disso, a sugestão de ‘análise de custos’ em vez de ‘custos’ também evidencia a capacidade do Annif de gerar sinônimos ou expressões variantes que preservam o sentido contextual e ampliam as possibilidades da representação temática.

Assim, os erros observados demonstram que o Annif tende a privilegiar os termos de maior frequência estatística, o que, sem refinamento semântico adicional, pode resultar em descritores genéricos, redundantes ou contextualmente deslocados. No entanto, parte dessas divergências reflete diferenças interpretativas plausíveis entre a indexação humana e a automática, e não necessariamente falhas do modelo.

4.3 OMISSÕES

As omissões correspondem aos casos em que os termos presentes no padrão-ouro não foram recuperados pelo Annif, revelando lacunas na compreensão semântica e na cobertura vocabular do sistema. Essas ausências indicam limitações tanto do vocabulário controlado utilizado quanto do modelo de aprendizado, que nem sempre consegue reconhecer variações lexicais ou conceituais relevantes para a representação temática completa dos documentos.

No documento D31, observou-se a ausência dos descritores ‘internet’ e ‘infraestrutura de informação’, ambos constantes no padrão-ouro. A última falha de correspondência, entretanto, decorre de uma variação ortográfica, o termo aparece como ‘infra-estrutura’ no texto original, o que pode ter prejudicado o reconhecimento automático. Esse caso evidencia a importância do pré-processamento linguístico e da normalização de grafias em sistemas de indexação automática.

Em D33, descritores como ‘artigos de periódico’, ‘citações bibliográficas’, ‘produtividade científica’, ‘produtividade de autor’, ‘produtividade de periódicos’ e ‘revisões de literatura’ não foram sugeridos, o que comprometeu a representação do documento. Essas omissões indicam que o sistema teve dificuldade em captar conceitos ligados à produção científica e aos processos de comunicação do conhecimento, do eixo temático central do texto expresso pelos descritores ‘bibliotecas digitais’ e ‘bibliotecas virtuais’ que foram corretamente atribuídos.

Situação semelhante foi identificada em D45, no qual os termos ‘cienciometria’ e ‘webometria’ não foram reconhecidos pelo sistema devido a variações de grafia, resultando em falhas no alinhamento entre a indexação automática e o padrão-ouro. Essa divergência ocorreu porque, durante o treinamento do modelo, os termos foram aprendidos na forma ‘cientometria’ e ‘webmetria’, levando o sistema a não associá-los diretamente aos descritores. Essa limitação evidencia que, embora o Annif tenha identificado parcialmente o campo temático por meio de descritores correlatos, não captou a precisão conceitual e a completude da representação temática.

De modo geral, as omissões observadas reforçam a necessidade de ampliar a cobertura do vocabulário controlado e de ampliar o *corpus* de treinamento do sistema, incorporando termos especializados e variações lexicais que podem não estar entre os mais frequentes, mas tem capacidade de impactar diretamente na representação precisa do conteúdo.

5 CONCLUSÕES

Este estudo avaliou o desempenho do sistema de indexação automática Annif, configurado com o *backend* MLLM, em um *corpus* de artigos em Ciência da Informação em português. Os resultados demonstram um bom desempenho (F1=49,23%), indicando que a ferramenta é viável para ser utilizada em um cenário de indexação semiautomática, onde

atuaria como um sistema de sugestão para um indexador humano, potencialmente agilizando o fluxo de trabalho em repositórios digitais, reforçando a eficiência sem comprometer a precisão.

A análise qualitativa evidencia que o Annif tem bom desempenho em captar conceitos centrais e recorrentes, mas falha em dois pontos, a propensão em incluir termos genéricos ou deslocados, cuja capacidade de discriminação temática é limitada e que introduzem ruído à indexação, e a omissão de conceitos especializados, que compromete a cobertura semântica.

Essas constatações reforçam os achados consistentes na literatura internacional, segundo os quais o desempenho da indexação automática é sensível a fatores como o alinhamento lexical entre o texto e o vocabulário controlado, o tamanho e a representatividade do *corpus* de treinamento e a natureza do domínio documental (Golub *et al.*, 2024; Kerketta, Mukhopadhyay, 2025).

As principais limitações deste estudo residem no tamanho reduzido do *corpus* de treinamento, que restringe a capacidade de generalização do modelo, e na avaliação centrada em correspondência lexical, que pode subestimar relações semânticas mais sutis. Além disso, a escolha por um único *backend* (MLLM), ainda que metodologicamente justificada para fins comparativos, não explorou plenamente o potencial da arquitetura multimétodo do Annif.

Para pesquisas futuras, recomenda-se:

a) ampliar o *corpus* de treinamento, garantindo maior representatividade temática e lexical;
b) comparar diferentes *backends* (ex.: Omikuji, fastText) e explorar combinações via *ensembles*;
c) integrar modelos semânticos baseados em *embeddings*, capazes de capturar correspondências conceituais não literais;
d) realizar estudos empíricos em fluxos reais de indexação semiautomática, com a participação de bibliotecários, de modo a avaliar a qualidade das sugestões automáticas e o impacto prático da adoção do Annif em contextos institucionais.

A força do Annif não reside em um único algoritmo, mas em sua arquitetura flexível, capaz de combinar modelos complementares e ajustar-se a diferentes ambientes institucionais, linguísticos e documentais. Essa característica o consolida como uma

solução versátil e escalável para a indexação automática baseada em aprendizado de máquina, permitindo equilíbrio entre precisão, revocação e contextualização semântica.

Em síntese, os resultados confirmam que o Annif é uma ferramenta de indexação automática promissora para a Ciência da Informação no Brasil, especialmente em fluxos de indexação semiautomática. Contudo, seu uso pleno em diferentes domínios depende de investimentos contínuos na qualidade dos dados de treinamento, na customização de vocabulários e na exploração de sua arquitetura multimétodo. Este estudo oferece, assim, um ponto de partida empiricamente fundamentado para a consolidação de pesquisas em indexação automática em língua portuguesa envolvendo a aplicação do *software* Annif.

REFERÊNCIAS

AHMED, M. Automatic indexing for agriculture: designing a framework by deploying Agrovoc, Agris and Annif. **Journal of Information and Knowledge**, [*s.l.*], v. 60, n. 2, April 2023, p.85-95. DOI:10.17821/srels/2023/v60i2/170966. Disponível em: https://www.srels.org/index.php/sjim/article/view/170966. Acesso em: 01 nov. 2025.

AHMED, M.; MUKHOPADHYAY, M; MUKHOPADHYAY, P. Automated knowledge organization: AI/ML-based subject indexing system for libraries. **Journal of Library & Information Technology**, [*s.l.*], v. 43, n. 1, p. 45-54, January 2023. DOI:10.14429/djlit.43.01.18619. Disponível em: https://www.proquest.com/openview/706bd22421815bd8de889f92ef99d612/1?pq-origsite=gscholar&cbl=2028807. Acesso em: 19 out. 2025.

BRITO, J. C. B.; MARTINS, D. L. Framework genérico para geração automática de assuntos e indexação em repositório digital. **Perspectiva em Ciência da Informação**, Belo Horizonte, v. 28, Fluxo Contínuo, 2023a. DOI: 10.1590/1981-5344/46629. Disponível em: https://www.scielo.br/j/pci/a/cnCL89BRvwqnQyGwJ69rXGk/?lang=pt. Acesso em: 01 nov. 2025.

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. 2023b. DOI: 10.55905/rcssv12n4-029. Disponível em: https://revistacaribena.com/ojs/index.php/rccs/article/view/2995. Acesso em: 01 nov. 2025.

DERMENTZI, M.; BRYANT, M.; ROVIGO, F.; GARCÍA-GONZÁLEZ, H. Multilingual automated subject indexing: a comparative study of LLMs vs alternative approaches in the context of the EHRI project. **Hal**, [*s. l.*], 2025. Disponível em: https://hal.science/hal-05142136v1. Acesso em: 19 out. 2025.

GIL-LEIVA, I.; ALONSO-ARROYO, A. Keywords given by authors of scientific articles in database descriptors. **Journal of the American Society for Information Science and**

Technology, [*s.l.*], v. 58, n. 8, p. 1175-1187, 2007. DOI: 10.1002/asi.20595. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/asi.20595. Acesso em: 01 nov. 2025.

GOLUB, K.; SOERGEL, D.; BUCHANAN, G.; TUDHOPE, D.; HIOM, D.; LYKKE, M. A framework for evaluating automatic indexing or classification in the context of retrieval. **J. Assn. Inf. Sci. Tec.**, [*s.l.*], n. 67, p. 3-16, 2016. DOI: 10.1002/asi.23600. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.23600. Acesso em: 01 nov. 2025.

GOLUB, K.; SUOMINEN, O.; MOHAMMED, A. T.; AAGAARD, H.; OSTERMAN, O. Automated Dewey Decimal Classification of swedish library metadata using Annif software. **Journal of Documentation**, [*s.l.*], v. 80, n. 5, p. 1057-1079, 2024. DOI: 10.1108/JD-01-2022-0026. Disponível em: https://www.emerald.com/jd/article/80/5/1057/1236186/Automated-Dewey-Decimal-Classification-of-Swedish. Acesso em: 01 nov. 2025.

HAHN, J. Semi-automated methods for BIBFRAME work entity description. **Cataloging & Classification Quarterly**, [*s.l.*], v. 59, n. 8, p. 853-867, 2021. DOI: 10.1080/01639374.2021.2014011. Disponível em: https://www.tandfonline.com/doi/full/10.1080/01639374.2021.2014011. Acesso em: 01 nov. 2025.

INKINEN, J. (ed.). Backends. **NatLibFi / Annif**, 30 jun. 2025. Disponível em: https://github.com/NatLibFi/Annif/wiki/Backends. Acesso em: 13 out. 2025.

KERKETTA, S.; MUKHOPADHYAY, P. Automated class numbers prediction for books: an AI/ML based approach using Annif. *In*: INTERNATIONAL CONFERENCE ON MARCHING BEYOND THE LIBRARIES (ICMBL): Leadership, Creativity, and Innovation, 2024, Bhubaneswar, India. **Proceedings** [...]. [*s.l.*]: Atlantis Press, 2025. DOI: 10.2991/978-94-6463-712-0_12. Disponível em: https://www.atlantis-press.com/proceedings/icmbl-24/126011042. Acesso em: 01 nov. 2025.

KIVIMÄKI, V.; HAUTAKANGAS, S.; JÄRVELÄ, H.; MALTUSCH, P. Artificial intelligence enhanced course content labelling: conceptual model and testing with university teachers. **EPiC Series in Computing**, [*s.l.*], v. 105, p. 314-325, 2025. DOI: 10.29007/j4q6. Disponível em: https://easychair.org/publications/paper/vGfT. Acesso em: 01 nov. 2025.

LANCASTER, F. W. **Indexação e resumo**: teoria e prática. Brasília, DF: Briquet de lemos, 2004.

LAPPALAINEN, M.; HULKKONEN, J.; INKINEN, J.; KALLIO, A.; KOSKELA, M.; LEHTINEN, M.; SJÖBERG, M.; SUOMINEN, O.; YETUKURI, L. Automaattisen sisällönkuvailun ohjelmiston rakentaminen: case Annif. **Signum**, [*s.l.*], vol. 53, n. 4, p. 14–20, 2021. DOI: 10.25033/sig.113612. Disponível em: https://journal.fi/signum/article/view/113612. Acesso em: 01 nov. 2025.

LEHTONEN, T.; PIUKKULA, J. Automaattinen asiasanoitus radio, ja television, ohjelmatietokanta ritvassa. **Informaatiotutkimus**, [*s.l.*], v. 1, n. 39, p. 27-45, 2020. DOI: 10.23978/inf.88107. Disponível em: https://journal.fi/inf/article/view/88107. Acesso em: 01 nov. 2025.

MEDELYAN, O. **Human-competitive automatic topic indexing**. 2009. 214f. Tese (Doutorado) – The University of Waikato, Hamilton, New Zealand, 2009. Disponível em: https://www.researchgate.net/publication/40660732. Acesso em: 19 out. 2025.

MITRA, R.; MUKHOPADHYAY, P. Machine learning applications in digital humanities: designing a semi-automated subject indexing system for a low-resource domain. **Journal of Library & Information Technology**, [*s.l.*], v. 43, n. 4, p. 219-225, July 2023. DOI:10.14429/djlit.43.04.19227. Disponível em: https://www.researchgate.net/publication/380804005_Machine_Learning_Applications_in_Digital_Humanities_Designing_a_Semi-automated_Subject_Indexing_System_for_a_Low-resource_Domain. Acesso em: 19 out. 2025.

POLEY, C.; UHLMANN, S.; BUSSE, F.; JACOBS, J-H.; KÄHLER, M.; NAGELSCHMIDT, M.; SCHUMACHER, M. Automatic subject cataloguing at the German National Library. **Liber Quarterly**, [*s.l.*], v. 35, p. 1-29, 2025. DOI: 10.53377/lq.19422. Disponível em: https://liberquarterly.eu/article/view/19422. Acesso em: 01 nov. 2025.

SILVA, B. F. de M.; CORRÊA, R. F. Aplicação da folksonomia assistida na construção de corpus de referência em ciência da informação. **Em Questão**, Porto Alegre, v. 26, n. 2, p. 413-436, maio/ago. 2020. DOI: 10.19132/1808-5245262.413-436. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/90123. Acesso em: 19 out. 2025.

SILVA, B. F. de M.; CORRÊA, R. F. O processo de construção do corpus de referência em ciência da informação. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 24, n. 56, p. 01-27, set./dez., 2019. DOI: 10.5007/1518-2924.2019.e65166. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2019.e65166. Acesso em: 01 nov. 2025.

SILVA, S. R. de B.; CORRÊA, R. F.; GIL-LEIVA, I. Avaliação direta e conjunta de Sistemas de Indexação Automática por Atribuição. **Informação & Sociedade**: Est., João Pessoa, v. 30, n. 4, p. 1-27, 2020. DOI: 10.22478/ufpb.1809-4783.2020v30n4.57259. Disponível em: https://www.researchgate.net/publication/348702009. Acesso em: 19 out. 2025.

SUOMINEN, O. Annif: DIY automated subject indexing using multiple algorithms. **LIBER Quarterly**, [*s.l.*], v. 29, n. 1, 2019. DOI: 10.18352/lq.10285. Disponível em: https://liberquarterly.eu/article/view/10732. Acesso em: 01 nov. 2025.

SUOMINEN, O.; INKINEN, J.; LEHTINEN, M. Annif and Finto AI: developing and implementing automated subject indexing. **JLIS.it**, [*s.l.*], v. 13, n. 1, jan., 2022. DOI: 10.4403/jlis.it-12740. Disponível em: https://www.medra.org/servlet/MREngine?hdl=10.4403/jlis.it-12740. Acesso em: 01 nov. 2025.

SUOMINEN, O.; INKINEN, J.; LEHTINEN, M. Annif at SemEval-2025 task 5: traditional XMTC augmented by LLMs. *In*: International Workshop on Semantic Evaluation (SemEval-2025), 19., 31 July - 1 August 2025, Vienna, Austria. **Proceedings** [...]. Vienna, Austria: Association for Computational Linguistics, 2025. p. 2424-2431. DOI:

10.48550/arXiv.2504.19675. Disponível em: https://arxiv.org/abs/2504.19675. Acesso em: 01 nov. 2025.

SUOMINEN, O.; KOSKENNIEMI, I. Annif analyzer shootout: comparing text lemmatization methods for automated subject indexing. **Journal Code4lib**, [*s.l.*], n. 54, 2022. Disponível em: https://journal.code4lib.org/articles/16719. Acesso em: 19 out. 2025.

TOEPFER, M.; SEIFERT, C. Fusion architectures for automatic subject indexing under concept drift: analysis and empirical results on short texts. **International Journal on Digital Libraries**, [*s.l.*], v. 21, n. 2, p. 169-189, 2020. DOI: 10.1007/s00799-018-0240-3. Disponível em: https://link.springer.com/article/10.1007/s00799-018-0240-3. Acesso em: 01 nov. 2025.

NOTAS

CONTRIBUIÇÃO DE AUTORIA
Concepção e elaboração do manuscrito: R. C. Lapa, R. F. Correa
Coleta de dados: R. C. Lapa, R. F. Correa
Análise de dados: R. C. Lapa
Discussão dos resultados: R. C. Lapa
Revisão e aprovação: R. F. Correa

PREPRINTS
O manuscrito não é um *preprint*.

USO DE INTELIGÊNCIA ARTIFICIAL
O uso de inteligência artificial foi mencionado no resumo e na seção de métodos. Foi utilizada a ferramenta ChatGPT (versão GPT-5 Thinking mini) exclusivamente para revisão gramatical e de clareza textual, de forma pontual e supervisionada, em conformidade com as diretrizes éticas do periódico.

CONFLITO DE INTERESSES
As pessoas autoras declaram não haver interesses conflitantes.

DISPONIBILIDADE DE DADOS DE PESQUISA E OUTROS MATERIAIS
Os dados foram publicados no próprio artigo. Todo o conjunto de dados que dá suporte aos resultados deste estudo está incluído no corpo do artigo.

PUBLISHER
Universidade Federal de Santa Catarina. As ideias expressadas neste artigo são de responsabilidade das pessoas autoras, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES
Edgar Bisset Alvarez, Genilson Geraldo, Camila de Azevedo Gibbon, Amanda Santos Witt, Karyn Lehmkuhl, Daniela Capri.

HISTÓRICO
Recebido em: 01-11-2025
Aprovado em: 16-06-2026
Publicado em: 31-07-2026