

PRODUÇÃO CIENTÍFICA SOBRE ENRIQUECIMENTO SEMÂNTICO DE METADADOS UTILIZANDO A FERRAMENTA *OPEN KNOWLEDGE MAPS*

*The scientific production on semantic enrichment of metadata
using the open knowledge maps tool*

Ana Carolina Novaes de Mendonça

Mestranda em Ciência da Informação, Universidade
Federal de São Carlos (UFSCar), São Carlos, São
Paulo, Brasil.

Instituto Brasileiro de Informação em Ciência e
Tecnologia (IBICT), Brasília, Brasil.

mnovaesana@gmail.com

<https://orcid.org/0009-0004-0285-9932>

Ana Carolina Simionato Arakaki

Pós Doutorado em Ciência da Computação –

Universidade de São Paulo (USP), São Paulo, Brasil.

Doutorado em Ciência da Informação – Universidade
Estadual Paulista (UNESP), Marília, São Paulo, Brasil.

anaarakaki@ibict.br

<https://orcid.org/0000-0002-0140-9110>

RESUMO

Objetivo: o estudo aborda o enriquecimento semântico de metadados, processo que aprimora a qualidade e a utilidade dos dados bibliográficos por meio da adição de informações contextuais e relações semânticas. Esse processo contribui para a acessibilidade e a interoperabilidade dos dados, permitindo a integração entre diferentes bases e facilitando a descoberta de informações antes isoladas. O objetivo do estudo foi descrever a produção científica sobre a aplicação do enriquecimento semântico no contexto de dados bibliográficos. **Método:** a pesquisa é descritiva e qualitativa, com uso da ferramenta *Open Knowledge Maps* e consultas complementares em bases científicas. **Resultado:** foram selecionados 16 estudos que aplicam diferentes métodos, como ontologias, vocabulários controlados e *Linked Open Data*. **Conclusões:** o enriquecimento semântico de metadados ao agregar camadas de significado, possibilita buscas mais precisas, facilita a integração de diferentes fontes de informação e amplia a descoberta de conhecimento em diversos domínios.

Palavras-chave: Enriquecimento semântico. Metadados. Dados bibliográficos.

ABSTRACT

Objective: the study addresses the semantic enrichment of metadata, a process that enhances the quality and usefulness of bibliographic data by adding contextual information and semantic relationships. This process contributes to data accessibility and interoperability, enabling the integration of different databases and facilitating the discovery of previously isolated information. The objective of the study was to describe the scientific production on the application of semantic enrichment in the context of bibliographic data. **Method:** this is a descriptive and qualitative study, using the Open Knowledge Maps tool and complementary searches in recognized scientific databases. **Results:** sixteen studies were selected, applying different methods such as ontologies, controlled vocabularies, and Linked Open Data. **Conclusions:** by adding layers of meaning, semantic enrichment of metadata

enables more precise searches, facilitates the integration of diverse information sources, and broadens knowledge discovery across various domains.

Keywords: *Semantic enrichment. Metadata. Bibliographic data.*

1 INTRODUÇÃO

O enriquecimento semântico é um processo que aprimora metadados ao agregar significados mais contextualizados, seja por meio da criação de novos metadados ou da melhoria dos já existentes (Isaac *et al.*, 2015). Essa abordagem representa uma evolução significativa no tratamento dos dados, tornando-os mais compreensíveis, interoperáveis e úteis, além de possibilitar sua conexão com outras bases de dados e a disponibilização de informações antes isoladas e desconectadas. O enriquecimento semântico fornece camadas adicionais de informação aos metadados, e pode auxiliar na tomada de decisões, tendo em vista que frequentemente decisões são tomadas sob pressão ou com base em dados incompletos e desatualizados (Arakaki; Isotani, 2024; Pfeffer; Sutton, 2006).

Os metadados organizam, descrevem e facilitam o acesso aos recursos armazenados. De acordo com Zeng e Qin (2008, p. 07), eles correspondem a informações que descrevem qualquer documento ou objeto, tanto em formatos digitais quanto tradicionais. Quando bem estruturados, os metadados permitem uma descrição dos itens, fornecendo detalhes como autor, data, formato e tema, o que facilita sua recuperação e utilização.

A evolução da catalogação foi impulsionada pela adoção de sistemas especializados, que aprimoraram significativamente a organização e recuperação de registros bibliográficos e de autoridade. Os catálogos desempenham um papel central na Ciência da Informação, pois garantem a estruturação e a acessibilidade dos dados bibliográficos, que são essenciais para a construção de registros organizados e confiáveis.

Diante desse cenário, o enriquecimento semântico de metadados é um processo que aprimora a organização e a recuperação da informação em ambientes digitais. No contexto dos dados bibliográficos, essa prática possibilita a adição de significados estruturados e com maiores significados aos metadados, favorecendo sua interoperabilidade e conexão com outras fontes de conhecimento. No entanto, a

literatura ainda apresenta uma lacuna no que diz respeito à produção científica sobre essa temática, especialmente no que se refere às abordagens adotadas. Surge a seguinte questão de pesquisa: como a produção científica tem abordado o enriquecimento semântico de metadados bibliográficos?

Com base na problemática, este estudo tem como objetivo descrever a produção científica sobre o enriquecimento semântico de metadados bibliográficos. Ao fazê-lo, busca-se contribuir para a compreensão do panorama atual da pesquisa na área, fornecendo subsídios para investigações futuras e para o aprimoramento das estratégias de enriquecimento semântico aplicadas a dados bibliográficos

2 FUNDAMENTAÇÃO TEÓRICA

Os dados desempenham um papel fundamental na catalogação e na *Web*, quando são estruturados e padronizados, eles representam os atributos fundamentais de um item e podem ser amplamente acessados e disponibilizados. Santos e Sant'Ana (2013, p. 205) definem dado como

[...] uma unidade de conteúdo necessariamente relacionada a determinado contexto e composta pela tríade entidade, atributo e valor, de tal forma que, mesmo que não esteja explícito o detalhamento sobre contexto do conteúdo, ele deverá estar disponível de modo implícito no utilizador, permitindo, portanto, sua plena interpretação.

Portanto, ao identificar o contexto em que os dados estão inseridos, torna-se possível compreender o significado e realizar uma interpretação correta das informações transmitidas. Isso é especialmente importante considerando também os elementos linguísticos, uma vez que o significado pode variar dependendo do contexto que é utilizado.

Os dados se diferenciam em estruturados, semiestruturados ou não estruturados. Schöch (2013) destaca a importância de diferenciar esses tipos de dados. Os dados estruturados são armazenados em bancos de dados e seguem um modelo de dados explícito, onde pares chave/valor possuem identificadores bem definidos e relações claras. Já os dados não estruturados, como textos livres, imagens ou áudio, não apresentam uma organização formal, tornando suas relações e significados implícitos. Entre esses dois estão os dados semiestruturados, frequentemente encontrados em arquivos *eXtensible Markup Language* (XML) ou JSON. Eles

possuem algum nível de organização, mas não seguem um esquema rígido, podendo variar em estrutura dependendo do contexto e das regras aplicadas.

Os dados assumem o papel de caracterizar outros dados, sendo denominados metadados. Pomerantz (2015, p. 26, tradução nossa) afirma que a definição de dado sobre dado não é útil, e, por isso, define como "[...] uma declaração sobre um objeto potencialmente informativo". Conforme mencionado por Zen e Qin (2008), os metadados organizam e descrevem recursos, sendo assim são a base sobre a qual novas camadas de informações são adicionadas.

Para Gilliland (2016, p. 2, tradução nossa), os metadados são "informações de valor agregado utilizadas para organizar, descrever, rastrear e melhorar o acesso a objetos de informação, itens físicos e coleções relacionadas a esses objetos". Já Arakaki (2019), com base em diversos autores, classifica os metadados de acordo com suas tipologias e funções, categorizando-os em administrativos, de autenticação, preservação, proveniência, técnicos, meta-metadados, descritivos, de direitos, de acesso e uso, estruturais e *markup languages*.

Diante desse panorama, a utilização de metadados vai além da organização e recuperação da informação, desempenhando um papel estratégico na tomada de decisão. Em diversos contextos, a disponibilidade de metadados estruturados auxilia na resolução de problemas e na definição de estratégias para alcançar objetivos específicos. Isso se torna ainda mais relevante em situações em que decisões administrativas são tomadas sob pressão, muitas vezes baseadas em informações incompletas, desatualizadas ou fundamentadas apenas na experiência individual (Arakaki; Isotani, 2024; Pfeffer; Sutton, 2006).

Para fortalecer a qualidade e minimizar distorções nos metadados, o processo de enriquecimento de metadados pode ser utilizado para ampliar a riqueza dos dados pela adição de contextos e relações semânticas, que vão além das informações explícitas contidas nos metadados básicos (Zeng, 2018).

O processo pode ser utilizado em diversos contextos, como em dados bibliográficos, geográficos, biomédicos e etc., cada um com seus padrões de metadados e tipo de dados. Por exemplo, no contexto de dados bibliográficos, o enriquecimento semântico possibilita, além das informações tradicionais como título, autor e data de criação, a conexão entre diferentes obras, autores,

influências culturais, gênero e temas, criando redes de conhecimento que facilitam a descoberta de novos materiais e a construção de relações entre documentos.

Existem diversas metodologias e abordagens para o enriquecimento de dados, permitindo aprimorar sua qualidade e capacidade de descoberta. De acordo com Zeng (2018, p. 02, tradução nossa) os métodos são

[...] como a utilização de vocabulários de sistemas de organização do conhecimento (KOS) e outros recursos que adotaram *Linked Open Data* (LOD); transformar dados semi-estruturados em dados estruturados por meio de análise semântica baseada em entidades e anotação para ampliar os pontos de acesso; explorar dados não estruturados e gerar bases de conhecimento para apoiar a descoberta de conhecimento; possibilitar o uso de dados de LAM (Bibliotecas, Arquivos e Museus) em vários contextos, em silos de dados, ao mesmo tempo em que oferece interfaces de usuário intuitivas online; e tornar os conteúdos heterogêneos de diversos provedores semanticamente interoperáveis por meio de uma infraestrutura de ontologia compartilhada.

Dessa maneira, o enriquecimento semântico de metadados se apresenta como uma abordagem para garantir a qualidade e a acessibilidade das informações, promovendo a interoperabilidade entre diferentes fontes de dados e ampliando sua capacidade de descoberta. A adoção de metodologias estruturadas e ferramentas especializadas nesse processo aprimora a organização dos metadados e auxilia na tomada de decisão em diversos contextos informacionais. Assim, o contínuo desenvolvimento e aplicação dessas técnicas representam um avanço significativo para a gestão da informação, possibilitando uma recuperação precisa e contextualizada dos dados bibliográficos.

3 PROCEDIMENTOS METODOLÓGICOS

A abordagem desta pesquisa é descritiva e qualitativa, pois busca caracterizar a produção científica sobre o enriquecimento semântico de metadados bibliográficos. De acordo com Gil (2002, p. 42), a pesquisa descritiva tem como finalidade “a descrição das características de determinada população ou fenômeno ou, então, o estabelecimento de relações entre variáveis”. Assim, este estudo descreve a literatura existente sem a intenção de testar hipóteses ou propor modelos explicativos.

O levantamento bibliográfico foi realizado utilizando a ferramenta *Open Knowledge Maps*, uma plataforma gratuita e de código aberto que facilita a navegação em artigos científicos. A busca foi realizada no dia 10 de agosto de 2024, utilizando as palavras-chave '*semanticenrichment*' AND '*metadata*'. O sistema gerou um mapa do conhecimento com os 100 artigos mais relevantes, abrangendo todas as disciplinas, mas filtrado para incluir apenas artigos científicos. A seleção dos 100 artigos mais relevantes foi realizada automaticamente pela ferramenta e os resultados são organizados em grupos temáticos (clusters) que evidenciam áreas de concentração sobre o tema.

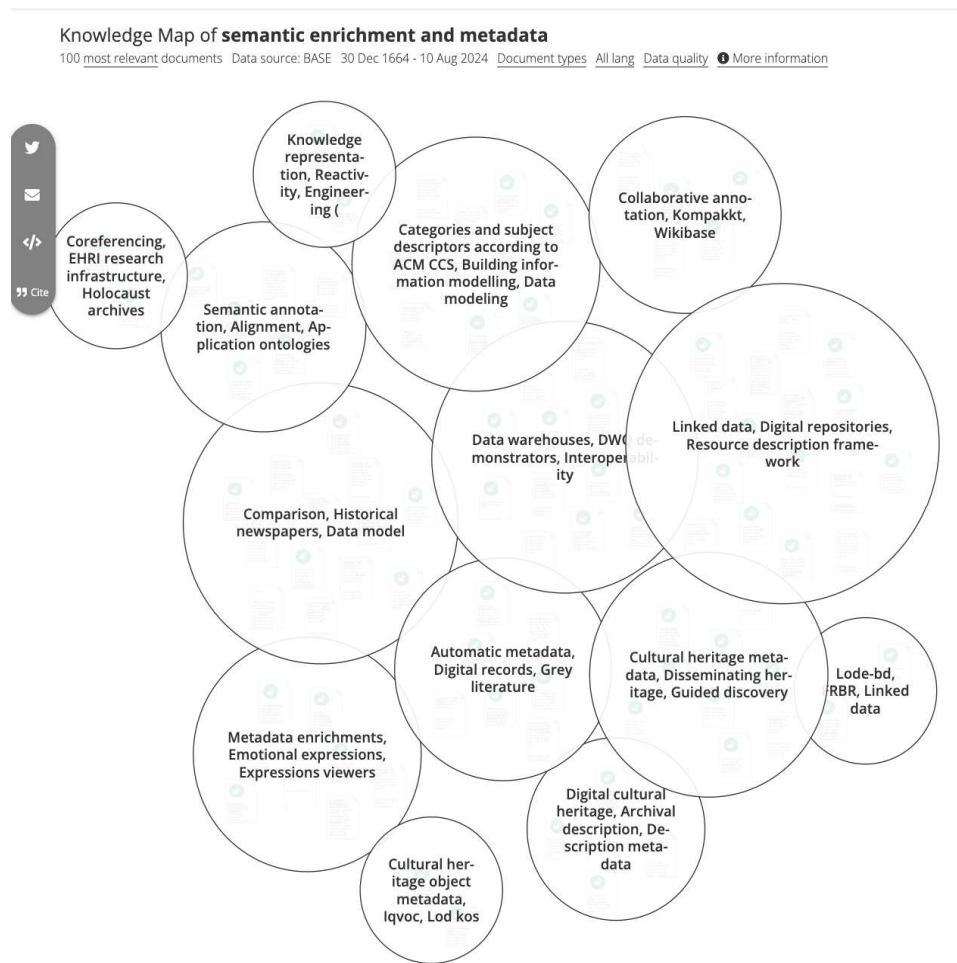
Além da busca realizada na *Open Knowledge Maps*, foram consultadas as bases: Base de Dados em Ciência da Informação (BRAPCI), *Google Scholar* e *Web of Science* para complementar a revisão da literatura. A busca foi realizada com as palavras-chave 'enriquecimento semântico', 'dados' e 'metadados', considerando apenas artigos completos, publicados entre 2013 e 2024, e indexados em bases científicas reconhecidas. Os critérios de inclusão consideraram artigos em português e inglês, que abordassem diretamente a temática do enriquecimento semântico aplicado a metadados bibliográficos.

4 RESULTADOS

A partir da utilização do *Open Knowledge Maps*, foram recuperados 100 artigos considerados mais relevantes sobre enriquecimento semântico e metadados em diferentes contextos. As publicações analisadas abrangem um período de 2000 a 2024, refletindo a evolução das abordagens e das ferramentas utilizadas ao longo dos anos. Esse levantamento evidencia a diversidade de aplicações do enriquecimento semântico, abrangendo áreas como patrimônio cultural, bibliotecas digitais, dados biomédicos e geoespaciais, entre outras.

Os artigos foram organizados em 14 grupos temáticos, agrupados de acordo com a similaridade de suas abordagens. A Figura 1 apresenta a distribuição desses grupos, permitindo uma visão mais clara das principais tendências e direções da pesquisa na área.

Figura 1 - Knowledge map of semantic enrichment



Fonte: Open Knowledge Maps (2024).

Os grupos temáticos (clusters) foram traduzidos para o português e uma breve descrição de cada um está presente no Quadro 1, abaixo.

Quadro 1 - Grupos temáticas (clusters)

Grupo temático	Descrição
Metadados automáticos, Registros digitais, Literatura cinzenta	Reúne documentos sobre a geração automática de metadados e enriquecimento semântico, com foco em bibliotecas digitais e literatura cinzenta. As abordagens incluem a utilização de ontologias, reconhecimento de metadados e migração de dados relacionais.
Metadados de objetos de patrimônio cultural, Iqvoc, Lodkos	Reúne documentos sobre o enriquecimento semântico de metadados para repositórios universitários, que incluem estratégias para melhorar o uso e o papel de repositórios de objetos de patrimônio cultural.
Enriquecimentos de	Agrupa documentos que abordam sobre ferramentas de

metadados, Expressões emocionais, Visualizadores de expressões.	enriquecimento semântico de metadados, abordando temas como a busca semântica, integração de conjuntos de dados e expressões emocionais em sistemas de metadados.
Patrimônio cultural digital, Descrição arquivística, Metadados de descrição.	Reúne documentos que abordam a melhoria da qualidade dos metadados culturais e arquivísticos por meio de crowdsourcing e tecnologias digitais. Discutem também a europeização das instituições culturais e a publicação de dados abertos relacionados ao patrimônio cultural.
Metadados de patrimônio cultural, Disseminação de patrimônio, Descoberta guiada.	Agrupa documentos que o enriquecimento de metadados para descoberta orientada e a integração de dados ligados para gestão de patrimônio cultural, linguístico e musical.
Lode-bd, FRBR, <i>Linked Data</i>	Reúne documentos que avaliam questões relativas à gestão de metadados em bibliotecas digitais, focando em conceitos-chave e perspectivas para melhorar a interoperabilidade e o gerenciamento de recursos online através de dados ligados e normas como FRBR.
<i>Linked Data</i> , repositórios digitais, Resource Description Framework (RDF)	Reúne documentos sobre linked data e metadados em repositórios digitais, explorando o enriquecimento semântico, a interoperabilidade de sistemas e a melhoria no acesso e gestão de informações.
<i>Data warehouses</i> , demonstrações DWQ, interoperabilidade	Aborda pesquisas sobre interoperabilidade de metadados, armazéns de dados e demonstrações de Data Warehouse Quality (DWQ), além das anotações semânticas, integração de serviços e dados em domínios como o biomédico e genômico, ferramentas para controle de autoridade e métodos para melhorar a interoperabilidade entre sistemas.
Comparação, jornais históricos, modelo de dados	Reúne estudos focados no enriquecimento semântico de conteúdos científicos e históricos, explorando como os metadados podem ser usados para melhorar a recuperação e análise de dados em áreas como as ciências da vida e publicações acadêmicas.
<i>Anotação colaborativa, Kompakkt, Wikibase</i>	Aborda a anotação colaborativa e enriquecimento semântico de mídias, integrando ferramentas de código aberto e grafos de conhecimento para a troca e modelagem de anotações culturais e dados 3D.
Categorias e descritores de assunto de acordo com a ACM CCS, Modelagem da informação, Modelagem de dados	Reúne pesquisas sobre o enriquecimento semântico e modelagem de dados em contextos como informações geográficas e patrimônio cultural. Os estudos abordam técnicas para aprimorar a representação e integração de dados em sistemas digitais por meio de anotações semânticas.
Representação do conhecimento, Reatividade, Engenharia	Reúne estudos sobre linguagens para expressar reatividade e modelar comportamento dinâmico na Web Semântica, focando na representação do conhecimento e na engenharia de sistemas interativos.
Anotação semântica, Alinhamento, Ontologias de aplicativos	Reúne pesquisas sobre anotação semântica, alinhamento de ontologias e seu uso em aplicativos, com foco em enriquecer dados espaciais, processar respostas factuais e modelar informações geoespaciais.

Coreferenciamento, infraestrutura de pesquisa EHRI, arquivos do Holocausto	Aborda a integração semântica de arquivos sobre o Holocausto, com foco na infraestrutura de pesquisa EHRI. Os artigos exploram técnicas de coreferenciamento e interoperabilidade para unificar o acesso a fontes históricas dispersas.
--	---

Fonte: Elaborado pelas autoras (2024).

Os clusters reúnem questões específicas relacionadas ao enriquecimento semântico de metadados. Por exemplo, o cluster "Metadados automáticos, Registros digitais, Literatura cinzenta" foca na geração automática e migração de dados em bibliotecas digitais e literatura cinzenta, enquanto o cluster "Metadados de objetos de patrimônio cultural, Iqvoc, Lodkos" explora o enriquecimento de metadados em repositórios de patrimônio cultural usando ontologias e dados ligados. Já o cluster "Enriquecimentos de metadados, Expressões emocionais, Visualizadores de expressões" discute a integração de expressões emocionais em visualizadores e ferramentas de busca semântica.

Posteriormente, foram selecionados os documentos que abordam o enriquecimento dos dados bibliográficos. Como resultado, 16 documentos abordam essa temática, que agrupados na forma crescente de data no Quadro 2.

Quadro 2 – Artigos que abordam o enriquecimento de dados bibliográficos

Autor e ano	Título	Descrição
Shotton <i>et al.</i> (2009)	Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article.	O artigo explora como aprimorar a acessibilidade e reutilização de dados bibliográficos em artigos científicos, aplicando melhorias semânticas. São integrados DOIs ativos, marcações semânticas, figuras interativas e dados baixáveis, promovendo uma experiência de pesquisa mais rica e conectada.
Solodovnik (2011)	Metadata issues in Digital Libraries: key concepts and perspectives.	O artigo destaca a importância da normalização de metadados e do uso de vocabulários controlados para melhorar a gestão de dados bibliográficos. Também enfatiza o enriquecimento semântico, que aprimora a descrição e conexão desses dados na Web.
Schaible; Mayr (2012)	Discovering Links for Metadata Enrichment on Computer Science Papers.	Propõe uma forma de adicionar informações a bibliografias iniciais, que geralmente contêm apenas metadados básicos, como título, autores e data de publicação. Utilizando a ferramenta Silk, o estudo conecta registros de ciência da computação a fontes externas, como DBLP e ACM, facilitando a descoberta e o preenchimento de dados bibliográficos adicionais.
Vlachidis <i>et al.</i> (2013)	Automatic metadata generation in an archaeological digital library: Semantic annotation of grey literature.	O artigo aborda a geração automática de metadados ricos para relatórios de escavações arqueológicas, utilizando o CIDOC CRM como ontologia central. Metadados são extraídos e enriquecidos semanticamente, melhorando a busca e a interoperabilidade dos dados através de anotações em XML e RDF.

Leiva-Mederos <i>et al.</i> (2013)	AUTHORIS: a tool for authority control in the Semantic Web.	O artigo apresenta o AUTHORIS, uma ferramenta que melhora o controle de autoridade e o enriquecimento de dados bibliográficos usando regras de aprendizagem e Linked Data. Baseada em padrões como Dublin Core e RDA, a ferramenta facilita a desambiguação e a reutilização eficiente de registros bibliográficos.
Anibaldiet <i>al.</i> (2015)	Migrating bibliographic datasets to the Semantic Web: The AGRIS case.	O AGRIS é um catálogo global com 5 milhões de registros bibliográficos estruturados, indexados pelo tesouro AGROVOC. Após sua conversão para SKOS e publicação como Linked Open Data (LOD), o conjunto de registros foi triplicado e divulgado como LOD.
Decourselle (2016)	Towards a Pattern-based Semantic Enrichment of Bibliographic Entities.	Aborda os desafios relacionados ao processo de migração de metadados e enriquecimento semântico de dados bibliográficos.
Brisebois (2017)	A semantic metadata enrichment software ecosystem (SMESE): its prototypes for digital libraries, metadata enrichments and assisted literature reviews.	O artigo apresenta o SMESE V1, um ecossistema de enriquecimento semântico de metadados para bibliotecas digitais, que integra mineração de dados, aprendizado de máquina e uma ontologia de mapeamento para aprimorar a descrição e a usabilidade dos dados bibliográficos.
Briseboiset <i>al.</i> (2017)	A semantic metadata enrichment software ecosystem based on topic metadata enrichments.	Apresenta o design, a implementação e a avaliação do modelo e algoritmo SATD (Scalable Annotation-based Topic Detection) usando metadados da web, dados abertos vinculados e regras de concordância.
Candela <i>et al.</i> (2018)	A linked open data framework to enhance the discoverability and impact of culture heritage.	Propõe uma estrutura para enriquecer e desambiguar locais em texto usando Wikidata e GeoNames, aplicada com sucesso na Biblioteca Virtual Miguel de Cervantes, demonstrando como o enriquecimento semântico melhora a recuperação de informações.
Al-Natsheh, <i>et al.</i> (2018)	Metadata Enrichment of Multi-Disciplinary Digital Library: A Semantic-based Approach.	O artigo aborda com uma categorização científica padrão e uma amostra de artigos marcados com tópicos para encontrar artigos semanticamente relevantes e enriquecer seus metadados de acordo.
Arakaki <i>et al.</i> (2019)	Availability of metadata in linked data for digital repositories.	O objetivo do artigo é discutir possibilidades de relacionamentos de conjuntos de dados e enriquecimento semântico de metadados em repositórios digitais e apresentar um modelo de conversão de registros em RDF.
Bellotto; Bekesi (2019)	Enriching metadata for a university repository by modelling and infrastructure: a new vocabulary server for phaidra.	Este artigo ilustra um passo inicial em direção ao enriquecimento semântico do repositório Phaidra da Universidade de Viena.
Labropoulou <i>et al.</i> (2023)	The LexMeta Model for Lexical Resources: Theoretical and Implementation Issues.	O artigo apresenta o LexMeta, baseado nos modelos FRBR e META-SHARE, ele visa melhorar a interoperabilidade semântica e a troca de metadados entre catálogos. Implementado em RDF/OWL e como uma ontologia Wikibase, o LexMeta facilita o enriquecimento semântico e a interoperabilidade dos dados bibliográficos.

Massariet <i>al.</i> (2024)	OpenCitations Meta.	O Open Citations Meta é um base de dados de metadados bibliográficos abertos que usa tecnologias da Web Semântica para garantir interoperabilidade e enriquecimento semântico dos dados. Ele atribui identificadores persistentes (OMIDs), elimina dependências externas e melhora a confiabilidade e o desempenho dos dados.
Ibañez; Giménez-Tol edo(2024)	Exploring named-entity recognition techniques for academic books.	O artigo propõe uma metodologia que aplica o reconhecimento de entidades nomeadas em livros acadêmicos para enriquecer semanticamente os dados, melhorando a descoberta e consulta das informações por meio de uma estrutura ontológica na web semântica.

Fonte: Elaborado pelas autoras (2024).

Shotton *et al.* (2009) demonstram, por meio do enriquecimento semântico de um artigo da *PLoS Neglected Tropical Diseases*, como tecnologias da Web Semântica podem melhorar o acesso, a reutilização e a interoperabilidade da informação científica. A proposta resultou em uma versão enriquecida do artigo, com integração de DOIs, marcação de termos, figuras interativas e dados reutilizáveis. Apesar de limitações técnicas como a ausência de metadados estruturados, necessidade de curadoria manual, falta de integração com sistemas editoriais e restrições no acesso a dados sensíveis, foi validado o potencial da publicação semântica e a necessidade de infraestrutura, padronização e ferramentas para sua adoção em larga escala.

Já Solodovnik (2011) discute a importância da normalização de metadados bibliográficos em bibliotecas digitais. Destaca uso de vocabulários controlados, ontologias, padrões de metadados para melhorar a gestão de dados bibliográficos em ambientes digitais e na Web Semântica. Analisa os modelos LODE-BD e SWAP como aplicações voltadas à publicação e descrição de dados bibliográficos em ambientes digitais e na Web Semântica. Aponta como desafios a fragmentação de padrões, a conversão de dados legados e a falta de mecanismos unificados para publicação e integração de dados.

Já Schaible Mayr (2012) propõem o uso da ferramenta Silk para enriquecer metadados bibliográficos por meio de links com fontes externas como DBLP, ACM e *Semantic Web Conference Corpus*. São exploradas duas abordagens, adição de links aos registros ou incorporação direta de dados. O Silk se mostrou funcional, mas como a necessidade de conhecer os esquemas RDF e dificuldades com

grandes volumes de dados. Apesar disso, os autores destacam a viabilidade da abordagem e sugerem sua aplicação em outros domínios.

Vlachidis *et al.* (2013) apresentam um método para a geração automática de metadados semântico para relatórios de escavações arqueológicas em literatura cinza do *Archaeological Data Service* (OASIS), integrando o CIDOC CRM, CRM-EH e Processamento de Linguagem Natural (NLP). A metodologia compreende as etapas de pré-processamento, extração de informações e conversão dos dados para o formato RDF, assegurando a conformidade com padrões semânticos. A avaliação realizada por especialistas confirmou a precisão e a capacidade de recuperação das informações, evidenciando que o método aprimora o acesso e a integração de dados arqueológicos.

Na sequência, Leiva-Mederos *et al.* (2013) apresentam o AUTHORIS, ferramenta baseada em *Linked Data* e tecnologias da *Web Semântica* voltada à geração e conversão de registros de autoridade, com o objetivo de aprimorar o controle de autoridades e promover a interoperabilidade entre entidades bibliográficas e não-bibliográficas. Testada com registros provenientes de 32 bibliotecas e editoras, a ferramenta demonstrou desempenho satisfatório, com maior ocorrência de erros nos formatos ISAD(G) e EAD. O estudo também destaca limitações do *Virtual International Authority File* (VIAF), como acesso restrito, baixa flexibilidade e exclusão de organizações não bibliotecárias, evidenciando a necessidade de modelos abertos, inclusivos e colaborativos para o controle de autoridades na *Web*.

No mesmo contexto, Anibaldiet *al.* (2015) relatam a conversão do catálogo *International Information System for the Agricultural Sciences and Technology* (AGRIS), que contém 5 milhões de registros bibliográficos, para o *Linked Open Data* (LOD), por meio da modelagem dos registros em RDF. A iniciativa envolveu a criação de URIs persistentes, enriquecimento semântico e o desenvolvimento do *Open AGRIS*, aplicação que integra dados externos via SPARQL e *Web services*. Os principais desafios incluíram a ausência de metadados administrativos para tratar de proveniência e confiança, além de limitações na interoperabilidade semântica e na disponibilidade de serviços RDF no domínio agrícola.

Decourselle (2016) propõe um modelo orientado a padrões para migrar e enriquecer registros bibliográficos com base no modelo FRBR. A abordagem facilita

a transformação de dados *Machine Readable Cataloging* (MARC) em entidades FRBR por meio da identificação de padrões bibliográficos recorrentes. Implementada na ferramenta CoM3T, permite automatizar parte do processo, incluindo análise e extração. Os principais desafios envolvem a complexidade do MARC, a criação de regras específicas, a escassez de métricas de avaliação e as limitações na fusão e enriquecimento com dados externos.

Brisebois (2017) propõe o *Semantic Metadata Enrichment Software Ecosystem* (SMESE V1), para bibliotecas digitais, baseado em uma arquitetura de *software* orientada a metadados e no modelo de linhas de produto de *software*. A proposta integra componentes de mineração de textos e dados, aprendizado de máquina e ontologias para apoiar o mapeamento e enriquecimento semântico de documentos oriundos da *Web*.

Baseado no estudo anterior, Brisebois *et al.* (2017), propõem um ecossistema de *software* para enriquecimento semântico de metadados com base na detecção de tópicos, denominado SATD. O modelo combina análise textual, grafos de coocorrência e relações semânticas extraídas de ontologias e *thesauri*, permitindo identificar tópicos por meio de similaridade com documentos anotados. Os principais desafios identificados referem-se à limitação da extração baseada apenas em palavras-chave, à ausência de coocorrências latentes, à falta de incorporação de conhecimento semântico estruturado e à dificuldade de equilibrar precisão com desempenho computacional.

Já Candela *et al.* (2018) propõem uma estrutura para enriquecer a identificação de locais nos registros bibliográficos, utilizando recursos como *Wikidata* e *Geo Names*, com aplicação na Biblioteca Virtual Miguel de Cervantes. Tiveram desafios, como a heterogeneidade dos dados, que exigiu ajustes para garantir a compatibilidade entre diferentes formatos, as variações linguísticas, que dificultaram a correspondência automática entre termos, e as limitações na disponibilidade e precisão dos dados externos, que impactam a qualidade do enriquecimento.

Al-Natsheh *et al.* (2018) aplicam uma categorização científica para enriquecer metadados em uma biblioteca digital multidisciplinar. A abordagem combina classificação supervisionada com vetores semânticos extraídos de títulos e resumos e expansão de consulta baseada em sinônimos obtidos da rede *Babel Net*. A fusão dos resultados permite atribuir múltiplos rótulos temáticos por artigo, mesmo em

contextos com vocabulários variados entre disciplinas. Avaliada com dados da biblioteca ISTEEX, a proposta supera métodos tradicionais como a LDA supervisionada. Os principais desafios envolvem a heterogeneidade terminológica, a limitação das redes semânticas para jargões técnicos e a incompletude dos dados de teste para avaliação.

Arakaki *et al.* (2019) propõem um modelo de conversão de dados de repositórios digitais para *Linked Data*, com foco na melhoria da interoperabilidade e no enriquecimento semântico de metadados. A abordagem identifica bases de dados, vocabulários e padrões compatíveis com os princípios da *Web Semântica*, estruturando o processo em etapas de extração, transformação para RDF e enriquecimento.

Bellotto e Bekesi (2019) descrevem o enriquecimento semântico do repositório *Phaidra* da Universidade de Viena, por meio da adoção de um novo vocabulário controlado em RDF alinhado aos princípios de LOD. A integração com vocabulários externos e a implementação de mecanismos de consulta estruturada ampliaram a interoperabilidade com outras plataformas. Para garantir a consistência e evolução do vocabulário, os autores destacam a necessidade de diretrizes de manutenção, incluindo regras para edição, inclusão de termos e documentação de apoio.

Labropoulou *et al.* (2023) apresentam o modelo Lex Meta para descrição de recursos lexicais em catálogos linguísticos e bibliográficos, com base nos modelos FRBR e META-SHARE. A proposta estrutura os metadados em três classes principais, permitindo representar diferentes versões e formatos de um mesmo recurso. O modelo incorpora vocabulários controlados, como o Lex Voc, para classificar aspectos como tipo de entrada e microestrutura do conteúdo. A implementação ocorre em dois formatos, RDF/OWL e *Wikibase*, com fluxos automatizados para sincronização. Os principais desafios envolvem a interoperabilidade entre modelos heterogêneos, o equilíbrio entre flexibilidade e padronização dos valores, a necessidade de extensibilidade e a adoção por comunidades com práticas descritivas distintas.

Massariet *al.* (2024), que apresentam o *Open Citations Meta*, uma base de dados de metadados bibliográficos de publicações acadêmicas indexadas pela *Open Citations*. A proposta adota identificadores persistentes próprios para desambiguar

publicações com múltiplos identificadores externos e incluir documentos sem DOI ou PMID, ampliando a cobertura das citações. O processo envolve curadoria automatizada, enriquecimento, correção de erros e conversão para RDF e uso de ontologias SPAR. Os principais desafios referem-se à ausência de identificadores em determinadas áreas do conhecimento, à ambiguidade entre identificadores múltiplos, à baixa qualidade dos dados de origem, à desambiguação de autores sem ORCID, à interoperabilidade entre formatos e ao escalonamento das atividades de curadoria.

Por fim, Ibañez e Giménez-Toledo (2024) exploram o uso do reconhecimento de entidades nomeadas (NER) para aprimorar a consulta e descoberta de informações em livros acadêmicos em espanhol, com o objetivo de ampliar a recuperabilidade e a visibilidade de conteúdos nas áreas de Humanidades e Ciências Sociais. A metodologia envolve conversão de PDF para XML, extração e indexação de parágrafos, mapeamento das entidades reconhecidas em uma ontologia desenvolvida para representar livros, autores, seções, parágrafos e entidades nomeadas. As entidades são publicadas como triplas RDF e disponibilizadas em *endpoint* SPARQL. Os principais desafios referem-se à indisponibilidade de textos completos, à necessidade de modelos linguísticos robustos por idioma, à ambiguidade entre entidades, à vinculação com grafos de conhecimento, à complexidade técnica dos mapeamentos e à adoção limitada de ontologias no setor editorial.

Conforme apontado por Zeng (2018), o enriquecimento semântico pode ser realizado por meio de diferentes métodos e estratégias, refletindo a diversidade de abordagens aplicadas para agregar significado e contexto aos dados bibliográficos. Os 16 estudos analisados evidenciam o impacto com aplicação de distintas técnicas, modelos e ferramentas.

O *Linked Data* é uma das abordagens mais frequentes, permitindo a vinculação de registros a fontes externas. Anibaldi *et al.* (2015) aplicam ao catálogo AGRIS, enquanto Schaible *et al.* (2012) utilizam a ferramenta Silk para enriquecer dados de artigos científicos. Arakaki *et al.* (2019) também propõem a conversão de metadados de repositórios institucionais com foco em interoperabilidade. Em complemento, Massari *et al.* (2023) desenvolvem uma base de dados semântica que

centraliza metadados de publicações científicas e mantém rastreamento de alterações com controle de versão em RDF.

O uso de ontologias e vocabulários controlados se destaca em iniciativas como a de Leiva-Mederos *et al.* (2013), que propõem o AUTHORIS para controle de autoridade, e Bellotto e Bekesi (2019), que implementam o vocabulário iQvoc no repositório Phaidra para melhorar a organização semântica dos registros. Labropoulou *et al.* (2023) apresentam o LexMeta como modelo orientado à descrição de recursos lexicais, visando interoperabilidade entre catálogos bibliográficos e linguísticos.

Abordagens baseadas em anotação semântica e reconhecimento de entidades nomeadas também são exploradas. Ibañez *et al.* (2024) aplicam essas técnicas a livros acadêmicos, utilizando modelos de linguagem treinados para extrair entidades e estruturar as informações em uma ontologia. Estratégias orientadas à classificação e rotulação de tópicos, como a proposta por Al-Natsheh *et al.* (2018), e à detecção de tópicos emergentes por meio de redes semânticas, como no modelo de Brisebois *et al.* (2017), mostram a capacidade do enriquecimento semântico de ampliar as possibilidades de descoberta e reuso de conteúdos científicos.

Apontam também desafios no processo de enriquecimento semântico de metadados, como a heterogeneidade dos formatos de origem, a necessidade de alinhamento entre modelos e vocabulários distintos, e a limitação de acesso ao texto completo dos documentos. Também se destacam questões relacionadas à desambiguação de entidades, à curadoria de dados em larga escala, à interoperabilidade entre sistemas e à complexidade técnica envolvida na modelagem e publicação dos dados em formato estruturado.

Como observado por Arakaki e Isotani (2024), a tomada de decisões muitas vezes ocorre com base em informações incompletas ou na experiência individual, contexto no qual o enriquecimento semântico de metadados pode desempenhar um papel fundamental, fornecendo um suporte mais estruturado e preciso para a gestão da informação.

5 CONCLUSÃO

O enriquecimento semântico de metadados representa um avanço para a melhoria da qualidade descritiva e da interoperabilidade entre sistemas. Mais do que

a simples adição de informações, trata-se da criação de camadas semânticas que tornam os dados mais compreensíveis, conectáveis e reutilizáveis. Conforme evidenciado nas publicações analisadas, há uma diversidade crescente de abordagens, com aplicações que vão desde o patrimônio cultural e dados geoespaciais até domínios biomédicos, demonstrando o caráter interdisciplinar dessa prática.

No contexto bibliográfico, o enriquecimento semântico contribui para aprimorar a recuperação da informação ao permitir buscas mais precisas e sensíveis ao contexto. Registros enriquecidos fornecem dados mais completos, favorecendo análises mais profundas e fundamentadas. Essa qualificação dos metadados beneficia a tomada de decisão em ambientes científicos, administrativos e culturais.

Além disso, a adoção de ferramentas baseadas em ontologias e *Linked Data* promove a integração de fontes heterogêneas e melhora a gestão da informação, especialmente diante do volume crescente de dados. O enriquecimento semântico, nesse sentido, se consolida como um processo relevante para tornar a informação mais acessível, interoperável e significativa.

Questões éticas envolvidas no processo de enriquecimento semântico, embora relevantes, não serão abordadas neste momento.

REFERÊNCIAS

AL-NATSHEH, H. T. *et al.* Metadata enrichment of multi-disciplinary digital library: a semantic-based approach. In: INTERNATIONAL CONFERENCE ON THEORY AND PRACTICE OF DIGITAL LIBRARIES, 2018, Cham. **Proceedings...** Cham: Springer International Publishing, 2018. p. 32–43. Disponível em: https://doi.org/10.1007/978-3-030-00066-0_4. Acesso em: 14 abr. 2025.

ANIBALDI, S. *et al.* Migrating bibliographic datasets to the Semantic Web: The AGRIS case. **Semantic Web**, Amstredã, v. 6, n. 2, p. 113–120, 2015. Disponível em: <https://doi.org/10.3233/SW-140148>. Acesso em: 14 abr. 2025.

ARAKAKI, A. C. S.; ISOTANI, S. Metadados e a tomada de decisão com base em evidências. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA – WIDAT, 6., 2023, Brasília. **Anais...** Brasília, v. 6, 2023. DOI: <https://doi.org/10.22477/vi.widat.52>. Disponível em: <https://labcotec.ibict.br/widat/index.php/widat2023/article/view/52>. Acesso em: 10 ago. 2024.

ARAKAKI, F. A. *et al.* Availability of metadata in linked data for digital repositories. **Anales de Documentación**, Murcia, v. 22, n. 2, 2019. Disponível em: <https://doi.org/10.6018/analesdoc.327441>. Acesso em: 14 abr. 2025.

BELLOTTO, A.; BEKESI, J. Enriching metadata for a university repository by modelling and infrastructure: a new vocabulary server for Phaidra. **Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare**, Viena, v. 72, n. 2, p. 446–459, 2019. Disponível em: <https://journals.univie.ac.at/index.php/voebm/article/view/5403>. Acesso em: 14 abr. 2025.

BRISEBOIS, R. **A Semantic Metadata Enrichment Software Ecosystem (SMESE)**: its prototypes for Digital Libraries, metadata enrichments and assisted literature reviews. 2017. Tese (Doutorado) – École de technologie supérieure, Montreal, 2017. Disponível em: <https://espace.etsmtl.ca/id/eprint/1967/>. Acesso em: 14 abr. 2025.

BRISEBOIS, R.; ABRAN, A.; NADEMBEGA, A.; N'TECHOBO, P. A semantic metadata enrichment software ecosystem based on topic metadata enrichment. **International Journal of Data Mining & Knowledge Management Process**, Chennai, v. 7, n. 3, p. 1–23, 2017. DOI: <https://doi.org/10.5121/ijdkp.2017.7301>. Acesso em: 14 abr. 2025.

CANDELA, G. *et al.* A linked open data framework to enhance the discoverability and impact of cultural heritage. **Journal of Information Science**, Londres, v. 45, n. 6, p. 756–766, 2019. DOI: <https://doi.org/10.1177/0165551518812658>. Acesso em: 14 abr. 2025.

DECOURSELLE, J. Towards a pattern-based semantic enrichment of bibliographic entities. **IEEE TCDL Bulletin**, Piscataway, v. 12, n. 2, 2016. Disponível em: <http://www.ieee-tcdl.org/Bulletin/v12n2/papers/decourselle.pdf>. Acesso em: 14 abr. 2025.

IBAÑEZ, P. C.; GIMÉNEZ-TOLEDO, E. Exploring named-entity recognition techniques for academic books. **Learned Publishing**, Oxford, v. 37, n. 3, p. e1610, 2024. DOI: <https://doi.org/10.1002/leap.1610>. Acesso em: 14 abr. 2025.

ISAAC, A.; MANGUINHAS, H.; STILLER, J.; CHARLES, V. **Report on enrichment and evaluation**. Haia: Europeana Task Force on Enrichment and Evaluation, 2015. Disponível em: http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/FinalReport_EnrichmentEvaluation_102015.pdf. Acesso em: 14 abr. 2025.

LABROPOULOU, P. *et al.* The Lexmeta Model for Lexical Resources: Theoretical and Implementation Issues. **Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje**, Zagreb, v. 49, n. 1, p. 137–154, 2023. Disponível em: <https://hrcak.srce.hr/292655>. Acesso em: 14 abr. 2025.

LEIVA-MEDEROS, A. *et al.* Authoris: a tool for authority control in the semantic web. **Library Hi Tech**, Bradford, v. 31, n. 3, p. 536–553, 2013. DOI: <https://doi.org/10.1108/LHT-12-2012-0135>. Acesso em: 14 abr. 2025.

MASSARI, A. *et al.* OpenCitations Meta. **Quantitative Science Studies**, Cambridge, v. 5, n. 1, p. 50–75, 2024. DOI: https://doi.org/10.1162/qss_a_00292. Acesso em: 14 abr. 2025.

OPEN KNOWLEDGE FOUNDATION. **Open definition**. Cambridge, [s. d.]. Disponível em: <https://opendefinition.org>. Acesso em: 10 ago. 2024.

OPEN KNOWLEDGE MAPS. **Knowledge map for research on semantic enrichment and metadata**. Viena, 2024. Disponível em: <https://openknowledgemaps.org/map/3e05c9f29d39a3d6d1f55bd1fd610acf?area=1>. Acesso em: 10 ago. 2024.

POMERANTZ, J. **Metadata**. Cambridge: The MIT Press, 2015. 239 p.

SANTOS, P. L. V. A. C.; SANT'ANA, R. C. G. Dado e granularidade na perspectiva da informação e tecnologia: uma interpretação pela ciência da informação. **Ciência da Informação**, Brasília, v. 42, n. 2, 2013. Disponível em: <https://brapci.inf.br/index.php/res/v/17173>. Acesso em: 10 ago. 2024.

SCHAIBLE, J.; MAYR, P. Discovering links for metadata enrichment on computer science papers. **arXiv**, Ithaca, NY, 2012. (Preprint). Disponível em: <https://arxiv.org/abs/1212.3677>. Acesso em: 10 ago. 2024.

SHOTTON, D. *et al.* Adventures in semantic publishing: exemplar semantic enhancements of a research article. **PLoS Computational Biology**, São Francisco, v. 5, n. 4, p. e1000361, 2009. DOI: <https://doi.org/10.1371/journal.pcbi.1000361>. Acesso em: 14 abr. 2025.

SOLODOVNIK, I. Metadata issues in digital libraries: key concepts and perspectives. **JLIS.it: Italian Journal of Library, Archives and Information Science**, Florença, v. 2, n. 2, p. 1–27, 2011. DOI: <https://doi.org/10.4403/jlis.it-4663>. Acesso em: 14 abr. 2025.

VLACHIDIS, A. *et al.* Automatic metadata generation in an archaeological digital library: semantic annotation of grey literature. In: PRZEPIÓRKOWSKI, A.; PIASECKI, M.; JASSEM, K.; FUGLEWICZ, P. (ed.). **Computational Linguistics**. Heidelberg: Springer, 2013. p. 187–202. (Studies in Computational Intelligence, v. 458). DOI: https://doi.org/10.1007/978-3-642-34399-5_10. Acesso em: 14 abr. 2025

ZENG, M. L. Semantic enrichment for enhancing LAM data and supporting digital humanities. **Profesional de la Información**, [S. l.], v. 28, n. 1, 2019. DOI: <https://doi.org/10.3145/epi.2019.ene.03>. Acesso em: 10 ago. 2024.

ZENG, M. L.; QIN, J. **Metadata**. Nova Iorque: Neal-Schuman Publishers, 2008.

NOTAS E CRÉDITOS DO ARTIGO

- **Reconhecimentos:** os autores gostariam de agradecer ao Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
- **Financiamento:** este estudo foi financiado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)
- **Conflitos de interesse:** Não se aplica.
- **Aprovação ética:** Não se aplica.
- **Disponibilidade de dados e material:** Não se aplica.
- **Manuscrito publicado como *preprint*:** o manuscrito foi originalmente publicado como trabalho completo no Seminário Nacional de Catalogação e Tecnologia (SNCat), em 2024. Posteriormente, passou por nova avaliação *double-blindpeer review*, além de receber ajustes e atualizações de conteúdo.
- **Contribuições dos autores:**

Contribuição	MENDONÇA, A. C. N.	ARAKAKI, A. C. S
Concepção do estudo	X	
Conceitualização	X	X
Metodologia	X	X
Coleta de dados / investigação	X	
Curadoria de dados	X	
Análise dos dados	X	
Discussão dos resultados	X	
Visualização (gráficos, tabelas e outros)	X	
Rascunho original	X	
Revisão e edição final	X	X
Supervisão e administração	X	X
Aquisição de financiamento	X	X

- **Licença de uso**

Os autores cedem ao **Ciência da Informação Express – CIExpress** direitos exclusivos de primeira publicação, com o trabalho simultaneamente licenciado sob a Licença *Creative Commons Attribution (CC BY) 4.0 International*. Esta licença permite que terceiros remixem, adaptem e criem a partir do trabalho publicado, atribuindo o devido crédito de autoria e publicação inicial neste periódico.

- **Publicador**

Universidade Federal de Lavras (UFLA).

As ideias expressas neste artigo são de responsabilidade de sua autoria, não representando, necessariamente, a opinião dos editores ou da universidade.

- **Editor do canal de comunicação e divulgação científica Ciência da Informação Express - CIExpress**

Nivaldo Calixto Ribeiro, Universidade Federal de Lavras (UFLA).

- **Revisor de língua portuguesa**

Revisado pelos autores.

- **Histórico**

Recebido em: 14/11/2024

Aceito em: 14/04/2025

Publicado em: 15/04/2025