UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

FELIPE IVO DA SILVA

PROVENIÊNCIA DE DADOS E METADADOS EM REPOSITÓRIOS DE DADOS DE PESQUISA

São Carlos
2025

FELIPE IVO DA SILVA

PROVENIÊNCIA DE DADOS E METADADOS EM REPOSITÓRIOS DE DADOS DE PESQUISA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de São Carlos (UFSCar), como parte das exigências para obtenção do título de Mestre em Ciência da Informação.

Área: Conhecimento, Tecnologia e Inovação.
Linha: Tecnologia, Informação e Representação.

Orientador: Prof. Dr. Felipe Augusto Arakaki
Coorientadora: Profa. Dra. Ana Alice Rodrigues Pereira Baptista

São Carlos
2025

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Ciência da Informação

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Felipe Ivo da Silva, realizada em 03/04/2025.

Comissão Julgadora:

Prof. Dr. Felipe Augusto Arakaki (UnB)

Prof. Dr. Rogério Aparecido Sá Ramalho (UFSCar)

Prof. Dr. Caio Saraiva Coneglian (Unimar)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Informação.

A Deus,
Á minha amada esposa, Sara,
Á D. Marinez Ivo – minha avó – pelos valores ensinados (*in memoriam*),
Ao Laércio Sertori – meu sogro – pela generosidade e honestidade ensinados (*in memoriam*).

AGRADECIMENTOS

Agradeço a Deus por todas as bênçãos imerecidas que surgiram em meu caminho durante todo esse estudo.

Agradeço à minha esposa, Sara, companheira de vida, por me dar força e encorajamento durante os momentos de dificuldades.

Estendo meus agradecimentos aos meus pais, Roberto e Karina, que sempre me guiaram e apoiaram em minha jornada em direção ao conhecimento e à ciência.

Ao meu orientador, Felipe, por sua paciência inabalável durante minha jornada acadêmica, me incentivando e guiando na trajetória cientifica.

À minha coorientadora, Ana Alice, por abrir novos horizontes e contribuir significativamente para o bom desenvolvimento deste trabalho.

À minha banca, composta pelos professores Dr. Rogério Aparecido Sá Ramalho e Dr. Caio Saraiva Coneglian, gostaria de expressar minha gratidão pelas contribuições para o desenvolvimento do meu trabalho.

Aos meus estimados professores do Programa de Pós-Graduação em Ciência da Informação da UFSCar, que foram minha inspiração.

Agradeço, também, ao Instituto Brasileiro de Informação em Ciência e Tecnologia, na pessoa da professora Ana Carolina Simionato Arakaki.

RESUMO

Os repositórios de dados de pesquisa são ambientes de armazenamento, preservação e compartilhamento de dados científicos, e garantem acessibilidade e reutilização dos dados. A proveniência, que documenta a origem, histórico e transformações dos dados, é fundamental para assegurar a autenticidade, confiabilidade e rastreabilidade das informações. No entanto, a falta de padronização e interoperabilidade entre os principais padrões de metadados utilizados nesses repositórios pode comprometer a eficácia da proveniência. Nesse contexto, a questão central desta pesquisa foi: os padrões de metadados utilizados para a descrição de dados em repositórios de pesquisa asseguram a rastreabilidade e autenticidade das informações ao longo do tempo? O objetivo geral do estudo foi avaliar a aderência dos metadados de proveniência, com base na Família PROV, aos principais padrões de metadados utilizados em repositórios de dados de pesquisa segundo o Re3data, como *Dublin Core*, *DataCite* e DDI. A pesquisa buscou identificar como esses padrões podem auxiliar na identificação da autenticidade e rastreabilidade das informações, propondo recomendações para sua aplicação. A metodologia adotada foi teórica e exploratória, utilizando levantamento bibliográfico em bases de dados reconhecidas, análise de literatura especializada e o método *Crosswalk* para mapear e comparar os padrões de metadados em relação à PROV-O. A análise focou na interoperabilidade entre os padrões e a Família PROV, visando identificar possíveis adaptações para melhorar a interoperabilidade semântica. Os resultados principais indicaram que os padrões analisados possuem diferentes graus de interoperabilidade com a Família PROV. O *Dublin Core* apresentou a maior correspondência relativa, enquanto *DataCite* e DDI mostraram níveis mais baixos de interoperabilidade. Apesar da viabilidade de integração, foram identificadas limitações que exigem adaptações para atender aos princípios de interoperabilidade semântica promovidos pelo W3C. Em conclusão, o estudo reforça a importância da proveniência para a confiabilidade e rastreabilidade dos dados em repositórios de dados pesquisa, destacando a necessidade de colaboração interdisciplinar entre Ciência da Informação e Ciência da Computação para aprimorar a gestão de dados em ambientes digitais. A pesquisa sugere a implementação de metadados de proveniência para promover a veracidade e longevidade dos dados.

Palavras-chave: Metadados. Dados. Proveniência. Repositórios de dados de pesquisa.

ABSTRACT

Research data repositories are environments for storing, preserving, and sharing scientific data, and ensure accessibility and reuse of data. Provenance, which documents the origin, history, and transformations of data, is essential to ensure the tradition, reliability, and traceability of information. However, the lack of standardization and interoperability between the main metadata standards in these repositories can compromise the effectiveness of provenance. In this context, the central question of this research was: do the metadata standards used to describe data in research repositories ensure traceability and prevent information from becoming obsolete over time? The general objective of the study was to evaluate the adherence of provenance metadata, based on the PROV Family, to the main metadata standards used in research data repositories according to Re3data, such as Dublin Core, DataCite, and DDI. The research sought to identify how these standards can assist in identifying deficiency and traceability of information, proposing recommendations for their application. The methodology adopted was theoretical and exploratory, using a bibliographic survey in recognized databases, analysis of specialized literature and the Crosswalk method to map and compare metadata standards in relation to PROV-O. The analysis focused on the interoperability between the standards and the PROV Family, eventually identifying possible adaptations to improve semantic interoperability. The main results indicated that the developed standards have different degrees of interoperability with the PROV Family. Dublin Core presented the highest relative correspondence, while DataCite and DDI presented lower levels of interoperability. Despite the integration suggestions, limitations were specified that exclude adaptations to meet the semantic interoperability principles promoted by the W3C. In conclusion, the study reinforces the importance of provenance for the reliability and traceability of data in research data repositories, highlighting the need for interdisciplinary collaboration between Information Science and Computer Science to improve data management in digital environments. The research suggests the implementation of provenance metadata to promote data veracity and longevity.

Keywords: *Metadata. Data. Provenance. Research data repositories.*

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

CFB	Conselho Federal de Biblioteconomia
DC	Terms *Dublin Core Metadata Initiative Metadata*
Terms	
DCMI	*Dublin Core Metadata Initiative*
DOI	*Digital object identifier*
PROV- CONSTRAINTS	Constraints of the PROV Data Model
PROV-DM	PROV *Data Model*
PROV-N	*Provenance Notation*
PROV-O	PROV *Ontology*
RDF	*Resource Description Framework*
Re3data	*Registry of Research Data Repositories*
W3C	*World Wide Web Consortium*
WoS	*Web of Science*
XML	*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

As informações desempenham um papel importante para o desenvolvimento global, e com o crescimento da tecnologia, o volume de informações digitais cresceu concomitantemente. Diante deste cenário, o estudo da proveniência torna-se primordial para descrição e confiabilidade das informações, já que estas permitem a comprovação e a autenticidade dos recursos informacionais, em especial, no contexto digital.

A proveniência é descrita por Moreau e Groth (2013, p. 4) “[...] como um registro que engloba informações sobre pessoas, instituições, entidades e atividades relacionadas à produção, influência ou entrega de um determinado dado ou objeto. Arakaki e Santos (2021, p. 3) complementam que o termo proveniência é utilizado para identificar o responsável pela criação, guarda e gerenciamento de informações e recursos em diversas áreas.

Coneglian e Segundo (2018) pontuam que computadores e humanos possuem formas distintas de se comunicar e expressar. Embora as máquinas sejam altamente eficientes em acessar e processar dados, gerando informações úteis para os usuários, é indispensável dispor de métodos que facilitem a interação entre ambos.

Neste processo de identificação, os metadados apresentam-se como uma solução para que o usuário possa recuperar com melhor precisão as informações sobre um determinado recurso informacional. Ao modo que, auxilia na garantia da qualidade e veracidade dos dados.

De acordo com Joudrey, Taylor e Wisser (2018), os metadados não se limitam apenas às informações descritivas tradicionais usadas para facilitar a descoberta de recursos, mas também incluem dados que são essenciais para a gestão, o uso e a preservação desses recursos. Além disso, eles podem abranger informações sobre o contexto, a qualidade e as condições dos dados, como sua localização, forma de exibição online, propriedade e estado de conservação.

Neste sentido, os metadados são uma parte do processo de descrição e podem ser guiados a partir dos princípios da Catalogação. A definição e padronização dos metadados dos recursos informacionais institui a criação de instrumentos que garantem a identificação e a preservação dos recursos informacionais, além da busca, acesso, uso e reuso da informação.

No contexto digital, alguns estudos foram realizados e apresentaram

discussões relevantes sobre a importância da proveniência da informação. Pode-se citar como trabalhos correlatos os estudos de Curty e Gama (2007), que discutem as relações entre a diplomática e a XML para garantia da proveniência de documentos jurídicos digitais. Freund, Sembay e Macedo (2019), que discutem as relações interdisciplinares entre a proveniência e a segurança da informação. Arakaki e Santos (2019), que realizaram um mapeamento dos metadados de proveniência em padrões de metadados no domínio bibliográfico. Arakaki (2020), que apresenta o modelo PROV para descrição da proveniência dos dados. Dondi, Lefferts e Delft (2022), que apresentam discussões sobre a proveniência no contexto do *Consortium of European Research Libraries*, Simmhan, Plale e Gannon (2005), que estabeleceram uma taxonomia das diversas possibilidades de informações da proveniência e os estudos de proveniência dos dados tratada pelo *World Wide Web Consortium* (W3C).

Dessa forma, os metadados apresentam-se como um meio para que o usuário possa avaliar com melhor precisão sobre a escolha de um determinado recurso informacional, salvaguardando a propriedade primária dos dados. Desse modo, os metadados são vistos como uma aplicação prática relacionada à catalogação, indexação, desenvolvimento de banco de dados e gravação de transações digitais.

Na atualidade, com o uso massivo de meios tecnológicos e com a abundância de dados e informação, a área de Ciência da Informação possui diversos desafios, sendo um destes desafios a garantia da representação da proveniência dos recursos informacionais.

Assim sendo, a Ciência da Informação deve contemplar um vasto conjunto teórico e metodológico para gestão, compartilhamento, organização, acesso, uso e reuso da informação, de modo a propiciar que o usuário encontre a informação desejada. Por este motivo, profissionais de diversas áreas têm se mobilizado para criar instrumentos para representação e tratamento da informação. No entanto, a questão da representação da proveniência de metadados e dados no contexto digital é escassa, Arakaki (2019, p. 18), diz que estudos teóricos sobre os metadados administrativos, em especial a proveniência dos dados no Brasil não é muito discutida, e essa conduta pode gerar uma falta de padronização, acerca da proveniência.

A representação da proveniência de metadados refere-se à forma como a origem, o histórico e as alterações de metadados são documentados e estruturados, enquanto os metadados de proveniência são os próprios dados que descrevem esses aspectos, ou seja, informações sobre a criação, modificação, transferência e

manutenção dos metadados.

1.1 Questão da pesquisa

Ao considerar as diversas publicações que são depositadas e disponibilizadas nos repositórios de dados de pesquisa, a proveniência faz-se necessária para garantir a autenticidade e confiabilidade das informações disponibilizadas. Dessa forma, acredita-se que os metadados de proveniência forneçam bases referenciais de apoio aos processos de gestão do ciclo de vida dos recursos informacionais bem como a garantia de sua origem. Haynes (2018, p. 134, tradução nossa) complementa que “No contexto dos materiais digitais, fornecer informações de proveniência pode ajudar a demonstrar que um registro não foi adulterado e que a evidência que ele apresenta é, portanto, confiável”.

A escolha dos padrões de metadados adotados pelos repositórios de dados de pesquisa, bem como sua interoperabilidade, interfere diretamente na garantia de informações. Baptista (2010, p. 88) aponta que seguir diretrizes é essencial para alcançar um nível de interoperabilidade que possibilite aos fornecedores de serviços lidar de maneira eficiente com os dados agregados oriundos de diferentes repositórios, e isso implica, essencialmente, que os serviços oferecidos à comunidade científica, possuam grande potencial para alcançar melhorias significativas em termos de qualidade.

Dessa forma, a adequação de metadados com a devida preocupação com descrição informacional, pode auxiliar o desenvolvedor dos registros informacionais a contribuir com a vida útil dos dados, principalmente os de proveniência. Diante deste contexto, este estudo tem como problema de pesquisa:

Os padrões de metadados utilizados para a descrição de dados em repositórios de dados de pesquisa asseguram a rastreabilidade e autenticidade das informações ao longo do tempo?

1.2 Objetivos

O objetivo geral é avaliar a aderência dos metadados de proveniência, com base na Família PROV, aos principais padrões de metadados utilizados para a descrição de informações no contexto de repositórios de dados, verificando quais

padrões de metadados podem auxiliar na identificação da autenticidade e rastreabilidade das informações em repositórios de dados de pesquisa.

Os objetivos específicos definem-se dessa forma:

A. Apresentar o contexto de metadados de proveniência na literatura;
B. Identificar e mapear os padrões de metadados mais utilizados com os metadados de proveniência estabelecidos pelo PROV-O em repositórios de dados de pesquisa;
C. Propor recomendações de aplicação de metadados de proveniência em repositórios de dados de pesquisa.

1.3 Justificativa

Esta pesquisa, cujo tema é proveniência de dados e metadados em repositórios de dados de pesquisa; e em particular, os repositórios digitais de dados, justifica-se a partir de um contexto profissional, social e tecnológico, ao qual cresce exponencialmente e em um curto período, juntamente com a facilidade de acesso a estes recursos informacionais dentro das instituições públicas e privadas, e também, da sociedade de modo geral.

Espera-se que, com este estudo, possa-se ampliar as discussões sobre a questão da proveniência da informação, e auxiliar os desenvolvedores, em especial os gestores de repositórios de pesquisa a melhorar a representação desses recursos nestes ambientes.

O aspecto inovador da pesquisa reside em sua proposta de analisar e relacionar os principais modelos de metadados de repositórios de dados, abordando a representação da proveniência, com base nas Ciências da Informação e da Computação. Do ponto de vista científico, a pesquisa contribui para o avanço da compreensão sobre a proveniência de dados e metadados em repositórios de dados.

No âmbito social e político, a pesquisa fortalece a confiança na gestão de dados científicos, promovendo a transparência e a autenticidade das informações, essenciais para decisões informadas em políticas públicas e no desenvolvimento acadêmico.

Tecnologicamente, a pesquisa contribui para a correlação de modelos e padrões de metadados e auxilia na garantia, integridade e o rastreamento de dados

em repositórios de dados, o que pode otimizar a utilização de recursos informacionais na ciência e na inovação. Do ponto de vista econômico, ao garantir a confiabilidade e a rastreabilidade dos dados, a pesquisa ajuda a evitar desperdícios e duplicações de esforços em projetos de pesquisa e desenvolvimento, promovendo uma gestão mais eficiente dos recursos.

Acredita-se que a proteção da representação da proveniência será uma contribuição significativa para os usuários e agentes computacionais, que poderão ter maiores garantias da confiabilidade da produção do recurso informacional.

Como explicitado, a proveniência desempenha um papel fundamental para todo o contexto de informação, pois é necessário identificar a origem desse objeto – a partir dos metadados – para auxiliar se o recurso informacional é autêntico, ou se não foi modificado ou alterado.

Diferencia-se este projeto de outros trabalhos pois tem como proposta a análise dos metadados de proveniência em repositório digitais. Destaca-se que, até o momento, não foram identificados estudos sobre a temática. Os repositórios de dados de pesquisa possuem uma vasta discussão, partindo principalmente do movimento de Acesso Aberto e da Comunicação Científica. Dessa forma, justifica-se o desenvolvimento desta pesquisa, ao considerar que os repositórios de dados de pesquisa oferecem informações significativas para o ambiente científico e é um objeto indispensável para o desenvolvimento das Ciências, e dessa forma, a qualidade informacional é um fator salutar na edificação e desenvolvimento de quaisquer campos acadêmicos e tecnológicos.

Assim, o uso dos metadados de proveniência nos repositórios de dados, é o objeto de estudo que busca abordar o tema no contexto digital, e que estuda a representação dos metadados de proveniência, com base teórica na Ciência da Informação e Ciência da Computação.

No contexto dos metadados, a proveniência da informação é uma das informações-chave que deve ser coletada e registrada. Isso permite que as informações sejam rastreadas de volta à sua fonte original, e que sejam tomadas decisões informadas sobre como usá-las e compartilhá-las. Além disso, a proveniência da informação pode ajudar a resolver disputas sobre a autoria ou a propriedade de informações e, consequentemente, a evitar a duplicação de esforços em pesquisas ou projetos.

Uma outra justificativa é a possibilidade de rastrear a história de um

determinado conjunto de dados. Isso é importante para entender como os recursos informacionais foram criados, gerenciados e compartilhados ao longo do tempo, e para identificar qualquer problema ou discrepância que possa ter ocorrido no processo.

Em resumo, o estudo de metadados com relação à proveniência é importante, porque ajuda a identificar responsabilidades, estabelecer regras de acesso e uso de dados, garantir a integridade dos dados e rastrear a história dos dados ao longo do tempo: isso é decisivo para garantir a qualidade e a confiabilidade dos dados, bem como apoiar a tomada de decisões baseadas em informações.

1.4 Teoria de base

A partir dos apontamentos sobre a pesquisa, destaca-se que a teoria de base possibilita a discussão e o aprofundamento teórico sobre metadados conforme abordado por Méndez Rodríguez (2002); Liu (2007); Alves (2010); Miller (2011); Alves e Santos (2013); Pomerantz (2015); Baca (2016); Zeng e Qin (2016); Riley (2017); Haynes (2018).

A questão da proveniência é pouco explorada na literatura, em especial no Brasil, e os estudos estão focados nas bases da arquivologia, no entanto, esta pesquisa não se restringirá apenas à perspectiva arquivística; ele buscou ampliar a análise, considerando principalmente a abordagem da Ciência da Informação. Dessa forma, a base teórica sobre proveniência será estabelecida em especial por autores como: Bellotto (2015), que apresenta uma fundamentação baseada na Diplomática (disciplina teórica arquivística que apresenta princípios para proveniência; a organicidade; a unicidade e a indivisibilidade/integridade), Albuquerque e Souto (2013), que fazem um panorama histórico sobre a questão da proveniência. Além de Tognoli e Guimarães (2019), que fazem uma discussão sobre o que é a proveniência.

No âmbito digital, em especial na Web, foram estabelecidos um conjunto de documentos que tratam da identificação da proveniência, chamado de família PROV. A família PROV é composta por quatro recomendações, além de oito notas que auxiliam no mapeamento e nas informações sobre o modelo PROV (WORLD WIDE WEB CONSORTIUM, 2013).

Dessa forma, a pesquisa irá traçar o seu referencial teórico a partir das discussões sobre metadados, proveniência, em especial na representação da

proveniência no contexto digital.

Para maior clareza das definições utilizados na pesquisa, podem ser analisadas no quadro a seguir:

Quadro 1 – Definição dos termos

Termo	Definição
Proveniência	Origem e histórico de um recurso de informação.
Proveniência de Dados	Documentação da origem e história dos dados e informações.
Proveniência de Recursos	Histórico da origem e modificações de recursos específicos.
Metadados de Proveniência	Metadados que documentam a história e a origem dos dados e recursos.

Fonte: Elaborado pelo autor.

Esse quadro reflete a interpretação dos teóricos citados, oferecendo um panorama dos principais conceitos relacionados à proveniência e metadados de proveniência.

1.5 Estrutura da dissertação

A dissertação é composta pela introdução, que inclui a questão da pesquisa, a justificativa, os objetivos, as teorias de bases, além das seguintes seções:

- **SEÇÃO 2: REVISÃO DE LITERATURA –** apresenta o conceito de metadados, sua importância e suas descrições. Discute, também, como os metadados facilitam a organização, busca e recuperação das informações em sistemas de informação. Além disso, nessa seção, são apresentadas as definições de metadados de proveniência e os seus impactos nos repositórios dados de pesquisa.

- **SEÇÃO 3: PROCEDIMENTOS METODOLÓGICOS –** esta seção apresenta os procedimentos metodológicos adotados na pesquisa, detalhando as etapas do desenvolvimento da pesquisa, os materiais e ferramentas utilizados, bem como o tipo de análise de dados aplicada. Além disso, serão descritas as abordagens teóricas e práticas

que fundamentam o estudo.

- **SEÇÃO 4: COLETA E ANÁLISE DE DADOS –** são apresentados os dados da pesquisa, os quais envolvem exemplos de metadados de proveniência bem como os metadados considerados de proveniência. Com o método de análise Crosswalk, foram traçadas comparações entre os metadados e seus resultados. Além disso, serão discutidas as implicações desses resultados e suas contribuições para a área de estudo.

- **SEÇÃO 5: CONSIDERAÇÕES FINAIS –** a partir dos resultados obtidos no estudo, essa seção apresenta as conclusões do estudo, fazendo uma breve contextualização e síntese dos resultados obtidos ao longo do trabalho.

2 REPOSITÓRIOS DE DADOS DE PESQUISA

Os repositórios de dados de pesquisa, também chamados de dados científicos, Gava, *et al*. (2024), têm ganhado relevância na era da informação, sendo amplamente utilizados para armazenar, preservar e disseminar dados de diversas naturezas. Eles vão além das bibliotecas digitais ao adicionar recursos suplementares, como ferramentas para preservação a longo prazo e o compartilhamento de metadados, conforme apontado por Martins, Silva, Santarém Segundo e Siqueira (2017).

Estes autores argumentam que um repositório digital, como os repositórios de dados de pesquisa, não apenas incorpora as principais características das bibliotecas digitais, mas também oferecem uma abordagem mais robusta em termos de preservação documental e integração de metadados, seguindo normas nacionais ou internacionais.

Um repositório digital pode ser definido como um sistema de informação que armazena, preserva e proporciona acesso a objetos digitais. Viana, Márdero Arellano e Shintaku (2005) destacam que esses repositórios possuem a capacidade de manter e gerenciar materiais por longos períodos, assegurando o acesso contínuo e adequado a esses materiais.

Já Braga (2010) complementa ao definir repositórios digitais como sistemas que armazenam e divulgam a produção intelectual de comunidades científicas, enfatizando o papel desses ambientes na preservação do conhecimento e na sua disseminação global.

Além disso, o Conselho Federal de Biblioteconomia (CFB), pontua que, os repositórios digitais se diferenciam de outras bibliotecas digitais principalmente pelo seu foco na preservação de longo prazo e no acesso contínuo a dados. Embora bibliotecas digitais também disponibilizem materiais para consulta e pesquisa, os repositórios têm a responsabilidade adicional de garantir que os dados estejam em conformidade com protocolos específicos de metadados e sejam acessíveis a outros pesquisadores, independentemente do contexto. [1]

Em relação aos repositórios de dados de pesquisa que exigem o manejo de um grande volume de dados no contexto científico, o que gera a demanda por repositórios

[1] Resolução CFB nº 240, de 30 de junho de 2021. Disponível em: https://www.in.gov.br/web/dou/-/resolucao-cfb-n-240-de-30-de-junho-de-2021-330702470.

digitais (plataformas de preservação) para a administração desse tipo particular de dado, ou seja, dados produzidos durante a realização de atividades de pesquisa e destinados ao reuso:

> Dados e informações digitais gerados pelas atividades de pesquisa necessitam de cuidados específicos, tornando-se necessário a criação de novos modelos de custódia e de gestão de conteúdos científicos digitais que incluam ações de arquivamento seguro, preservação, formas de acrescentar valor a esses conteúdos e de otimização da sua capacidade de reuso. No intuito de por em prática soluções para o problema, observa-se, no âmbito de várias disciplinas, um esforço em torno do desenvolvimento de repositórios digitais orientados especialmente para uma gestão ativa de dados de pesquisa (Sayão; Sales, 2012, p. 180).

Como acontece com muitos conceitos dentro de uma vasta área do conhecimento, existem diversas definições para o conceito de “dados de pesquisa” na CI, Gava, *et al*. (2024). Neste contexto a definição do termo se faz fundamental, visto que a expressão "dados de pesquisa" pode ter diferentes significados e varia conforme o contexto científico em que é aplicada.

> De fato, o termo “dado de pesquisa” tem uma amplitude de significados que vão se transformando de acordo com domínios científicos específicos, objetos de pesquisas, metodologias de geração e coleta de dados e muitas outras variáveis. Pode ser o resultado de um experimento realizado num ambiente controlado de laboratório, um estudo empírico na área de ciências sociais ou a observação de um fenômeno cultural ou da erupção de um vulcão num determinado momento e lugar. Dados digitais de pesquisa ocorrem na forma de diferentes tipos de dados, como números, figuras, vídeos, softwares; com diferentes níveis de agregação e de processamento, como dados crus ou primários, dados intermediários e dados processados e integrados; e em diferentes formatos de arquivos e mídias (Sayão; Sales, 2020, p. 33).

Em 2007, a *Organization Economic Co-operation and Development* (OECD), definiu explicitamente o termo “dados de pesquisa”, como registros de fatos usados como fontes primárias para a pesquisa científica e que são comumente aceitos na comunidade como necessários para validar resultados da pesquisa. (Gava, *et al*. 2024).

Grant (2017, tradução nossa) acredita que

> [...] embora os dados de pesquisa sejam muitas vezes pensados em termos de pesquisas científica, eles podem também representar outros contextos, por exemplo, dados de censos, histórias orais e pesquisas longitudinais de Ciências Sociais e resultados de pesquisas em artes e humanidades.

Sayão e Sales (2020, p. 32) definem que os dados de pesquisa são:

> [...] dados brutos coletados diretamente por um instrumento ou um sensor e agregados a partir de múltiplas fontes; ou podem ser produtos de um modelo

> teórico, simulação ou visualização; ou de experimentos conduzidos na bancada de um laboratório [...].

De acordo com Semeler; Pinto (2019, p.116) os dados de pesquisa podem ser:

> documentos, questionários, avaliações, registros de casos, protocolos de estudo, planilhas, notas de laboratório, notas de campo, diários, filmes, imagens, arquivos digitais de áudio e vídeo, sequências genéticas, coordenadas geográficas, banco de dados, algoritmos, metodologias, protocolos, entre outros tipos de manifestação de pesquisa.

Com base nas definições apresentadas, é possível compreender que os dados de pesquisa permeiam diversos contextos e áreas do conhecimento, principalmente nos ambientes digitais.

No entanto, segundo Curty (2017), os dados de pesquisa se constituem como matéria-prima essencial para a ciência e são indispensáveis para novos ciclos de formação e atualização do conhecimento científico, tornando possível a inovação científica e tecnológica. Nesse contexto, emergem modos específicos para a publicação de dados, que promovem tanto o compartilhamento quanto o reaproveitamento dessas informações. Esses modos, são denominados *Data Publishing* ou publicação de dados, que são implementados por meio de periódicos especializados, artigos dedicados a dados e repositórios voltados para dados de pesquisa.

Arakaki e Santos (2022), explicam que, os repositórios de dados de pesquisa estão incluídos os metadados que permitem a representação de dados com vistas a promover sua recuperação. Os metadados correspondem a recursos informacionais que, nos repositórios de dados de pesquisa, devem sempre ser organizados de forma eficiente. Essa organização é fundamental para garantir a descrição, referência, interpretação, utilização e reutilização dos dados em contextos de pesquisa científica.

Na esfera científica, marcada pelo avanço das tecnologias e pela crescente complexidade das pesquisas, as novas formas de produção, organização e análise de dados têm assumido um papel central, permitindo avanços significativos em diversas áreas do conhecimento. É nesse contexto que surge o conceito de *e-science*, caracterizado pelo uso intensivo de dados e pela integração de ferramentas tecnológicas no processo de pesquisa.

Para Gray (2009), a *e-science* é um dos quatro paradigmas nos quais a ciência pode ser categorizada, refletindo a forma como ela foi desenvolvida e continua sendo

praticada ao longo do tempo e com base nas tecnologias disponíveis.

O primeiro paradigma está relacionado à ciência experimental, o segundo à ciência teórica, enquanto o terceiro diz respeito à ciência computacional, marcada pela incorporação de computadores ao processo científico. Já o quarto paradigma representa o modelo atual, caracterizado pelo uso extensivo de máquinas e softwares para gerar dados em grande escala, um modelo amplamente conhecido como *e-science*, ou ciência orientada por dados.

Sendo assim, os dados passam a ser o destaque no fazer científico, sobretudo em ambientes colaborativos, em que o compartilhamento é essencial para o reuso (Felipe e Santos, 2022). Para que essa prática seja eficaz, é fundamental que os dados sejam devidamente representados e compartilhados, permitindo que pesquisadores possam reutilizá-los, economizando tempo e recursos que seriam necessários para uma nova coleta, além disso, os dados podem ser integrados com outros dados e surgiram novas perspectivas de investigação.

Nesse sentido, a implementação de estruturas e ferramentas tecnológicas torna-se indispensável para promover a divulgação e o compartilhamento de dados. Diferentemente de objetos informacionais, como livros ou artigos, os dados possuem estruturas específicas que requerem abordagens próprias.

2.1 Tipos de repositórios de dados

Repositórios podem ser classificados em diferentes tipos, dependendo do seu propósito e da natureza dos objetos digitais que armazenam. Em geral, eles podem ser categorizados em repositórios institucionais, temáticos, de dados e de *software.*

De acordo com Azambuja (2019), os repositórios institucionais, por exemplo, são utilizados para armazenar e disponibilizar a produção acadêmica e científica de uma instituição específica, enquanto os repositórios temáticos reúnem conteúdo de áreas específicas do conhecimento, sendo úteis para comunidades científicas com interesses focados em temas determinados. Os repositórios de dados, por sua vez, armazenam dados de pesquisa e são essenciais para a preservação e o compartilhamento de dados científicos, enquanto os repositórios de *software* guardam código-fonte de programas desenvolvidos no contexto de projetos científicos ou acadêmicos.

Segundo Sayão e Sales (2016), existe um consenso entre os pesquisadores

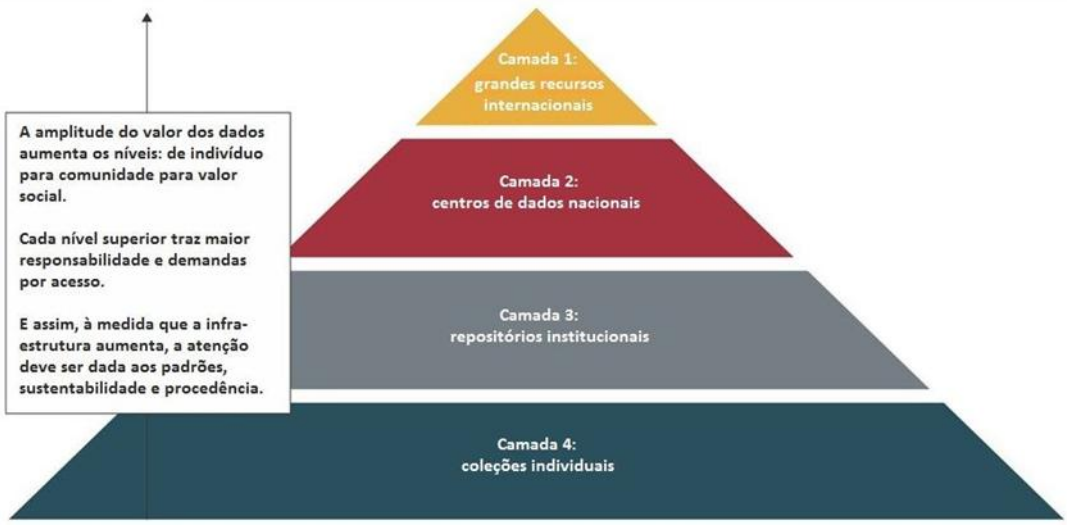
em relação à classificação dos repositórios de dados, que pode ser dividida em quatro categorias principais.

- Repositórios institucionais de dados de pesquisa: são administrados dentro de uma instituição acadêmica ou de pesquisa e, normalmente, armazenam apenas a produção científica gerada pela própria instituição. Devido a essas características, costumam ser multidisciplinares na maioria dos casos.
- Repositórios disciplinares, ou temáticos, de dados de pesquisa: são voltados ao armazenamento de dados de áreas específicas de estudo e geralmente são gerenciados por grupos ou organizações dedicadas a um determinado campo do conhecimento científico.
- Repositórios multidisciplinares de dados de pesquisa: são aqueles que recebem dados de diferentes áreas do conhecimento e de múltiplas fontes, incluindo pesquisadores de diversas instituições.
- Repositórios de dados de pesquisa orientados por projetos: são destinados ao armazenamento dos dados gerados por um grupo ou projeto específico, concentrando-se exclusivamente na produção vinculada a esse projeto.

Azambuja (2019) diz que, essa classificação é útil para diferenciar os tipos de repositórios de dados com base em quem pode submeter materiais e no tipo de conteúdo armazenado. No entanto, ela não reflete a complexidade das infraestruturas desses repositórios em termos de durabilidade, acesso e abrangência dos dados armazenados.

De acordo com o The Royal Society (2012), os repositórios podem ser organizados em uma estrutura hierárquica, semelhante a uma pirâmide, o que se demonstra na figura 1, de acordo com o alcance dos dados que eles armazenam, sua importância percebida e o custo associado à sua manutenção.

Figura 1 – Pirâmide de dados segundo a The Royal Society



Fonte: Science as na open enterprise – The Royal Society (2012, p. 60)

Essa hierarquia é composta por quatro níveis: no topo estão os repositórios de âmbito internacional, seguidos pelos nacionais, depois pelos institucionais, e, na base, encontram-se os repositórios comunitários ou individuais. Cada camada reflete um alcance e relevância distintos, assim como diferentes exigências de recursos para preservação e gerenciamento dos dados.

Azambuja (2019), diz que, para que haja um bom repositório de dados é necessário que ele seja capaz de suprir as necessidades de descrever, preservar e permitir a descoberta dos dados, para isso, Bonetti *et al.* (2024) aponta que é necessário que seguir boas práticas, como os princípios FAIR, para potencializar seus benefícios. Essa necessidade, está alinhada diretamente com os princípios FAIR, ou seja, isso indica que os dados devem ser localizáveis (*Findable*), acessíveis (*Accessible*), interoperáveis (*Interoperable*) e reutilizáveis (*Re-usable*).

De acordo com a FORCE11 (2016), os princípios norteadores FAIR são os seguintes:

Quadro 2 – Apresentação dos Princípios Norteadores FAIR

Para o dado ser localizável:	**Para o dado ser acessível:**	**Para o dado ser interoperável:**	**Para o dado ser reutilizável:**
•F1 - os metadados estão atribuídos a um identificador globalmente exclusivo e eternamente	•A1 - os metadados são recuperáveis pelo seu identificador usando um protocolo de comunicação padronizado;	•I1 - os metadados usam uma linguagem formal, acessível, compartilhada e amplamente	•R1 - os metadados tem uma pluralidade de atributos precisos e relevantes; – R1.1 - os metadados são

persistente; •F2 - os dados estão descritos com metadados ricos; •F3 - os metadados estão registrados ou indexados em um recurso pesquisável; •F4 - os metadados especificam o identificador de dados.	– A1.1 - o protocolo é aberto, gratuito e universalmente implementável; – A1.2 - o protocolo permite um procedimento de autenticação e autorização, quando necessário; •A2 - os metadados estão acessíveis, mesmo quando os dados não estão mais disponíveis.	aplicável para a representação do conhecimento; •I2 - os metadados usam vocabulários que seguem os princípios FAIR; •I3 - os metadados incluem referências qualificadas a outros metadados.	liberados com uma licença de uso de dados clara e acessível; – R1.2 - os metadados estão associados à sua proveniência; – R1.3 - os metadados atendem aos padrões da comunidade relevantes ao domínio.

Fonte: Elaborado pelo autor, baseado FORCE11, 2016, (tradução nossa).

Os princípios FAIR fornecem diretrizes fundamentais para orientar produtores de dados, ampliando o valor dos dados gerados a partir de publicações. Esses princípios estão alinhados diretamente com o uso de padrões de metadados, evidenciando a relevância dos metadados de proveniência para a contextualização e reutilização dos dados. Além disso, os repositórios de dados, por sua vez, constituem ambientes propícios para o uso e reuso de informações, que podem ser aprimorados em conformidade com os princípios FAIR (WILKINSON, 2016).

2.2 O surgimento dos repositórios de dados de pesquisa

A era da transformação digital trouxe consigo um aumento significativo na produção de dados em ambientes digitais, gerados tanto por cidadãos quanto por instituições em diversos setores da sociedade. Esse cenário exige uma organização cada vez mais detalhada e precisa das informações, com foco em uma abordagem orientada por dados. (Gava, *et al*., 2024).

Como resultado, surgem novas oportunidades para oferecer serviços inovadores e personalizados aos usuários, baseados nessa vasta quantidade de informações estruturadas. Junto a essa transformação digital, emerge o conceito de dataísmo, ou *dataism*, que promove uma análise dos contextos informacionais com ênfase nos dados. Esse enfoque busca alcançar maior precisão e identificar padrões de comportamento até então não observados, proporcionando compreensões mais profundos sobre os fenômenos analisados (Brooks, 2013).

Segundo Semeler e Pinto (2019, p. 115) “A geração de dados está atrelada a todas as coisas usadas no dia a dia. Os dados são colecionados sobre qualquer coisa,

a qualquer momento e em qualquer lugar.” Como resultado, cresce a preocupação em analisar os diversos tipos de dados gerados pelo uso cada vez mais intenso das Tecnologias de Informação e Comunicação (TIC).

Nesse cenário, emerge o conceito de curadoria digital, que pode ser entendido como "um conjunto de atividades de gestão e preservação de dados, com o propósito de torná-los acessíveis de maneira rápida e a qualquer momento" (Silva et al., 2021, p. 568). A proveniência desses dados, ou seja, o rastreamento de sua origem, transformações e contextos de uso, é fundamental para garantir sua autenticidade, confiabilidade e reutilização em pesquisas futuras.

As pesquisas sobre curadoria digital começaram a ganhar destaque com o foco na gestão de dados de pesquisa, especialmente no contexto da *eScience*. De acordo com Costa e Cunha (2014), o termo pode ser encontrado sob outras nomenclaturas, como ciência orientada a dados, computação intensiva em dados, ciberinfraestrutura ou quarto paradigma, entre outros. Essas abordagens destacam a importância de documentar a proveniência dos dados, garantindo que sua trajetória seja transparente e verificável, o que é essencial para a credibilidade científica.

A curadoria digital se refere ao tratamento de grandes volumes de dados no ambiente científico, o que gera a necessidade de criar repositórios digitais, ou plataformas de preservação, para gerenciar esses dados específicos. Esses dados são produzidos ao longo de atividades de pesquisa e têm como principal objetivo o reuso em diferentes contextos científicos (Gava, et al., 2024). Nesse sentido, a proveniência desempenha um papel crucial, pois permite que os usuários compreendam a origem e as transformações dos dados, facilitando sua interpretação e aplicação em novos estudos. A integração de metadados de proveniência, como os propostos pela Família PROV, em repositórios digitais é, portanto, uma prática essencial para fortalecer a confiabilidade e a interoperabilidade dos dados científicos.

De acordo com, Costa e Braga (2016), os dados de pesquisa também são elementos essenciais para o cumprimento dos quatro objetivos da ciência aberta. Os quatro elementos da ciência aberta são definidos por Gezelter (2009): o primeiro ponto aborda a importância de garantir transparência na metodologia, na observação e na coleta de dados. O segundo destaca a necessidade de tornar os dados de pesquisa acessíveis ao público e disponíveis para reutilização. O terceiro enfatiza a importância de oferecer acesso aberto às publicações científicas. Por fim, o quarto ponto ressalta a utilização de ferramentas baseadas na web como meio de promover a colaboração

científica.

Conforme observado por Walport e Brest (2011), os repositórios de dados de pesquisa têm se mostrado uma abordagem eficaz para organizar, preservar e compartilhar informações. Contudo, os autores ressaltam que esses sistemas precisam ser bem estruturados e equipados com ferramentas adequadas para descrever e disseminar os dados de forma a facilitar seu acesso amplo e sua reutilização.

Portanto, ao considerar os conceitos de dados de pesquisa, este trabalho tem como um dos focos centrais refletir sobre o papel dos repositórios de dados de pesquisa no contexto da Ciência da Informação, destacando a importância da utilização de metadados de proveniência que auxiliam na preservação dos dados em repositórios de dados de pesquisa.

Considerando a importância dos dados de pesquisa, a relação ao uso de metadados de proveniência para busca, recuperação e interoperabilidade dos recursos informacionais.

2.3 Contexto de Metadados

As dificuldades para a compreensão dos metadados advêm, muitas vezes, da complexidade de compreensão estrutural e das inúmeras variações possíveis, de acordo com a maneira em que estejam organizados e apresentados. Além disso, uma vez que diferentes comunidades e sistemas podem adotar abordagens distintas na definição e na organização de seus dados, é importante adotar medidas documentais de instrução.

Embora o uso de metadados já estivesse presente em diversos contextos antes da década de 1960, o termo "metadados" foi introduzido por Jack E. Myers em 1960 para designar um conjunto de dados relacionados a outros dados. No entanto, o conceito explícito de metadados, como entendemos hoje, não se consolidou nessa época, mas sim na década de 90, com o surgimento do padrão de metadados *Dublin Core* (DC) em 1995. Durante os anos 60 e 70, os metadados eram utilizados em sistemas de bancos de dados e outras áreas, mas ainda de forma implícita, sem a formalização do termo e das definições que ganhariam destaque mais tarde (Vellucci, 1998; Haynes, 2018).

No entanto, seu uso pode variar de acordo com a área de conhecimento em

que é aplicado, já que os metadados são gerados com objetivos específicos e dentro de contextos e domínios distintos.

Nessa circunstância, Joudrey, Taylor e Wisser (2018, p. 181-182, tradução nossa) acrescentam que:

> Os metadados podem incluir informações descritivas sobre o contexto, qualidade e condição, ou características dos dados. Esta definição implica que os metadados incluem não apenas informações descritivas, como aquelas encontradas em ferramentas de recuperação tradicionais para fins de descoberta de recursos, mas também informações necessárias para a gestão, uso e preservação do recurso de informação (por exemplo, dados sobre onde o recurso está localizado, como ele é exibido on-line, sua propriedade, sua condição).

Sendo assim, a interpretação e o uso eficaz dos metadados são essenciais para a descrição, recuperação, gerenciamento, interoperabilidade e autenticidade digital de um recurso (Chowdhury; Chowdhury, 2007; Haynes, 2004).

Na literatura, contudo, não há um comum entendimento do que caracteriza a definição de metadados. Diversos autores tentaram conceituar o termo, sendo "dados sobre dados" a definição clássica e mais amplamente difundida. Entretanto, essa definição é considerada simplista e contribui pouco para a compreensão aprofundada do conceito (Silva, 2013).

Um elemento de metadado é constituído por um atributo e pelo seu conteúdo. O atributo determina o significado do elemento e suas especificações, enquanto o conteúdo é o dado descrito pelo atributo. A sintaxe do elemento estabelece sua estrutura, e a semântica define seu significado, ou seja, um esquema de metadados é precedido por regras outrora definidas, isso garante com que o preenchimento dos campos seja padronizado, proporcionando sentido e coerência na descrição (Duval et al., 2002; Grácio, 2002; Alves, 2010).

Autores, como Grácio (2002, p. 21), defendem o entendimento de metadados como um conjunto de elementos de dados:

> [...] cujo número é variável de acordo com o padrão, e que descreve o conteúdo de um recurso, possibilitando a um usuário ou a um mecanismo de busca acessar e recuperar esse recurso. Esses elementos descrevem informações como nome, descrição, localização, formato, entre outras, que possibilitam um número maior de campos para pesquisas.

Ademais, estudiosos como, Boeres e Arellano (2005, p.7) defendem que “para preservação digital as informações devem ser preservadas de modo a evitar que sejam, corrompidas, ao criar uma estrutura que guarde o conteúdo e a estrutura da informação, para isto podemos usar os metadados”.

> Todas as formas de preservação digital, exceto as mais simples, podem se beneficiar pela criação, manutenção e evolução de Metadados detalhados para apoio aos processos de preservação. Por exemplo, metadados podem documentar o processo técnico associado com a preservação especificar informações de direitos autorais e estabelecer o conteúdo digital. Eles podem registrar a cadeia de custódia de um objeto digital e identificá-lo individualmente tanto interna como externamente em relação ao arquivo em que reside. Em resumo, a criação e instalação de metadados para preservação parece ser um componente chave para as estratégias de preservação. (OCLC/RGL, 2001, p.2 apud Bodê, 2008, p.63)

O termo pode ter sentido e definição distintos em relação as áreas de estudo, contudo, o consenso que as reúne é que seu objetivo principal é a descrição de recurso de informacionais. Diante do escopo deste estudo, a definição que mais se adequa é apresentada por Grácio (2002, p.23), em sua dissertação, “Conjunto de elementos que descrevem as informações contidas em um recurso, com o objetivo de possibilitar sua busca e recuperação.”

Ainda de acordo com Grácio (2002, p.23), complementa que:

> O conjunto de elementos ou o conjunto semântico de campos representa o conteúdo do recurso descrito, ou seja, as informações que possibilitam identificar o que o recurso representa e o que ele contém. Esse conjunto pode ter um número de elementos variável de acordo com o padrão de metadados utilizado. Os elementos devem conter dois tipos de informações: - descritivas, ou seja, aquelas referentes às características explícitas do recurso, tais como título, data, formato, tipo etc; - e temáticas, de conteúdo intelectual, ou seja, aquelas que expressam o conteúdo do recurso, tais como palavras-chave e referências cruzadas.

A criação dos diferentes tipos de metadados respeitam, definitivamente, uma lógica, isso possibilita a estruturação, disponibilização e interoperabilidade com outros padrões. Duval et al. (2002, p. 1, tradução nossa) define princípios como “[...] conceitos considerados comuns a todos os domínios de metadados e que dão suporte ao desenvolvimento de qualquer esquema de metadados ou aplicação”.

Para viabilizar a interoperabilidade entre sistemas de informação, é fundamental utilizar padrões de metadados. Alves (2010, p. 47-48), define os padrões de metadados,

> Os padrões de metadados são estruturas de descrição constituídas por um conjunto predeterminado de elementos de metadados (atributos codificados ou identificadores de uma entidade) metodologicamente construídos e padronizados. O objetivo do padrão de metadados é descrever uma entidade gerando uma representação unívoca e padronizada que possa ser utilizada para recuperação da mesma.

Logo, os padrões de metadados podem ser compreendidos como recursos informacionais que:

> [...] possibilitam a troca de informações entre instituições que utilizam o mesmo

> padrão ou até mesmo entre aquelas que utilizam padrões diferentes. Isso é importante, pois além de diminuir o trabalho de descrição de recursos, permite que um usuário possa, em uma única pesquisa, buscar informações em diferentes instituições. Grácio (2002, p.25).

Dessa forma, diante das contribuições dos autores mencionados, adota-se o conceito de metadados como um conjunto de elementos que descrevem as informações contidas em um recurso, com o objetivo de facilitar sua busca e recuperação, conforme apresentado por Grácio (2002). Além disso, os padrões de metadados são compreendidos como estruturas padronizadas que viabilizam a descrição unívoca de entidades e a interoperabilidade entre diferentes sistemas de informação, como definido por Alves (2010). Portanto, a abordagem adotada neste estudo está alinhada com a concepção de que metadados e seus padrões são fundamentais para a organização, preservação e compartilhamento eficiente de recursos informacionais.

3 METADADOS E PROVENIÊNCIA

Nesta seção, foram discutidos os aspectos relacionados aos metadados, incluindo suas definições e o conceito de padrão de metadado, bem como a análise da produção científica do termo. Além disso, foram explorados os princípios que definem a questão da Proveniência, sua importância para a manutenção e gestão da informação. Também foram apresentados a Família PROV difundida pelo W3C e os benefícios da adoção desses metadados nos repositórios de dados.

3.1 Metadados de Proveniência

Diante desse cenário, a gestão eficaz da informação não se limita apenas à sua representação e segurança, mas também à preservação de sua informação de proveniência por meio de metadados específicos. Conforme destacado pelo Arquivo Nacional (2015, p. 140), proveniência é o “Termo que serve para indicar a entidade coletiva, entidade coletiva pessoa ou família produtora de arquivo”. A inclusão e identificação de metadados de proveniência em sistemas de gestão de informação é crucial, pois não apenas autentica a origem e a autoria dos dados, mas também possibilita a rastreabilidade e a contextualização histórica necessárias, oferecendo assim, credibilidade aos repositórios.

Na área da Biblioteconomia, os estudos sobre proveniência estão tradicionalmente ligados à propriedade de exemplares individuais de livros, especialmente aqueles considerados especiais. A proveniência dos materiais bibliográficos é singular para compreender sua história e contexto, facilitando a gestão eficiente das coleções e a preservação do patrimônio cultural. Segundo Haynes (2018, p. 134, tradução nossa) “Quando se trata de estabelecer a autenticidade de um item, sua história torna-se importante, sua proveniência: as circunstâncias de sua criação, quem a possuiu e as condições sob as quais sua propriedade foi transferida.” Metadados detalhados sobre a proveniência não apenas enriquecem o conhecimento sobre os itens da coleção, mas também auxiliam na autenticação, na valorização do material e na promoção de seu uso educacional e acadêmico.

Segundo Gil e Miles (2013), a proveniência descreve o contexto de como os dados serão coletados, permitindo o uso concreto dessas informações. Ela também é crucial para determinar a propriedade e os direitos sobre um objeto, além de facilitar

julgamentos sobre a confiabilidade das informações. Além disso, as informações de proveniência ajudam a verificar se o processo utilizado para obter um resultado está em conformidade com requisitos específicos, possibilitando a reprodução ou replicações do método utilizado na geração dos dados.

Assim como na arquivologia, em outras disciplinas como a museologia, a proveniência desempenha um papel fundamental na garantia da origem e da autenticidade de itens museológicos e obras de arte.

> A proveniência é algo que dominou o comércio no mundo da arte. A ideia de que uma pintura é o que ela pretende ser (por exemplo, saber por quem ela foi pintada, quando, que não é uma falsificação ou cópia) afetará seu valor percebido. Essa ideia também se aplica a livros impressos e outros artefatos físicos, onde pode haver um valor associado a um manuscrito original ou a uma primeira edição. Essa ideia foi adotada no mundo comercial e se aplica à documentação associada a transações comerciais. (HAYNES, 2018, p. 134, tradução nossa)

Josserand (2016) enumera três aspectos sob os quais as informações de proveniência são capazes de ser pontuadas como informações relevantes:

> O conhecimento da proveniência é útil em três níveis: 1º do ponto de vista administrativo, fornece informações sobre a situação do documento; 2º do ponto de do ponto de vista da investigação, é uma ferramenta de reflexão múltipla; 3º do ponto de vista de coleções, melhora o conhecimento delas, contribui para a história das bibliotecas e pode ser uma ferramenta para promovê-los ao público. Apresenta também um certo interesse na proteção e segurança das coleções.

Na arquivologia, a questão da proveniência, auxilia de maneira intrínseca na gestão dos acervos. Grande parte da literatura que versa sobre proveniência se encontra nas áreas em torno da arquivologia, deve se destacar o uso de bases de dados que utilizam sempre metadados para descrição do objeto informacional. O meio tecnológico proporciona uma ampla gama de recursos para a gestão eficiente dos repositórios de dados de pesquisa. Segundo Arakaki (2019), com a facilidade de copiar informações digitais, os metadados de proveniência tornam-se importantes na medida em que asseguram a fonte original que a informação está sendo extraída.

Esses metadados incluem informações sobre a proveniência dos documentos, o que é essencial para entender sua origem, histórico de custódia e contextos relacionados. No ambiente tecnológico atual, diversas ferramentas e sistemas são utilizados para armazenar, organizar e recuperar informações de maneira precisa e eficaz. Segundo Pomerantz (2015), existem várias normas e padrões que podem ser empregados para representar a proveniência.

> Diversos esquemas de metadados que existem atualmente, e a padronização que ocorreu em outros domínios e para outros casos de uso (*Dublin Core* para uso geral, *Art & Architecture Thesaurus* para objetos de arte, EXIF para imagens digitais, etc.) ainda tem de emergir para proveniência. Esses esquemas de proveniência compartilham muitas características: todos eles são compostos de conjuntos de elementos que identificam características do recurso ou de entidades que o influenciaram, e todos categorizam relacionamentos entre recursos e entidades. (POMERANTZ, 2015, p. 103, tradução nossa).

Dessa forma, a proveniência se estende por diversos contextos. Segundo Lemieux e ImProvenance Group (2016) é fundamental abordar a questão da proveniência de maneira multidisciplinar ou interdisciplinar, já que diferentes áreas frequentemente utilizam as mesmas tecnologias para representar informações de proveniência.

O contexto de metadados de proveniência na literatura acadêmico-científica, em especial no campo da Ciência da Informação, destaca a crescente relevância do registro da origem e da trajetória dos dados em ambientes de pesquisa.

3.2 Análise da Produção Científica sobre metadados de proveniência

No ambiente acadêmico e científico contemporâneo, as bases de dados desempenham um papel essencial no processo de internacionalização do conhecimento, permitindo o acesso e a disseminação de informações em escala global. Nesse cenário, os metadados de proveniência emergem como um elemento crítico para assegurar a confiabilidade e a rastreabilidade dos dados utilizados em pesquisas, segundo Arakaki (2020), a representação da proveniência é essencial para garantir a confiabilidade dos dados, uma vez que oferece informações detalhadas sobre quem criou ou alterou um recurso, as ações que levaram à sua modificação e detalhes sobre o próprio recurso.

De acordo com Moreau e Missier (2013), os metadados de proveniência referem-se a informações sobre entidades, atividades e indivíduos envolvidos na criação de um dado, permitindo a avaliação de sua qualidade, segurança e confiabilidade.

Na Ciência da Informação, onde a organização e recuperação de informação são processos centrais, o papel dos metadados de proveniência se torna ainda mais relevante. Conforme Alves (2010), os padrões de metadados consistem em estruturas

descritivas que utilizam conjuntos predefinidos de informações, construídos de forma metodológica e padronizada. Dessa forma, a correta aplicação de padrões de metadados assegura uma descrição uniforme e de qualidade, promovendo o intercâmbio de informações, a interoperabilidade entre sistemas e a eficiência na recuperação dos dados.

Para verificar a produção científica sobre metadados de proveniência, foi conduzida uma busca nas bases de dados *Scopus* e *Web of Science* com o objetivo de avaliar, por meio de métricas, a utilização do termo em inglês *"Provenance metadata"*.

Conforme os autores, Bufrem, Silva, Sobral e Correia (2016), reforçam que ambas as bases de dados permitem pesquisar e classificar os resultados por parâmetros específicos, tais como, primeiro autor, citação, instituição, entre alternativas.

Além disso, as duas bases proveem a possibilidade de analisar as informações a partir de *rankings* e métricas construídos com base no alto grau de sistematização dos dados contidos nas bases (CHADEGANI et al., 2013).

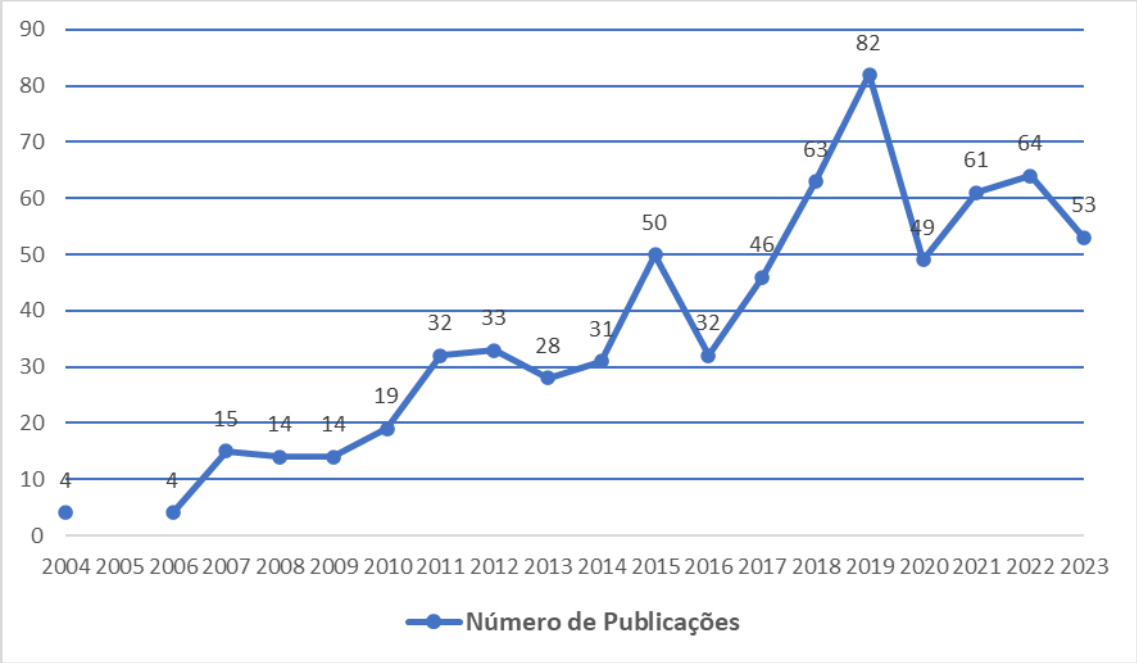
A análise incluiu todos os tipos de documentos disponíveis nessas bases, como artigos de periódicos, anais de conferências e outros materiais acadêmicos, para fornecer uma visão completa do impacto e da relevância do conceito de metadados de proveniência no cenário acadêmico global.

Além disso, a investigação se concentrou em identificar a frequência de publicações, contribuindo para uma visão mais clara sobre a evolução e a crescente adoção dos metadados de proveniência como um tema crucial para a confiabilidade e rastreabilidade de dados em ambientes de pesquisa.

Para atingir os objetivos desta pesquisa, foi realizado um tratamento para verificar a duplicidade de informações nos documentos, com os dados extraídos em formato Microsoft Excel®. Os registros foram destacados em relação ao DOI, autor, título e identificador das bases. Após essa etapa, os dados foram organizados e analisados utilizando ferramentas específicas para a verificação de duplicidades e repetições, visando oferecer uma visualização mais clara dos resultados. O período de análise abrange de 2004 a 2023 porque a primeira recuperação de dados ocorre a partir de 2004, e 2023 é utilizado para abranger um período anual completo de análise. Na base Scopus, foram encontrados 627 registros, enquanto a base Web of Science identificou 90 registros. Desses, 73 registros apresentaram sobreposição, com 4

casos de duplicidade na Scopus e 2 duplicidades na WoS.

Gráfico 1 - Publicações por ano na Web of Science e na Scopus



Fonte: Silva e Arakaki (2024)

Como resultado obteve-se na Scopus 606 publicações, enquanto a WoS contava com apenas 15 publicações. Dentre elas, 73 registros foram identificados em ambas as plataformas, com 6 casos de duplicidade, indicando uma concordância nas publicações indexadas. Assim, o total de publicações contabilizadas nas duas bases soma 694 documentos. Dessa forma, observa-se uma crescente produção científica acerca do tema.

3.3 PROV *Ontology* (PROV-O)

A Ontologia PROV (PROV-O) integra a família de documentos PROV, resultado dos esforços do *Provenance Working Group* do W3C, com o objetivo de possibilitar a representação e o intercâmbio de informações de proveniência geradas em diferentes sistemas (ARAKAKI, 2019; LEBO; SAHOO; MCGUINNESS, 2013).

No ano de 2013, esse mesmo grupo de trabalho, elaborou um conjunto de especificações que define um modelo para a interoperabilidade da proveniência de

dados na *Web*, a Família PROV. Esta compreende quatro recomendações principais: o Modelo de Dados PROV (PROV-DM), a Ontologia PROV (PROV-O), a Notação de Proveniência (PROV-N) e as Restrições do Modelo de Dados PROV (PROV-CONSTRAINTS). Além disso, foram publicadas oito (8) notas que fornecem orientações adicionais e informações sobre o modelo PROV, auxiliando no mapeamento e na compreensão detalhada do mesmo.

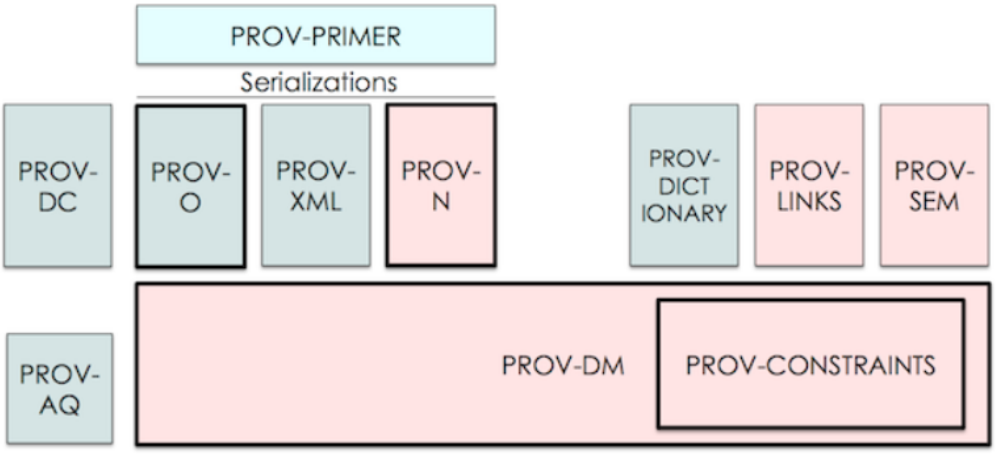
De acordo com Groth e Moreau (2013) a família de documentos PROV define um modelo, serializações e outras definições de apoio correspondente que permitem o intercâmbio de informações de proveniência em ambientes heterogêneos como a *Web*.

De acordo com o W3C, o documento da Família PROV oferece uma visão geral não normativa, estabelecendo padrões/sugestões para seu uso. Proveniência refere-se a informações sobre as entidades, atividades e pessoas envolvidas na criação de um dado ou objeto, sendo útil para avaliar sua qualidade e confiabilidade. O objetivo do PROV é facilitar a ampla publicação e troca de informações sobre procedência na *Web* e em outros sistemas de informação. Ele permite a representação e intercâmbio dessas informações em formatos amplamente usados, como o *Resource Description Framework* (RDF) e *Extensible Markup Language* (XML).

De acordo com Arakaki (2019) a proposta do padrão PROV não é abranger todas as especificidades de vários domínios, mas fornecer um conjunto de metadados para garantir um mínimo de informações de proveniência aplicável a todos domínios.

Para visualizar melhor a relação dos documentos mencionados acima, a Figura 2, ilustra os documentos da Família PROV.

Figura 2 - Família PROV



Fonte: Groth e Moreau (2013, não paginado)

O W3C contextualiza a figura 2, afirmando que, em seu núcleo está um modelo de dados conceituais (PROV-DM), que define um vocabulário comum para descrever a proveniência. Este modelo é instanciado por várias serializações, utilizadas por implementações para a troca de informações de proveniência. Para auxiliar desenvolvedores e usuários a expressar proveniência válida, um conjunto de restrições (PROV-Constraints) é definido, o que pode ser utilizado para criar validadores de proveniência. Esse conjunto é complementado por uma semântica formal (PROV-SEM).

Além disso, para fornecer suporte adicional à troca de proveniência, são disponibilizadas especificações adicionais, na W3C, para protocolos que permitem localizar e acessar a proveniência (PROV-AQ), conectar pacotes de descrições de proveniência (PROV-Links), representar coleções no estilo de dicionário (PROV-Dictionary) e definir como interoperar com o vocabulário amplamente utilizado *Dublin Core* (PROV-DC).

O PROV-O [...] expressa o modelo de dados PROV (PROV-DM) usando a OWL2 *Web Ontology Language* (OWL2). (LEBO; SAHOO; MCGUINNESS, 2013, n.p., tradução nossa), permitindo que seja realizado o mapeamento do PROV-DM para RDF. De acordo com Lebo, Sahoo e McGuinness (2013, n.p., tradução nossa), o PROV-O:

> [...] contém um conjunto de classes, propriedades e restrições que podem ser usadas para representar e trocar informações de proveniência criadas em diferentes sistemas e em diferentes contextos. Também pode ser especializado para criar novas classes e propriedades para modelar informações de proveniência para diferentes aplicações e domínios.

Os termos do PROV-O são divididos em três categorias: termos de ponto de partida (*starting point terms*), termos expandidos (*expanded terms*) e termos qualificados (*qualified terms*) (LEBO; SAHOO; MCGUINNESS, 2013).

Os termos de ponto de partida fornecem um conjunto fundamental de classes e propriedades que constituem a base do PROV-O. Essa categoria inclui três classes e nove propriedades. No quadro 3 nomeia-se as classes.

Quadro 3 – Classes de ponto de partida do PROV-O

Classe	**Descrição**
prov:Entity	Uma coisa física, digital, conceitual ou outro tipo

	de coisa com alguns aspectos fixos; entidades podem ser reais ou imaginárias.
prov:Activity	Algo que ocorre durante um período e atua sobre ou com entidades; pode incluir consumir, processar, transformar, modificar, realocar, usar ou gerar entidades.
prov:Agent	Algo com forma de responsabilidade por uma atividade que está ocorrendo, pela existência de uma entidade ou pela atividade de outro agente.

Fonte: adaptação e tradução de Lebo, Sahoo e McGuinness (2013).

O quadro 4, apresenta as propriedades de ponto de partida.

Quadro 4 – Propriedades de ponto de partida do PROV-O

Propriedade	**Descrição**
prov:wasGeneratedBy	Descreve o que gerou a entidade. Geração é a conclusão da produção de uma nova entidade por uma atividade.
prov:wasDerivedFrom	Descreve a entidade preexistente da qual uma derivação foi originada. Uma derivação é uma transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova ou a construção de uma nova entidade com base em uma entidade preexistente.
prov:wasAttributedTo	Indica a qual agente uma entidade foi atribuída. Atribuição éa designação de uma entidade a um agente.
prov:startedAtTime	Momento ou hora em que uma atividade teve início. Início é quando uma atividade é considerada iniciada por uma entidade, conhecida como gatilho. Qualquer uso, geração ou invalidação envolvendo uma atividade segue o início da atividade.
prov:used	Descreve uma entidade que foi usada por uma atividade. Uso é o começo da utilização de uma entidade por uma atividade.
prov:wasInformedBy	Descreve uma atividade a1 que informou uma atividade a2. Comunicação é a troca de uma entidade por duas atividades, uma atividade usando a entidade gerada pela outra.
prov:endedAtTime	Momento ou hora em que uma atividade teve fim. Fim é quando uma atividade é considerada finalizada por uma entidade, conhecida como gatilho. Qualquer uso, geração ouinvalidação envolvendo uma atividade precede o fim da atividade.
prov:wasAssociatedWith	Descreve o agente responsável pela ocorrência de uma atividade. Uma associação de atividades é uma atribuição de responsabilidade a um agente para uma atividade, indicando que o agente teve uma função na atividade. Além disso, permite que um plano seja especificado, que é o plano pretendido pelo agente para atingir alguns objetivos no contexto desta atividade.
prov:actedOnBehalfOf	Descreve o agente ao qual foi delegada responsabilidade. Delegação é a atribuição de autoridade e responsabilidade a um agente (por si ou por outro agente) para realizar uma atividade específica como delegado ou representante, enquanto o agente que ele atua em nome de mantém a responsabilidade pelo resultado do trabalho delegado. Por exemplo, um aluno agiu em nome de seu supervisor, que

	agiu em nome do presidente do departamento, que agiu em nome da universidade; todos esses agentes são responsáveis de alguma forma pela atividade que ocorreu, mas não dizemos explicitamente quem é o responsável e em que grau.

Fonte: adaptação e tradução de Lebo, Sahoo e McGuinness (2013).

Já na categoria de termos expandidos, está compreende 7 classes e 16 propriedades. As classes dessa categoria estão listadas no quadro 5.

Quadro 5 – Classes expandidas do PROV-O

Classe	**Descrição**
prov:Collection	Uma coleção é uma entidade que fornece uma estrutura para alguns constituintes, que também são entidades.
prov:EmptyCollection	Uma coleção vazia é uma coleção sem membros.
prov:Bundle	Um pacote é um conjunto nomeado de descrições de proveniência e é ele próprio uma Entidade, permitindo que a proveniência de proveniência seja expressa.
prov:Person	Agentes pessoais são pessoas.
prov:SoftwareAgent	Um agente de software é o *software* em execução.
prov:Organization	Uma organização pode ser instituição social ou jurídica, como empresa, sociedade etc.
prov:Location	Um local pode ser um local geográfico identificável (ISO 19112), mas também pode ser um local não geográfico, como um diretório, linha ou coluna. Como tal, existem inúmeras maneiras pelas quais o local pode ser expresso, como por uma coordenada, endereço, ponto de referência etc.

Fonte: adaptação e tradução de Lebo, Sahoo e Mccguinness (2013).

No quadro 6, estão detalhadas as propriedades expandidas do PROV-O, oferecendo uma visão mais abrangente dos termos dessa categoria.

Quadro 6 – Propriedades expandidas do PROV-O

Propriedade	**Descrição**
prov:alternateOf	Duas entidades alternativas apresentam aspectos da mesma coisa. Esses aspectos podem ser iguais ou diferentes, e as entidades alternativas podem ou não se sobrepor no tempo.
prov:specializationOf	Uma entidade que é uma especialização de outra compartilha todos os aspectos desta e, adicionalmente, apresenta aspectos mais específicos da mesma coisa que a última. Exemplos de aspectos incluem um período, uma abstração e um contexto associado à entidade.
prov:generatedAtTime	Momento ou hora em que uma entidade foi completamente criada e está disponível para uso.

prov:hadPrimarySource	Descreve a fonte primária de uma derivação. Uma fonte primária para um tópico refere-se a algo produzido por algum agente com experiência direta e conhecimento sobre o tópico, no momento do estudo do tópico, sem se beneficiar em retrospectiva. Como tal, é importante que as fontes secundárias façam referência às fontes primárias das quais elas derivaram, para que sua confiabilidade possa ser investigada.
prov:value	Descreve um valor que é uma representação direta de uma entidade.
prov:wasQuotedFrom	Descreve a entidade da qual derivou a citação. Uma citação é a repetição (de parte ou da totalidade) de uma entidade, como texto ou imagem, por alguém que pode ou não ser o autor original. A citação é um caso particular de derivação.
prov:wasRevisionOf	Descreve a entidade da qual derivou a revisão. Uma revisão é uma derivação para a qual a entidade resultante é uma versão revisada de algum original. A implicação aqui é que a entidade resultante contém conteúdo substancial do original. A revisão é um caso particular de derivação.
prov:invalidatedAtTime	Momento ou hora em que uma entidade foi invalidada. Invalidação é o início da destruição, cessação ou expiração de uma entidade existente por uma atividade. Qualquer geração ou uso de uma entidade precede sua invalidação.
prov:wasInvalidatedBy	Descreve o que foi responsável pela invalidação de uma entidade.
prov:hadMember	Uma coleção é uma entidade que fornece uma estrutura para alguns constituintes, que também são entidades. Diz-se que esses constituintes são membros das coleções.
prov:wasStartedBy	Descreve a entidade que deu início a uma atividade.
prov:wasEndedBy	Descreve a entidade que finalizou uma atividade.
prov:invalidated	Descreve a entidade que foi invalidada.
prov:influenced	Influência é a capacidade de uma entidade, atividade ou agente de afetar o caráter, desenvolvimento ou comportamento de outro por meio de uso, início, fim, geração, invalidação, comunicação, derivação, atribuição, associação ou delegação.
prov:atLocation	A localização de qualquer recurso.
prov:generated	Descreve a entidade que foi gerada.

Fonte: adaptação e tradução de Lebo, Sahoo e Mccguinness (2013).

As classes e propriedades expandidas, podem ser aplicadas para estabelecer conexões entre as classes dos termos iniciais, com muitas delas atuando como subclasses e subpropriedades dos termos de ponto de partida (LEBO; SAHOO; MCGUINNESS, 2013).

A categoria de termos qualificados abrange um total de 20 classes e 25 propriedades. As classes qualificadas estão organizadas e apresentadas no quadro 7, detalhando seus papéis e funcionalidades dentro do PROV-O.

Quadro 7 – Classes qualificadas do PROV-O

Classe	Descrição
prov:Influence	Influência é a capacidade de uma entidade, atividade ou agente de afetar o caráter, desenvolvimento ou comportamento de outro por meio de uso, início, fim, geração, invalidação, comunicação, derivação, atribuição, associação ou delegação.
prov:EntityInfluence	A capacidade de influência de uma entidade.
prov:Usage	Uso é o começo da utilização de uma entidade por uma atividade.
prov:Start	Começo é quando uma atividade é considerada iniciada por uma entidade, conhecida como gatilho.
prov:End	Fim é quando uma atividade é considerada finalizada por uma entidade, conhecida como gatilho.
prov:Derivation	Uma derivação é uma transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova ou a construção de uma nova entidade com base em uma entidade preexistente.
prov:PrimarySource	Uma fonte primária para um tópico refere-se a algo produzido por algum agente com experiência direta e conhecimento sobre o tópico, no momento do estudo do tópico, sem se beneficiar em retrospectiva.
prov:Quotation	Uma citação é a repetição (de parte ou da totalidade) de uma entidade, como texto ou imagem, por alguém que pode ou não ser o autor original.
prov:Revision	Uma revisão é uma derivação para a qual a entidade resultante é uma versão revisada de algum original.
prov:ActivityInfluence	A capacidade de influência de uma atividade.
prov:Generation	Geração é a conclusão da produção de uma nova entidade por uma atividade.
prov:Communication	Comunicação é a troca de uma entidade por duas atividades, uma atividade usando a entidade gerada pela outra.
prov:Invalidation	Invalidação é o início da destruição, cessação ou expiração de uma entidade existente por uma atividade.
prov:AgentInfluence	A capacidade de influência de um agente.
prov:Attribution	Atribuição é a designação de uma entidade a um agente.
prov:Association	Uma associação de atividades é uma atribuição de responsabilidade a um agente para uma atividade, indicando que o agente teve uma função na atividade.
prov:Plan	Um plano é uma entidade que representa um conjunto de ações ou etapas pretendidas por um ou mais agentes para atingir alguns objetivos.
prov:Delegation	Delegação é a atribuição de autoridade e responsabilidade a um agente (por si ou por outro agente) para realizar uma atividade específica como delegado ou representante, enquanto o agente que ele atua em nome de mantém a responsabilidade pelo resultado do trabalho delegado.
prov:InstantaneousEvent	Um evento instantâneo, ou evento abreviado, acontece no mundo e marca uma mudança no mundo, em suas atividades e em suas entidades. Os eventos incluem geração, uso ou invalidação de entidades, bem como início ou término de atividades.
prov:Role	O papel é a função de uma entidade ou agente em relação a uma atividade, no contexto de uso, geração, invalidação, associação, início e fim.

Fonte: adaptação e tradução de Lebo, Sahoo e Mccguinness (2013).

O quadro 8, apresenta-se de forma detalhada as propriedades qualificadas do PROV-O, destacando suas funções e como elas se relacionam com as demais categorias de termos.

Quadro 8 – Propriedades qualificadas do PROV-O

Propriedade	**Descrição**
prov:wasInfluencedBy	Indica entidade, atividade ou agente responsável por influenciar outro.
prov:qualifiedInfluence	Influência é a capacidade de uma entidade, atividade ou agente de afetar o caráter, desenvolvimento ou comportamento de outro por meio de uso, início, fim, geração, invalidação, comunicação, derivação, atribuição, associação ou delegação.
prov:qualifiedGeneration	Geração é a conclusão da produção de uma nova entidade por uma atividade.
prov:qualifiedDerivation	Uma derivação é uma transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova ou a construção de uma nova entidade com base em uma entidade preexistente.
prov:qualifiedPrimarySource	Uma fonte primária para um tópico refere-se a algo produzido por algum agente com experiência direta e conhecimento sobre o tópico, no momento do estudo do tópico, sem se beneficiar em retrospectiva.
prov:qualifiedQuotation	Uma citação é a repetição (de parte ou da totalidade) de uma entidade, como texto ou imagem, por alguém que pode ou não ser o autor original.
prov:qualifiedRevision	Uma revisão é uma derivação para a qual a entidade resultante é uma versão revisada de algum original.
prov:qualifiedAttribution	Atribuição é a designação de uma entidade a um agente.
prov:qualifiedInvalidation	Invalidação é o início da destruição, cessação ou expiração de uma entidade existente por uma atividade.
prov:qualifiedStart	Começo é quando uma atividade é considerada iniciada por uma entidade, conhecida como gatilho.
prov:qualifiedUsage	Uso é o começo da utilização de uma entidade por uma atividade.
prov:qualifiedCommunication	Comunicação é a troca de uma entidade por duas atividades, uma atividade usando a entidade gerada pela outra.
prov:qualifiedAssociation	Uma associação de atividades é uma atribuição de responsabilidade a um agente para uma atividade, indicando que o agente teve uma função na atividade.
prov:qualifiedEnd	Fim é quando uma atividade é considerada finalizada por uma entidade, conhecida como gatilho.
prov:qualifiedDelegation	Delegação é a atribuição de autoridade e responsabilidade a um agente (por si ou por outro agente) para realizar uma

	atividade específica como delegado ou representante, enquanto o agente que ele atua em nome de mantém a responsabilidade pelo resultado do trabalho delegado.
prov:influencer	Essa propriedade é usada como parte do padrão de influência qualificado. Subclasses de prov:Influence usam essas subpropriedades para referenciar o recurso (Entidade, Agente ou Atividade) cuja influência está sendo qualificada.
prov:entity	Essa propriedade (prov:entity) faz referência a uma entidade (prov:Entity) que influenciou um recurso. Ela se aplica a uma influência de entidade (prov:EntityInfluence), fornecida por uma subpropriedade de prov:qualifiedInfluence da entidade (prov:Entity), atividade (prov:Activity) ou agente (prov:Agent) influenciado.
prov:hadUsage	O uso opcional envolvido na derivação de uma entidade.
prov:hadGeneration	A geração opcional envolvida na derivação de uma entidade.
prov:activity	Essa propriedade (prov:activity) faz referência a uma atividade (prov:Activit)y que influenciou um recurso. Ela se aplica a uma influência de atividade (prov:ActivityInfluence), fornecida por uma subpropriedade de prov:qualifiedInfluence da entidade (prov:Entity), atividade (prov:Activity) ou agente (prov:Agent) influenciado.
prov:agent	Essa propriedade (prov:agent) faz referência a um agente (prov:Agent) que influenciou um recurso. Ela se aplica a uma influência de agente (prov:AgentInfluence), fornecida por uma subpropriedade de prov:qualifiedInfluence da entidade (prov:Entity), atividade (prov:Activity) ou agente (prov:Agent) influenciado.
prov:hadPlan	O plano opcional adotado por um agente em associação com alguma atividade
prov:hadActivity	A atividade opcional de uma influência, que usou, gerou, invalidou ou era de responsabilidade de alguma entidade
prov:atTime	Momento ou hora em que um evento instantâneo (prov:InstantaneousEvent) ocorreu, no formato xsd:dateTime.
prov:hadRole	A função opcional que uma entidade assumiu no contexto de uma atividade.

Fonte: adaptação e tradução de Lebo, Sahoo e Mccguinness (2013).

Após a apresentação da ontologia PROV, é possível relacionar e mapear seus elementos para os padrões do domínio bibliográfico em ambientes digitais como os repositórios de dados de pesquisa. No entanto, mesmo com essas limitações, observa-se que as definições das classes e propriedades apresentam dificuldades de entendimento, sendo muitas vezes genéricas ou repetitivas, o que gera confusão.

Esse detalhamento da ontologia PROV já foi explorado em outros trabalhos acadêmicos, como a tese de Arakaki (2019) e o estudo de Tomoyose (2021). Esses autores contribuíram significativamente para o entendimento e aplicação da ontologia em diferentes contextos, abordando de forma aprofundada as classes, propriedades e o mapeamento dos termos do PROV.

3.4 DUBLIN CORE

O termo *"Dublin Core"* originou-se no primeiro *Workshop* realizado em 1995 em Dublin, Ohio, EUA, onde um grupo de especialistas em informação começou a desenvolver um conjunto básico de elementos de metadados para descrever recursos digitais de forma simples e consistente.

Ressalta-se que o tratamento e a preocupação com os sistemas interoperáveis e as instituições que fazem uso dos metadados abrangem desde a forma de criação até sua divulgação, com ênfase na manutenção dos dados. De mesmo modo, como já apontado anteriormente, intensifica-se a preocupação com a confiança informacional, afetando diretamente a credibilidade das informações disponibilizadas em repositórios digitais, visto que os metadados possuem papel fundamental na infraestrutura para descrição de recursos informacionais (Arakaki, 2017).

No contexto do objetivo delineado por este trabalho, apresenta-se a seguir, os dados obtidos da coleta de metadados DC, considerados metadados de proveniência ou que podem conter informações de proveniência segundo o W3C, dentro da plataforma *DSpace* que sugere, predominantemente, o uso dos elementos DC.

Ainda segundo o W3C, muitos termos de DC podem ser usados para descrever informações de proveniência sobre um recurso: quando foi afetado no passado, quem o afetou e como foi afetado. O restante dos termos DCMI (metadados de descrição) nos informam o que foi afetado. No quadro 9, pode-se ver os termos DC de acordo com as quatro subcategorias (o quê?, quem?, quando? e como?).

Quadro 9 – Categorização dos Termos DC.

Categoria	**Subcategoria**	**Termos**
Metadados descritivos	O que	abstract, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, bibliographicCitation, conformsTo, coverage, description, educationLevel, extent, format, hasPart, isPartOf, identifier, instructionalMethod,

		isRequiredBy, language, mediator, medium, relation, requires, spatial, subject, tableOfContents, temporal, title, type
Proveniência	Quem	contributor, creator, publisher, rightsHolder
Proveniência	Quando	available, created, date, dateAccepted, dateCopyrighted, dateSubmitted, issued, modified, valid
Proveniência	Como	accessRights, hasFormat, hasVersion, isFormatOf, isVersionOf, license, isReferencedBy, isReplacedBy, references, replaces, rights, source

Fonte: Baseado no Grupo de Trabalho do W3C.

A classificação é, de certa forma, reduzida, pois pode-se argumentar que certos elementos incluídos nos metadados de descritivos também podem conter informações sobre a proveniência, dependendo de como são utilizados. As categorias são explicadas com mais detalhes seguir:

- Termos/metadados descritivos (O quê?): Esta categoria abrange todos os termos que descrevem um recurso sem referir-se à proveniência, totalizando 25 de 55 termos. Exemplos incluem *dct:title, dct:abstract, dct:description* de um recurso, ou *dct:format*, que indica o formato em que o recurso está disponível.
- Termos do agente (quem?): Esta categoria inclui termos relacionados a agentes. Todas as propriedades têm *dct:Agent* como alcance, ou seja, um recurso que age ou tem a capacidade de agir. Termos como *dct:contributor, dct:creator* e *dct:publisher* claramente influenciam o recurso, sendo fundamentais para determinar sua origem. Embora isso não seja tão evidente para *dct:rightsHolder*, a propriedade é considerada uma informação crucial sobre a proveniência de muitos recursos, como obras de arte, e por isso está incluída nesta categoria.
- Termos de data e hora (quando?): Esta categoria abrange termos relacionados a data e hora. As datas fazem parte do registro de proveniência de um recurso, pois indicam quando algo foi criado (*dct:created*), modificado (*dct:modified*), publicado (*dct:issued*), entre outros. Duas datas têm uma relevância especial para a proveniência*: dct:available e dct:valid*. Elas se diferenciam das outras datas porque, por definição, podem representar um intervalo de tempo. Frequentemente, o período de disponibilidade ou validade de um recurso é

inerente a ele e conhecido previamente – como a validade de um passaporte ou a duração de uma oferta especial publicada na *Web*. Nesses casos, não há nenhuma ação específica que torne o recurso inválido ou indisponível; é simplesmente determinado pelo período de validade.

- Termos de derivação e licenciamento (como?): Esta categoria inclui termos relacionados à derivação. Quando um recurso é derivado de outros, o original se torna parte da cadeia de proveniência do derivado. No DC, as derivações podem ser classificadas como versões (*dct:isVersionOf*), serializações de formato (*dct:isFormatOf*), substituições (*dct:replaces*) e fontes de informação (*dct:source*). A relação *dct:references* é mais fraca (referenciar um recurso não implica necessariamente que o conteúdo se baseia nele), mas pode-se presumir que um recurso referenciado influenciou o recurso descrito e, portanto, é relevante para sua proveniência. As propriedades inversas não contribuem diretamente para a proveniência do recurso descrito; por exemplo, um recurso geralmente não é afetado diretamente por ser referenciado ou utilizado como fonte. No entanto, as propriedades inversas são relevantes para descrever as relações entre os recursos. Além disso, licenciamento (*dct:license*), direitos (*dct:rights*) e acesso (*dct:accessRights*) também fazem parte da proveniência do recurso, pois determinam e explicam como o recurso pode ser usado para futuras derivações.

Em síntese, a W3C evidência os termos *Dublin Core* que podem ser considerados informações de proveniência, como descrito acima, vale ressaltar também, que os termos citados estão inclusos, em sua maioria, dentro da categoria de “Metadados Descritivos”, sendo que sua definição é semelhante entre os autores, o que diferencia são as subcategorias que são usadas para dividi-los (Arakaki, 2019). A definição de metadados administrativos foi relatada por Joundrey, *et al.* (2018, p. 191-192), como:

> [...] criados para fins de gerenciamento, tomada de decisão e manutenção de registros. Eles fornecem informações sobre os requisitos técnicos, de preservação e armazenamento de recursos de informações, especialmente objetos digitais. Metadados administrativos auxiliam no monitoramento, acesso, reprodução, digitalização e backup de recursos digitais.

Enquanto os metadados descritivos facilitam a identificação e a descoberta de recursos, os metadados administrativos garantem sua preservação e acessibilidade a

longo prazo. A relação entre essas categorias evidencia a complementaridade, importante para atender às demandas técnicas, gerenciais e de pesquisa, destacando a importância de padrões como *Dublin Core* e a constante evolução no entendimento e aplicação de conceitos relacionados à proveniência e à administração de dados no cenário atual.

3.5 DATACITE METADATA SCHEMA

O termo *DataCite Metadata Schema* refere-se a um padrão de metadados utilizado para a descrição, publicação e citação de dados de pesquisa e outros recursos acadêmicos. Para este trabalho, foi adotada a versão 4.6 do *DataCite Metadata Schema*, lançada em dezembro de 2024, conforme documentado pela organização *DataCite*[2].

O consórcio *DataCite* foi estabelecido em 2009 por instituições acadêmicas e centros de informação de destaque, com o objetivo de facilitar o acesso on-line a dados de pesquisa e aumentar sua os números de citação (Brase, 2009; Petritsch, 2017).

O *DataCite Metadata Schema* é uma estrutura desenvolvida para a descrição, publicação e citação de dados de pesquisa e outros recursos acadêmicos. Inspirado no padrão *Dublin Core* (Weibel, 1997), seus elementos são organizados em três níveis de obrigatoriedade: obrigatório (M), recomendado (R) e opcional (O). Essa estrutura visa garantir uma identificação precisa e consistente dos recursos digitais, permitindo sua citação e recuperação em repositórios acadêmicos.

A versão 4.6 do *DataCite Metadata Schema*, contém 19 elementos principais distribuídos entre seis campos obrigatórios, seis recomendados e sete opcionais. Elementos essenciais como "*Identifier*", "*Creator*", "*Title*", "*Publisher*", "*PublicationYear*" e "*ResourceType*" são obrigatórios e asseguram que cada recurso seja descrito com informações fundamentais e identificáveis.

[2] chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://datacite-metadata-schema.readthedocs.io/_/downloads/en/4.6/pdf/

Elementos como "*Subject*", "*Contributor*", "*Description*" e "*GeoLocation*" são recomendados para enriquecer a descoberta e reutilização dos recursos (*DataCite Metadata Schema* v4.6, 2024).

Além disso, de acordo com o manual do *DataCite Metadata Schema*, pontua que ele foi concebido para ser genérico e aplicável a uma ampla gama de conjuntos de dados, independentemente da disciplina acadêmica. Seu principal objetivo é sustentar a citação e a descoberta de dados, sem substituir esquemas de metadados específicos de uma área do conhecimento. Assim, ele é uma base mínima que pode ser expandida com metadados disciplinares mais detalhados. Os usuários são incentivados a fornecer metadados em inglês, além de outros idiomas exigidos por financiadores ou instituições.

Uma característica importante da versão 4.6 é o suporte aprimorado para publicações textuais. Tipos de recursos específicos como periódico, artigo de periódico, dissertação e caderno computacional foram incorporados para melhorar a identificação. Outra adição relevante foi a propriedade "*relatedItem*", projetada para descrever recursos relacionados ao item registrado, como um artigo de conferência ou uma série de livros. Isso permite registrar informações mesmo quando não há um identificador associado ao recurso relacionado.

Há disponível no site o mapeamento do *DataCite Metadata Schema* para *Dublin Core*[3], e isso permite a promoção da interoperabilidade entre diferentes esquemas de metadados, facilitando a descrição e o compartilhamento de recursos digitais em repositórios acadêmicos e plataformas de gestão de dados de pesquisa (*DataCite Metadata Schema* v4.6, 2024).

3.6 DDI – DATA DOCUMENTATION INITIATIVE

[3] https://datacite-metadata-schema.readthedocs.io/en/4.6/mappings/dublincore-qualified/

O *Data Documentation Initiative* (DDI)[4], é um padrão de metadados amplamente utilizado nas áreas de ciências sociais, comportamentais e econômicas. Desenvolvido em 1995, em Quebec, no Canadá, por um grupo internacional focado na gestão de metadados (Vardigan, 2013), o DDI visa descrever dados e processos de pesquisa de maneira estruturada, abrangendo diversas etapas do ciclo de vida dos dados.

O DDI é composto por três padrões principais que atendem a diferentes necessidades de documentação e integração de dados. O DDI *Codebook* (DDI-C) é destinado à documentação de dados de pesquisa simples ou únicos, contendo informações detalhadas sobre um único conjunto de dados, similar a um livro de códigos, sendo amplamente utilizado em repositórios como o *Dataverse*, baseado em XML, é ideal para uso em contextos mais específicos.

Já o DDI – *Lifecycle*, que oferece uma descrição detalhada de todas as etapas do ciclo de vida dos dados, desde a concepção inicial e coleta até o arquivamento. Diferentemente de padrões que se concentram apenas nos resultados da coleta, o DDI – *Lifecycle* também incorpora informações sobre os processos de criação, normalização e análise, utilizando o formato XML para possibilitar a integração dos dados com outros sistemas.

Por fim, o DDI *Cross-Domain Integration* (DDI-CDI) é focado na integração de dados entre disciplinas, enfatizando a descrição e processos dos dados, e é ideal para lidar com dados multidimensionais, big data e estruturas complexas. Baseado em UML (*Unified Modeling Language*), o DDI-CDI pode ser expresso em diferentes formatos, incluindo XML, sendo voltado para contextos interdisciplinares que exigem interoperabilidade.

A aplicação prática do DDI é evidenciada em repositórios como o *Indepth Data Repository*, utilizado para registrar dados demográficos e de saúde em populações de baixa e média renda LMICs, na sigla em inglês. Estudos como os de Sankoh et al. (2014) demonstram como o padrão é empregado para documentar informações como título da pesquisa "*titl*", número de identificação "*IDNo*" e data de publicação "*prodDate*", estruturando metadados em XML para facilitar o compartilhamento e reutilização.

O modelo do DDI apresenta características descritas por Ball (2012), que

[4] https://www.ddialliance.org/

organiza as fases do ciclo de vida em um esquema detalhado. No entanto, Gupta e Müller-Birn (2018) apontam limitações, como o foco predominante em dados primários, pouca ênfase na publicação de dados analisados e agilidade restrita em alguns cenários. Essas características fazem com que a adoção do DDI seja mais frequente em áreas onde a coleta de dados originais é prioritária.

Ainda que as descrições completas das fases do ciclo de vida não estejam detalhadas no site oficial do DDI, o padrão é amplamente reconhecido por sua relevância em projetos de ciências sociais, desempenhando um papel essencial na normalização e gerenciamento de metadados de pesquisa. Essa abordagem reflete a relevância do DDI como uma base para o planejamento e organização de informações em estudos complexos e interdisciplinares.

4 PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa classifica-se como de natureza teórica, ou seja, buscou levantar informações particulares referente a área de estudo, com o incentivo de ajudar o corpo científico e acadêmico. Sendo assim, o estudo tem como alicerce os conceitos de dados, metadados e proveniência tendo como fundamentos os teóricos e estudiosos da Ciência da Informação e Ciência da Computação.

Ao considerar os objetivos da pesquisa, esta constitui-se como exploratória. Trata-se de uma investigação de cunho científico a qual levantou uma série de documentos relevantes sobre o tema, a fim de dar o alicerce teórico e, também, auxiliar na compreensão das atribuições identificadas no texto. Assim sendo, a busca pela literatura fundamental não se reduziu ao âmbito nacional, mas também ao internacional, principalmente os de língua inglesa e espanhola, tendo em vista sua ampla contribuição acerca do tema.

A análise da literatura existente sobre proveniência possibilitou o desenvolvimento de uma base teórica relacionada aos metadados de proveniência. Esse estudo envolveu a investigação da proveniência em diferentes contextos, contribuindo para a compreensão do problema e a definição dos resultados esperados. Assim, foi conduzida uma pesquisa bibliográfica que abrangeu temas como: metadados, proveniência de dados e tipologias de metadados.

Nesse contexto, a busca pelo estado da arte, que é o produto que pode ser advindo por meio de uma revisão bibliográfica e visa "[...] identificar e debater uma determinada produção acadêmica em diferentes áreas do conhecimento [...]" (Ferreira, 2002, p. 258). O objetivo é identificar pesquisas não apenas no campo da Ciência da Informação, mas também em áreas correlatas, como a Ciência da Computação, que igualmente tratam de questões relacionadas à proveniência.

Os procedimentos técnicos são embasados em consultas de artigos, livros, teses, dissertações, manuais e trabalhos em eventos, em meios digitais e físicos.

Desse modo, o processo metodológico se configurou da seguinte maneira:

1ª **Etapa** - Levantamento bibliográfico com estratégia de busca, utilizando *strings* como por exemplo: "*provenance AND metadata*", "*metadata provenance*", "*metadata standards AND data repositories*", e "provenance *AND information science*".

Nas respectivas bases de dados:

a. BRAPCI - Base de Dados em Ciência da Informação - Acervo de Publicações Brasileiras em Ciência da Informação;
b. EBSCO - *Industries is an American Company*;
c. Portal de periódicos da CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior;
d. BDTD – Biblioteca Digital de Teses e Dissertações;
e. *Scopus*;
f. *Web of Science*.

2ª **Etapa** - Após o levantamento e seleção do corpus literários foram analisados alguns critérios seguindo a ordem citada:

g. Relevância do tema, classificada com base na temática e pertinência ao escopo do estudo, que aborda (metadados de proveniência e a proveniência).
h. Idiomas (português, inglês e espanhol);
i. Atualidade dos documentos (data de publicação).

3ª **Etapa** – Leitura e fichamento que consiste na identificação e organização da literatura coletada, a fim de agrupar e tabular de forma lógica os conteúdos abordados nessa pesquisa. A leitura das referências citadas e o levantamento do fichamento permitiram a reunião das informações necessárias para elaboração teórica.

4ª **Etapa** - Desenvolvimento da redação e apresentação de qualificação com resultados parciais.

5ª **Etapa** – Coleta de dados.

A coleta de dados consiste no processo de reunir informações relevantes e necessárias para responder à pergunta central da pesquisa, que é: como os esquemas de metadados da Família PROV estão sendo empregados nos repositórios de dados de pesquisa, para proporcionar a garantia da rastreabilidade e autencidade das informações ao longo do tempo.

6ª **Etapa** – Análise de Dados com a utilização do Método *Crosswalk*.

Este método facilita a interoperabilidade semântica, permitindo que sistemas transformem dados de um determinado formato para outros.

7ª **Etapa** – Conclusão da redação e da análise de dados para a defesa da dissertação.

4.1 Universo da pesquisa

Para o levantamento do *corpus* com a finalidade de analisar o trabalho, foram analisados os padrões de metadados mais utilizados nos repositórios de dados científicos disponibilizados pelo Re3data, através dos filtros de busca de padrões de metadados. O recorte justifica-se pela necessidade de compreender quais esquemas de metadados são mais prevalentes nos repositórios de dados de pesquisa e qual a relação deles com os metadados de proveniência, mais especificamente, com a Família PROV. Essa identificação permite uma análise mais aprofundada sobre o uso e a interoperabilidade dos metadados, contribuindo para melhorias na gestão e na recuperação de dados em repositórios digitais.

No universo da Ciência Aberta, o *Research Data Repositories Information*, comumente conhecido como Re3data, serve como uma plataforma que agrega dados de pesquisa de vários campos de estudo, oferecendo-os em vários formatos e tornando-os acessíveis a todos. Este diretório global, que se originou na Alemanha em 2012, simplifica o processo de busca de dados de pesquisa. De acordo com Re3data (2021), os parceiros fundadores foram a Escola de Biblioteconomia e Ciência da Informação de Berlim, o *Helmholtz Open Science Office* no Centro Alemão de Pesquisa em Geociências GFZ, a Biblioteca KIT do Instituto de Tecnologia de Karlsruhe (KIT) e as Bibliotecas da Universidade Purdue.

O Re3data abrange uma grande variedade de áreas do conhecimento, incentivando o compartilhamento, o acesso e a visibilidade ampliada de dados de científicos. Cada resultado obtido no diretório vem acompanhado de uma descrição ou resumo, fornecendo informações sobre as pesquisas relacionadas, métricas de diferentes diretórios, bem como detalhes como país de origem, tema ou palavra-chave e tipo de documento.

Outro fator para escolha do Re3Data é que ela participa do movimento de acesso

aberto, de modo a reunir e compartilhar artigos científicos, além de serem, em grande parte, instituições públicas, possibilitando a obtenção de informações a partir da lei de acesso à informação (Brasil, 2011).

Para análise, foram considerados os documentos norteadores dos repositórios como políticas e manuais, e quando necessário, foram analisados os tipos específicos de registros dos recursos informacionais, sobretudo, seus dados.

Os padrões de metadados selecionados para análise, nesta pesquisa, foram baseadas no critério de utilização nos repositórios de dados científicos, conforme indicado pelo Re3Data, priorizando aqueles que são empregados acima de 200 repositórios. Essa decisão se fundamenta no fato de que padrões como os elementos do Dublin Core, presente em 603 repositórios, o *DataCite Metadata Schema*, com 440 repositórios, o DDI (*Data Documentation Initiative*), com 280 repositórios e o *Repository-Developed Metadata Schema*, utilizado em 221 repositórios, demonstram uma ampla variedade de padrões em contextos de compartilhamento e organização de dados distintos. [5]

A relevância dos padrões de metadados selecionados pode ser entendida pela sua ampla adoção e uso em diferentes contextos científicos, exemplo, o Dublin Core, que é reconhecido por sua simplicidade e flexibilidade em descrever uma grande variedade de recursos digitais, não apenas para fornecer um conjunto básico de elementos de descrição que podem ser usados por catalogadores ou não catalogadores, mas também, para simples descrição de recursos de informação (Weibel, 1997).

Na seção 4, será abordada a explicação os padrões de metadados adotados, no contexto da proveniência, explorando sua origem, e o papel que desempenham na gestão de repositórios de dados, dessa forma, permitirá uma melhor compreensão de como esses padrões contribuem para a preservação, o acesso e a organização de informações dentro do universo dos repositórios de dados científicos.

4.2 Materiais e métodos

O método *crosswalk*, introduzido pela *National Information Standards*

[5] https://www.re3data.org/metrics/metadataStandards

Organization (NISO) em 1999, será empregado como método de análise. Essa abordagem permite a interoperabilidade semântica, permitindo que os diferentes tipos de metadados conversem com outros tipos de padrões de metadados facilitando pesquisas simultâneas em diversos bancos de dados como se fossem uma única entidade (Chan; Zeng, 2006; Baca, 2016). Consequentemente, foi determinado realizar um *crosswalk* da família PROV para os padrões citados no universo da pesquisa, com o objetivo de examinar a compatibilidade da procedência com base no PROV dentro dos padrões propostos pelo W3C no que se refere à proveniência.

Quadro 10 – Apresentação do método *crosswalk*

Etapa	Subetapa	Observação
1° etapa: Harmonização Extração da terminologia comum, propriedades, organização e processos utilizados pelos padrões de metadados, e criar um quadro genérico para que se possa desenvolver novos ou rever padrões de metadados já existentes.	Subetapa A: Terminologia	Utilização de terminologias diferentes dos padrões dificultam o mapeamento entre eles.
		É essencial chegar a um acordo sobre a terminologia dos padrões, além de estabelecer uma definição formal para cada termo.
	Subetapa B: Propriedades - As semelhanças das propriedades dos padrões são extraídas e os conceitos generalizados	Identificadores únicos para cada metadado, por exemplo, tag, etiqueta, identificador.
		Qual definição semântica de cada metadado?
		O metadado é obrigatório, opcional ou obrigatório em certas condições?
		Um metadado pode ocorrer várias vezes?
		Organização dos metadados em relação ao outro, por exemplo, as relações hierárquicas.
		Restrições impostas pelos valores do elemento (texto livre, escala numérica ou data)?
		Suporte opcional para elementos de metadados definidos localmente?
		As propriedades comuns podem ser expressas e utilizadas de uma forma similar dentro de cada padrão? Esta etapa simplifica o desenvolvimento do Crosswalk.
	Subetapa C: Organização	Para facilitar cada padrão deve ser organizado em forma similar, de modo que determinada seção de um padrão possa ser encontrada em uma seção de outro padrão.
	Subetapa D: Processo	Há ocasiões em que a escolha do processo selecionado seja arbitrária e não um processo análogo a outro padrão relacionado.

2° etapa: Mapa semântico	O mapeamento semântico é a especificação de cada elemento do padrão com o elemento semanticamente equivalente para o outro padrão. Para St.Pierre e LaPlant (1999) é o processo mais importante da harmonização e desenvolvimento do Crosswalk, pois determina o mapeamento semântico entre os padrões de metadados de origem e destino.
3° etapa: Mapeando elemento a elemento - Identificar os metadados opcionais e obrigatórios. Nesta fase considerar as propriedades de cada metadado	**Uma para muitos:** ocorrência de vários elementos de origem a uma única ocorrência no elemento alvo. A um elemento que se está verificando irá ser correspondente a diversos elementos do outro padrão de metadados.
	Muitos para um: muitos elementos de um padrão de metadados para apenas um metadado no padrão de destino. Devem-se aproximar todos os elementos do primeiro metadado e indicar a um único elemento do outro padrão. Se a resolução é mapear todos os valores do elemento de origem para um único valor no elemento alvo, regras explícitas são obrigadas a especificar como os valores serão anexados juntos. Caso seja apenas mapear um valor de elemento de origem para o destino, com a possível consequência de perda de informações, a resolução deve indicar os critérios para a seleção de elementos.
	Elementos extras na fonte: Outro caso importante que requer resolução é a manipulação de um elemento de origem que não é mapeado para qualquer elemento apropriado no padrão alvo. Uma vez que muitos padrões fornecem a capacidade de capturar informações adicionais, a resolução deve especificar exatamente como o valor do elemento deve ser adicionado.
	Elementos obrigatórios /não resolvidos em alvo: Em alguns casos, pode haver elementos obrigatórios no alvo que não têm mapeamento correspondente no padrão de metadados de origem. Porque o alvo requer um valor para os elementos obrigatórios, o Crosswalk deve fornecer uma resolução para os seus valores.
4° etapa: Hierarquia, Objeto e Visão Lógica	**Hierarquia:** A maioria dos padrões de metadados organizam seus metadados hierarquicamente. Em alguns casos, a profundidade da hierarquia pode ser fixada. Em outros casos a profundidade da hierarquia é ilimitada.
	Objeto: Item versus coleção. Item é um único documento, ou seja, os metadados associados a um documento. Coleção conjunto de itens, ou seja, os metadados referem-se a mais de um item.
	Visão Lógica: Permite ver um conjunto específico de metadados do padrão organizado de uma maneira específica
	Conversão de conteúdo: Padrões de metadados restringem o conteúdo de cada metadado para um determinado tipo de dado, intervalo de valores, ou vocabulário controlado. Muitas vezes, as conversões são baseadas não só nas propriedades que definem a fonte e os metadados alvo, mas também os conteúdos dos elementos de metadados de origem.
	Combinações de conversão: Quando as propriedades de conversão são consideradas de forma independente, as conversões de metadados podem parecer simples para especificar e processar. Na prática, vários problemas de conversão refletem em uma combinação, o que dificulta a especificação de conversão e processo. Deve considerar as transformações necessárias para converter um metadado alvo, onde várias propriedades são diferentes do metadado de origem

Fonte: Arakaki (2019) baseado em St.Pierre e LaPlant (1999).

Chan e Zeng (2006) enfatizam que a equivalência é crucial para criar um *crosswalk*, pois a correspondência entre os padrões garante a equivalência dos

elementos. Os autores propõem duas abordagens para implementar um *crosswalk*: "*crosswalkning* absoluto" e "*crosswalking* relativo". No primeiro, há uma correspondência exata entre os elementos nos esquemas de origem e destino. Por outro lado, o "*crosswalking* relativo" mapeia elementos do esquema de origem para pelo menos um elemento no esquema de destino. Essa abordagem relativa se mostra mais eficaz ao mapear um padrão de metadados complexo para um mais simples.

Ainda, de acordo com Chan e Zeng (2006), o processo de *crosswalk* encontra vários graus de equivalência, incluindo um-para-um, um-para-muitos, muitos-para-um e um-para-nenhum. O primeiro tipo, um-para-um, ocorre quando um elemento de metadados corresponde perfeitamente a um elemento no padrão de destino. No segundo tipo, um-para-muitos, um único elemento de metadados no padrão inicial pode corresponder a vários elementos no segundo padrão, indicando a existência de múltiplos elementos com contexto similar. O terceiro tipo, muitos-para-um, sugere que múltiplos elementos de metadados no padrão de origem mapeiam para um único elemento no padrão de destino. Por fim, a equivalência um-para-nenhum acontece quando um elemento de metadados no padrão de origem não tem correspondência correspondente no padrão de destino.

No entanto, conforme apresentado por Chan e Zeng (2006), o modelo de análise *crosswalk* não é exclusivo e pode ser adaptado a cada projeto, considerando diferentes tipos de interoperabilidade semântica. As autoras ressaltam que outros processos ou ideias também podem contribuir para a garantia da interoperabilidade.

Logo, para facilitar o entendimento sobre o grau de correspondência, usaremos uma terminologia mais compreensível, incluindo ‘correspondências exatas’, relações em que um elemento é ‘mais genérico’ que outro e relações em que um elemento é ‘mais específico’ que outro.

O primeiro tipo, ‘correspondência exata’, ocorre quando um elemento de metadados corresponde exatamente a um elemento no padrão de destino, ou seja, ambos têm o mesmo nível de abrangência e detalhamento semântico. No segundo tipo, ‘Mais genérico’, um único elemento de metadados no padrão inicial é mais amplo ou abrangente semanticamente do que vários elementos no segundo padrão, indicando que um conceito geral pode ser detalhado de múltiplas formas, isso não quer dizer que o padrão de destino não possa abranger mais que uma definição. No terceiro tipo, ‘Mais específico’, o elemento de metadado no padrão de origem é mais particular do que um único elemento no padrão de destino, sugerindo que o conceito

da origem é mais específico que o destino, pois o metadado de destino engloba mais possibilidades de definição. Por fim, foi adotado, ‘Não há correspondência’, para quando não há qualquer tipo de correspondência entre os padrões, devido a ausência de um padrão de destino.

Para exemplificar o mapeamento de correspondência é possível verificar no quadro 11 os elementos *Dublin Core* e DDI.

Quadro 11 – Exemplo de correspondência entre padrões.

Dublin Core	**DDI**	**Grau de correspondência**	**Observação**
title	titl	Correspondência exata	Ambos os elementos são correspondentes.
creator	AuthEnty	Mais genérico	O elemento *Dublin Core* é mais genérico, ou seja, é muito mais amplo seu significado do que o elemento DDI que indicado a entidade autora.
mediator	contact	Mais específico	O elemento Dublin Core é mais específico que o elemento DDI, ou seja, o elemento Dublin Core indica que um mediador, no contexto educacional, pode ser o pai, professor assistente de ensino ou cuidador, já o elemento DDI representa o contato telefônico, seja do autor, da instituição, ou de qualquer outra natureza não especifica.
format	-	Não há correspondência	Não há correspondência entre o padrão origem e o padrão de destino.

Fonte: Elaborado pelo autor.

O exemplo demonstrado evidencia como os diferentes tipos de correspondência impactam a interoperabilidade entre padrões de metadados. A categorização entre ‘correspondência exata’, ‘mais genérico’ e ‘mais específico’ permite compreender melhor as relações semânticas e tomar decisões mais informadas ao integrar sistemas baseados em diferentes padrões.

4.3 Análise de dados

Após o levantamento dos resultados, bem como suas especificidades, foram analisados os padrões de metadados segundo o conceito de proveniência, apontando características e modelos de proveniência digital e a relação da origem dos dados e metadados para a proveniência de documentos digitais.

Deste modo, este estudo utilizou o método comparativo do *crosswalk*, que consiste em observar as diferenças e semelhanças entre os padrões de metadados analisados e avaliar como os repositórios de dados de pesquisa estruturam a representação da proveniência dos registros informacionais para garantir a autenticidade e a confiabilidade das informações prestadas.

Após o levantamento dos metadados nos repositórios, foi realizado o mapeamento a partir do método *crosswalk*, com o intuito de estabelecer quais são os principais metadados utilizados. A técnica de documentação para a coleta de dados se torna fundamental para que se visualize os procedimentos utilizados nas análises.

O método *crosswalk* foi utilizado na análise pois é um método que contribui para os aspectos de interoperabilidade em ambientes informacionais digitais e entre os esquemas de metadados. Assim, de acordo com a definição fornecida por Riley (2004), Baca (2016), Simionato (2015) e (Castro e Simionato, 2020, p. 15-16), um *crosswalk* é um mapeamento dos elementos, semântica e sintaxe de um esquema de metadados para outros.

O seu objetivo é permitir a identificação formal dos elementos de metadados ou grupos de elementos equivalentes entre diferentes esquemas, de modo a garantir a semântica e a abrangência em cada contexto. Esse método começa com o estudo dos esquemas de metadados de cada domínio: a partir disso, é necessário comparar os elementos descritivos ou elementos fixos com um ou mais padrões, buscando compreender as diferenças e semelhanças em vários níveis. Essa análise tem como objetivo avaliar o grau de interoperabilidade entre os sistemas e representar visualmente as relações de semelhanças e diferenças encontradas. Em resumo, o *crosswalk* é uma abordagem que permite mapear elementos de metadados entre diferentes esquemas, facilitando a troca e a compreensão de informações entre sistemas heterogêneos. Essa técnica é importante para alcançar a interoperabilidade e garantir que os dados sejam adequadamente interpretados em diferentes contextos.

Para identificar os padrões de metadados, o *crosswalk* é uma boa opção, tendo

em vista que ele relaciona os elementos dos metadados bem como suas equivalências. Além disso, de acordo com Castro e Simionato (2020, p.4):

> Esse método também se caracteriza por uma análise descritiva e exploratória: através dela é possível identificar elementos conceituais, a partir da literatura científica das áreas de Ciência da Informação e da Ciência da Computação, a fim de apontar a necessidade da construção de ambientes informacionais digitais estruturados e padronizados, na utilização de ontologia e metadados.

Nesse sentido, St. Pierre e LaPlant (1998) propõem algumas etapas para o processo do *crosswalk*, sendo o mapeamento semântico considerado o mais importante, como apresentado por Silva (2022, p. 36-37):

> é a especificação da equivalência semântica de cada elemento dos padrões. Para um mapeamento significativo, é necessária uma definição clara e precisa dos elementos de cada padrão.

Com isso, é possível identificar como os repositórios de dados de pesquisa estão garantindo a proveniência dos registros informacionais, sustentando a criação de recomendações e incentivos para aplicação de metadados de proveniência em repositórios de dados de pesquisa.

5 Proveniência nos padrões de metadados

Na sequência, envolveram a análise detalhada dos padrões de metadados, que foram coletados do repositório Re3data. Cada um desses padrões, incluindo *Dublin Core, DataCite Metadata Schema e* DDI *(Data Documentation Initiative),* desempenha um papel crucial na descrição e documentação de dados, proporcionando uma base sólida para a compreensão de sua proveniência.

A abordagem adotada consistiu na realização de um mapeamento em *crosswalk* entre cada um desses padrões e a ontologia PROV. Essa prática permitirá identificar as correspondências entre os elementos dos padrões de metadados e as classes e propriedades da ontologia PROV, facilitando a interoperabilidade e a integração de informações.

O mapeamento em *crosswalk* é uma estratégia eficaz para traduzir as especificações de metadados de diferentes esquemas em uma estrutura comum, favorecendo a consistência e a clareza na documentação da proveniência dos dados.

Com a realização desse mapeamento, espera-se não apenas aprofundar o entendimento sobre como a proveniência é abordada em diferentes padrões de metadados, mas também contribuir para o desenvolvimento de práticas que promovam a transparência e a rastreabilidade das informações em ambientes de pesquisa e repositórios de dados. Junto a isso pretende-se dar sugestões de melhorias na descrição de objetos em repositórios de dados, com foco na proveniência.

5.1 Mapeamento DUBLIN CORE

O padrão Dublin Core possui vocabulários compostos por classes e propriedades, neste trabalho foram considerados ambas, tendo em vista abranger os elementos analisados. Para determinar a correspondência entre cada elemento, foram analisadas as definições, bem como as orientações e recomendações de uso fornecidas pelas organizações responsáveis por cada padrão. Esse levantamento possibilitou na identificação da correspondência entre a classes dos termos DC para a PROV, pelo método *Crosswalk.*

Embora a W3C[6] tenha realizado um mapeamento de correspondência entre os padrões de metadados Dublin Core e a família de metadados de proveniência PROV, essa abordagem não contemplou todas as classes e propriedades do Dublin Core, o que limitou o escopo da análise. Em contraste, a presente pesquisa se destaca por incluir a maioria das classes do Dublin Core no mapeamento *Crosswalk,* ampliando a profundidade da análise e proporcionando uma maior compreensão da interoperabilidade semântica. Ademais, não foram encontradas evidências de que o W3C tenha aplicado de forma completa o método de *Crosswalk.*

O Quadro 10, conforme apresentado no trabalho, foi desenvolvido com base nas informações disponibilizadas pela W3C sobre a correspondência de alguns metadados entre o padrão Dublin Core e a família PROV. Embora o texto mencione que a W3C realizou um mapeamento de correspondência entre esses padrões, destaca-se que o referido quadro foi elaborado no âmbito desta pesquisa, ampliando o escopo do mapeamento ao incluir a maioria das classes do *Dublin Core* e aplicando o método *Crosswalk* de forma mais abrangente.

Quadro 12 – *Crosswalk* DC *term* para PROV *term*

Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
Classes					
Agent	Um recurso que atua ou tem o poder de atuar.	Agent	Um agente é algo que tem alguma forma de responsabilidade por uma atividade que ocorre, pela existência de uma entidade ou pela atividade de outro	Correspondência exata	*Crosswalking* absoluto. Ambos dct:Agente prov:Agentreferem-se ao mesmo conceito: um recurso que tem o poder de agir (que então tem responsabilidade por uma atividade, entidade ou outro agente).

[6] https://www.w3.org/TR/prov-dc/
[7] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#http://purl.org/dc/terms/Agent
[8] https://www.w3.org/TR/2013/REC-prov-o-20130430/#wasDerivedFrom

Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
			agente.		
AgentClass	Um grupo de agentes.	-	-	Não há correspondência	-
BibliographicResource	Um livro, artigo ou outro recurso documental.	Entity	Uma entidade é um tipo físico, digital, conceitual ou outro tipo de coisa com alguns aspectos fixos; entidades podem ser reais ou imaginárias.	Mais específico	*Crosswalking* relativo. Um recurso bibliográfico refere-se a livros, artigos etc., que são entidades concretas do PROV.
FileFormat	Um formato de recurso digital.	-	-	Não há correspondência	-
Frequency	Uma taxa na qual algo se repete.	-	-	Não há correspondência	-
Jurisdiction	A extensão ou alcance da autoridade judicial, policial ou outra.	-	-	Não há correspondência	-
LicenseDocument	Um documento legal que dá permissão oficial para fazer algo com um recurso.	Entity	Uma entidade é um tipo físico, digital, conceitual ou outro tipo de coisa com alguns aspectos fixos; entidades podem ser reais ou imaginárias.	Mais específico	*Crosswalking* relativo. Documento que concede permissão para fazer algo em relação a um recurso. Assim, ele é mapeado como um tipo de prov:Entity.
LinguisticSystem	Um sistema de sinais, símbolos, sons, gestos ou regras usadas na	Plan	Um plano é uma entidade que representa	Mais genérico	*Crosswalking* relativo. A dct:LinguisticSysteé um sistema de símbolos, sons,

Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
	comunicação.		um conjunto de ações ou etapas pretendidas por um ou mais agentes para atingir alguns objetivos.		gestos, etc. usado na comunicação. Portanto, o sistema linguístico define o plano a seguir para aprender um determinado idioma.
Location	Uma região espacial ou lugar nomeado.	Location	Um local pode ser um local geográfico identificável (ISO 19112), mas também pode ser um local não geográfico, como um diretório, linha ou coluna.	Correspondência exata	*Crosswalking* absoluto. Ambos dct:Locatione prov:Locationdefinem locais como "regiões espaciais ou locais nomeados".
LocationPeriodOrJurisdicition	A extensão ou alcance da autoridade judicial, policial ou outra.	-	-	Não há correspondência	-
MediaType	Um formato de arquivo ou meio físico.	-	-	Não há correspondência	-
MediaTypeOrExtent	Um tipo ou extensão de mídia.	-	-	Não há correspondência	-
MethotdOfAccrual	O método pelo qual os itens são adicionados a uma coleção.	Plan	Um plano é uma entidade que representa um conjunto de ações ou etapas pretendidas por um ou mais agentes para atingir alguns objetivos.	Mais específico	*Crosswalking* relativo. dct:MethodOfAccrualdefine o método pelo qual os itens são adicionados a uma coleção (ou seja, o método prov:Planseguido na atividade de inserção).

Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
MethodOfInstruction	Um processo usado para gerar conhecimento, atitudes e habilidades que o recurso descrito foi criado para dar suporte.	Plan	Um plano é uma entidade que representa um conjunto de ações ou etapas pretendidas por um ou mais agentes para atingir alguns objetivos.	Mais específico	*Crosswalking* relativo. “Processo que serve para gerar conhecimento, atitude e habilidades”. Como dct:MethodOfInstructiondefine o método associado a uma atividade, ele é mapeado como prov:Plan.
PeriodOfTime	Características temporais do recurso.	-	-	Não há correspondência	-
PhysicalMedium	Um material físico ou transportador.	-	-	Não há correspondência	-
PhysicalResource	Uma coisa material.	Entity	Uma entidade é um tipo físico, digital, conceitual ou outro tipo de coisa com alguns aspectos fixos; entidades podem ser reais ou imaginárias.	Mais específico	Crosswalking relativo. Uma coisa material, que é um tipo concreto de prov:Entity.
Policy	Um plano ou curso de ação de uma autoridade, com a intenção de influenciar e determinar decisões, ações e outros assuntos.	Plan	Um plano é uma entidade que representa um conjunto de ações ou etapas pretendidas por um ou mais agentes para atingir alguns objetivos.	Mais específico	Crosswalking relativo. dct:Policyé definido como "um plano ou curso de ação de uma autoridade, destinado a influenciar e determinar decisões, ações e outros assuntos." Esta é uma especialização de prov:Plan.

Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
ProvenanceStatement	Quaisquer alterações na propriedade e custódia de um recurso desde sua criação que sejam significativas para sua autenticidade, integridade e interpretação.	Bundle	Um pacote é um conjunto nomeado de descrições de proveniência e é, em si, uma Entidade, permitindo assim que a poveniência da procedência seja expressa.	Correspondência exata	*Crosswalking* absoluto. A dct:ProvenanceStatement is defined as "A statement of any changes in ownership and custody of a resource since its creation", which is a container for any provenance related assertion.
RigthsStatement	Uma declaração sobre os direitos de propriedade intelectual (DPI) mantidos sobre um recurso, um documento legal dando permissão oficial para fazer algo com um recurso ou uma declaração sobre direitos de acesso.	Entity	Uma entidade é um tipo físico, digital, conceitual ou outro tipo de coisa com alguns aspectos fixos; entidades podem ser reais ou imaginárias.	Mais específico	Crosswalking relativo. Declaração sobre os direitos intelectuais de um recurso (por exemplo, um Documento). Assim, ele é mapeado como um tipo de prov:Entity.
SizerOrDuration	Uma dimensão ou extensão, ou um tempo gasto para reproduzir ou executar.	-	-	Não há correspondência	-
Standard	Um ponto de referência contra o qual outras coisas podem ser avaliadas ou comparadas.	-	-	Não há correspondência	-
Propriedades					
abstract	Um resumo do recurso.	-	-	Não há correspondência	Um resumo do recurso
accessRig	Informações	-	-	Não há	Informações sobre

Dublin Core	Definição DC[7]	PROV	Definição PROV[8]	Grau de correspondência	Observações
hts	sobre quem acessa o recurso ou uma indicação de seu status de segurança.			correspondência	quem pode acessar o recurso.
accrualMethod	O método pelo qual os itens são adicionados a uma coleção.	-	-	Não há correspondência	O método pelo qual os itens são adicionados a uma coleção.
accrualPeriodicity	A frequência com que os itens são adicionados a uma coleção.	-	-	Não há correspondência	A frequência com que os itens são adicionados a uma coleção.
accrualPolicy	A política que rege a adição de itens a uma coleção.	-	-	Não há correspondência	A política que rege a adição de itens a uma coleção.
alternative	Um nome alternativo para o recurso.	-	-	Não há correspondência	Um nome alternativo para o recurso.
audience	Uma classe de agentes para os quais o recurso é pretendido ou útil.	-	-	Não há correspondência	Uma classe de entidade para quem o recurso é destinado ou útil.
available	Data em que o recurso ficou ou ficará disponível.	-	-	Não há correspondência	Data (geralmente um intervalo) em que o recurso ficou ou ficará disponível.
bibliographicCitation	Uma referência bibliográfica para o recurso.	-	-	Não há correspondência	A prática recomendada é incluir detalhes bibliográficos suficientes para identificar o recurso da forma mais inequívoca possível.
conformsTo	Um padrão estabelecido ao qual o recurso descrito está em conformidade.	-	-	Não há correspondência	Um padrão estabelecido ao qual o recurso descrito está em conformidade.
coverage	O tópico espacial ou temporal do recurso, a aplicabilidade espacial do	-	-	Não há correspondência	Uma entidade responsável por fazer contribuições para o recurso.

Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
	recurso ou a jurisdição sob a qual o recurso é relevante.				
created	Data de criação do recurso.	generatedAtTime	Geração é a conclusão da produção de uma nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.	Mais específico	*Crosswalking* relativo. Propriedade utilizada para descrever o horário de criação de um recurso (ou seja, o horário de sua geração). Mapeamos isso como uma subpropriedade de prov:generatedAtTime porque “criação” é uma das muitas atividades que geram uma entidade (por exemplo, geração inclui modificação, emissão, aceitação etc.).
creator	Uma entidade responsável por criar o recurso.	wasAttributedTo	Attribution é a atribuição de uma entidade a um agente.	Correspondência exata	*Crosswalking* relativo. Um criador é um dos agentes que participou da criação de um recurso. Eles têm a atribuição pelo resultado dessa atividade.
date	Um ponto ou período de tempo associado a um evento no ciclo de vida do recurso.	-		Não há correspondência	Um ponto ou período de tempo associado a um evento no ciclo de vida do recurso
contributor	Uma entidade responsável por fazer contribuições ao recurso.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais específico	*Crosswalking* relativo. Um contribuidor está associado à atividade de criação ou à atualização do recurso. Portanto, ele/ela tem atribuição sobre o

Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
					resultado dessas atividades.
dateAccepted	Data de aceitação do recurso.	generatedAtTime	Geração é a conclusão da produção de uma nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.	Mais específico	*Crosswalking* relativo. Propriedade usada para descrever a data em que o recurso foi aceito. dct:dateAccepted é mapeado como uma subpropriedade de prov:generatedAtTime porque o recurso aceito foi gerado por uma atividade "Aceitar" que pode ter alterado seu estado anterior.
dateCopyrighted	Data de copyright do recurso.	generatedAtTime	Geração é a conclusão da produção de uma nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.	Mais específico	*Crosswalking* relativo. Propriedade usada para descrever a data em que o recurso foi protegido por direitos autorais. dct:dateCopyrighted é mapeado como uma subpropriedade de prov:generatedAtTime porque o recurso protegido por direitos autorais foi gerado por uma atividade "CopyRight" que pode tê-lo alterado em relação ao seu estado anterior.
dateSubmitted	Data de envio do recurso.	generatedAtTime	Geração é a conclusão da produção de uma nova entidade por uma atividade.	Mais específico	*Crosswalking* relativo. Propriedade usada para descrever a data em que o recurso foi enviado. dct:dateSubmitted é mapeado como uma

Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
			Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.		subpropriedade de prov:generatedAtTime porque o recurso enviado foi gerado por uma atividade "Enviar" que pode ter alterado seu estado anterior.
description	Um relato do recurso.	-	-	Não há correspondência	Uma conta do recurso.
educationLevel	Uma classe de agentes, definida em termos de progressão através de um contexto educacional ou de treinamento, para a qual o recurso descrito se destina.	-	-	Não há correspondência	Uma classe de entidade, definida em termos de progressão num contexto educativo ou de formação, à qual se destina o recurso descrito.
extent	O tamanho ou duração do recurso.	-	-	Não há correspondência	O tamanho ou duração do recurso.
format	O formato do arquivo, meio físico ou dimensões do recurso.	-	-	Não há correspondência	O formato do arquivo, meio físico ou dimensões do recurso.
hasPart	Um recurso relacionado que está incluído física ou logicamente no recurso descrito.	-	-	Não há correspondência	Um recurso relacionado que está incluído física ou logicamente no recurso descrito.
hasVersion	Um recurso relacionado que é uma versão, edição ou adaptação do recurso descrito.	-	-	Não há correspondência	Um recurso relacionado que é uma versão, edição ou adaptação do recurso descrito.
hasFormat	Um recurso relacionado que é substancialmente o mesmo que o recurso	alternateOf	Duas entidades alternativas apresentam aspectos da mesma	Mais genérico	*Crosswalking* relativo.

Dublin Core	Definição DC[7]	PROV	Definição PROV[8]	Grau de correspondência	Observações
	descrito pré-existente, mas em outro formato.		coisa. Esses aspectos podem ser os mesmos ou diferentes, e as entidades alternativas podem ou não se sobrepor no tempo.		
identifier	Uma referência inequívoca ao recurso dentro de um determinado contexto.	-	-	Não há correspondência	Uma referência inequívoca ao recurso dentro de um determinado contexto.
instructionalMethod	Um processo usado para gerar conhecimento, atitudes e habilidades que o recurso descrito foi criado para dar suporte.	-	-	Não há correspondência	Um processo, utilizado para gerar conhecimentos, atitudes e competências, que o recurso descrito se destina a apoiar.
isFormatOf	Um recurso relacionado preexistente que é substancialmente o mesmo que o recurso descrito, mas em outro formato.	wasDerivedFrom	Derivação é uma transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova, ou a construção de uma nova entidade com base em uma entidade pré-existente.	Mais específico	*Crosswalking* relativo. dct:isFormatOf refere-se a outro recurso "pré-existente" que é o mesmo, mas em outro formato (de acordo com dct:hasFormat), implicando que o novo recurso é baseado no primeiro.

Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
isPartOf	Um recurso relacionado no qual o recurso descrito está física ou logicamente incluído.	-	-	Não há correspondência	Um recurso relacionado no qual o recurso descrito está incluído física ou logicamente.
isReferenceBy	Um recurso relacionado que faz referência, cita ou aponta para o recurso descrito.	-	-	Não há correspondência	Um recurso relacionado que faz referência, cita ou aponta para o recurso descrito.
isReplacedBy	Um recurso relacionado que suplanta, desloca ou substitui o recurso descrito.	-	-	Não há correspondência	Um recurso relacionado que substitui, substitui ou substitui o recurso descrito.
isRequiredBy	Um recurso relacionado que requer o recurso descrito para dar suporte à sua função, entrega ou coerência.	-	-	Não há correspondência	Um recurso relacionado que requer que o recurso descrito suporte sua função, entrega ou coerência.
issued	Data de emissão formal do recurso.	generatedAtTime	Geração é a conclusão da produção de uma nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.	Correspondência exata	*Crosswalking* relativo. Propriedade usada para descrever a data em que o recurso foi emitido. dct:issued é mapeado como uma subpropriedade de prov:generatedAtTime porque o recurso emitido é uma entidade em si, que foi gerada em um determinado momento.
isVersionOf	Um recurso relacionado do qual o recurso descrito é uma versão, edição ou adaptação.	-	-	Não há correspondência	Um recurso relacionado do qual o recurso descrito é uma versão, edição ou adaptação.

Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
language	Uma linguagem do recurso.	-	-	Não há correspondência	Um idioma do recurso.
license	Um documento legal que dá permissão oficial para fazer algo com o recurso.	-	-	Não há correspondência	Um documento legal que dá permissão oficial para fazer algo com o recurso.
mediator	Uma entidade que media o acesso ao recurso.	-	-	Não há correspondência	Uma entidade que medeia o acesso ao recurso e para quem o recurso se destina ou é útil.
medium	O material ou portador físico do recurso.	-	-	Não há correspondência	O suporte material ou físico do recurso.
modified	Data em que o recurso foi alterado.	generatedAtTime	Geração é a conclusão da produção de uma nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.	Mais específico	*Crosswalking* relativo. Propriedade utilizada para descrever a data em que o recurso foi modificado. dct:modified é mapeado como uma subpropriedade de prov:generatedAtTime porque o recurso modificado foi gerado por uma atividade "Modificar" que o alterou de seu estado anterior.
provenance	Uma declaração de quaisquer alterações na propriedade e custódia do recurso desde sua criação que sejam significativas para sua autenticidade, integridade e interpretação.	-	-	Não há correspondência	Uma declaração de quaisquer alterações na propriedade e custódia do recurso desde a sua criação que sejam significativas para a sua autenticidade, integridade e interpretação.
publisher	Uma entidade responsável por disponibilizar o	wasAttributedTo	*Attribution* é a atribuição de uma	Mais específico	*Crosswalking* relativo. Um editor tem a atribuição do

Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
	recurso.		entidade a um agente.		recurso publicado após participar da atividade editorial que o gerou.
references	Um recurso relacionado que é referenciado, citado ou de outra forma apontado pelo recurso descrito.	wasDerivedFrom	Derivação é uma transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova, ou a construção de uma nova entidade com base em uma entidade pré-existente.	Mais específico	*Crosswalking* relativo. No PROV, uma derivação é definida como "uma transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova, ou a construção de uma nova entidade baseada em uma entidade pré-existente". Se um recurso n1 faz referência a outro recurso o1 então a construção de n1 é baseada em o1, mesmo que o1 não influencie n1 significativamente. A remoção da referência a o1 em n1 levaria à construção de outro recurso n1, diferente de n1.
relation	Um recurso relacionado.	-	-	Não há correspondência	Um recurso relacionado que é referenciado, citado ou apontado de outra forma pelo recurso descrito.
replaces	Um recurso relacionado que é suplantado, deslocado ou substituído pelo recurso descrito.	-	-	Não há correspondência	Um recurso relacionado.
requires	Um recurso relacionado que é exigido pelo recurso descrito para dar suporte à sua função,	-	-	Não há correspondência	Um recurso relacionado que é suplantado, deslocado ou substituído pelo recurso descrito.

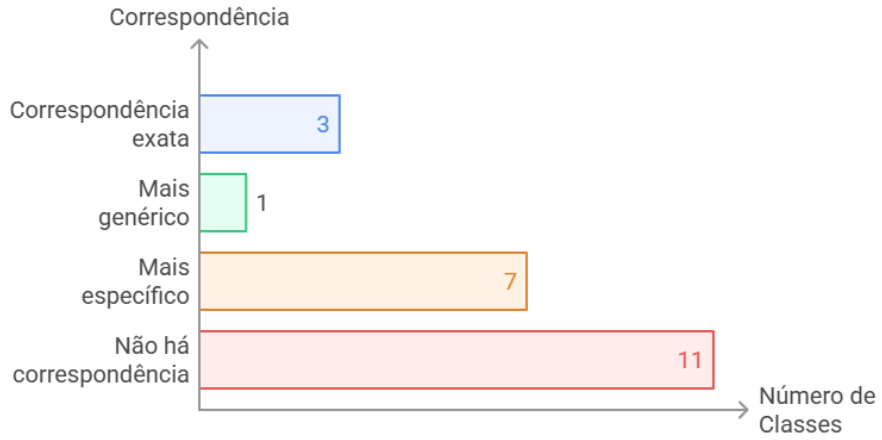
Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
	entrega ou coerência.				
rights	Informações sobre direitos mantidos sobre o recurso.	-	-	Não há correspondência	Um recurso relacionado que é exigido pelo recurso descrito para apoiar sua função, entrega ou coerência.
rightsHolder	Uma pessoa ou organização que possui ou gerencia direitos sobre o recurso.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais específico	*Crosswalking* relativo. Ao titular dos direitos cabe a atribuição da licença associada a um recurso. Assim, podemos dizer que o recurso é atribuído em parte ao titular dos direitos.
source	Um recurso relacionado do qual o recurso descrito é derivado.	wasDerivedFrom	Derivação é uma transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova, ou a construção de uma nova entidade com base em uma entidade pré-existente.	Mais específico	*Crosswalking* relativo. dct:source é definido como um "recurso relacionado do qual o recurso descrito é derivado", o que corresponde à noção de derivação em PROV-DM ("uma transformação de uma entidade em outra"). No entanto, prov:wasDerivedFrom também abrange derivações mais amplas, como "uma atualização de uma entidade resultando em uma nova", que não é coberta por dct:source.
spatial	Características espaciais do recurso.	-	-	Não há correspondência	Características espaciais do recurso.
subject	Um tópico do recurso.	-	-	Não há correspondência	O tópico do recurso.

Dublin Core	**Definição DC[7]**	**PROV**	**Definição PROV[8]**	**Grau de correspondência**	**Observações**
tableOfContents	Uma lista de subunidades do recurso.	-	-	Não há correspondência	Uma lista de subunidades do recurso.
temporal	O tópico espacial ou temporal do recurso, a aplicabilidade espacial do recurso ou a jurisdição sob a qual o recurso é relevante.	-	-	Não há correspondência	Características temporais do recurso.
title	Um nome dado ao recurso.	-	-	Não há correspondência	Um nome dado ao recurso.
type	O formato do arquivo, meio físico ou dimensões do recurso.	-	-	Não há correspondência	A natureza ou gênero do recurso.
valid	Data (geralmente um intervalo) de validade de um recurso.	-	-	Não há correspondência	Data (geralmente um intervalo) de validade de um recurso.

Fonte: Elaborado pelo autor.

De acordo com os dados coletados a partir da realização do *Crosswalk* do padrão Dublin Core para o modelo sugerido da Família PROV, foi observado, nas classes, que dos 22 elementos DC, apenas metade possui equivalência, sendo que apenas 3 elementos são tidos como *Crosswalking* absoluto e 8 como *Crosswalking* relativo. Já o Grau de Equivalência apontado é: 3 Correspondência exata, 1 Mais genérico, 7 Mais específico e 11 Não há correspondência.

Gráfico 2 - Correspondência de Classes do Dublin Core



Fonte: Elaborado pelo autor (2025)

Em resumo:

- ***Crosswalking* Absoluto (Correspondência Exata):** *Agent, Location, ProvenanceStatement.*

- ***Crosswalking* Relativo:**

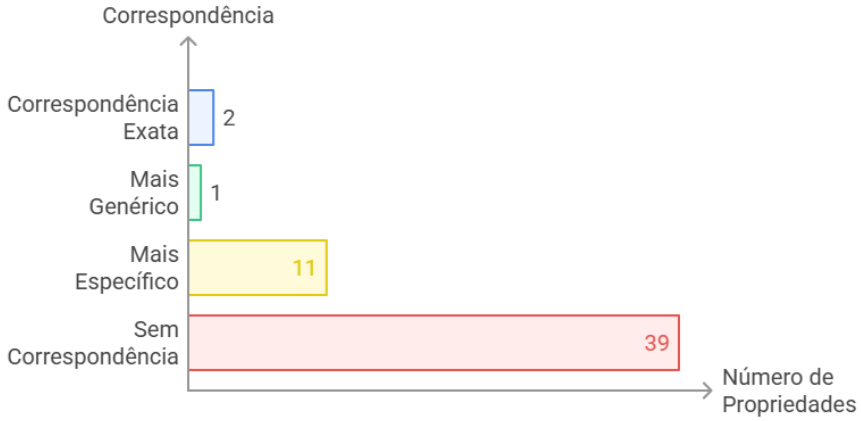
    - **Mais específico:** *BibliographicResource, LicenseDocument, MethodOfAccrual, MethodOfInstruction, PhysicalResource, Policy, RightsStatement.*

    - **Mais genérico:** *LinguisticSystem.*

- **Não há correspondência:** *AgentClass, FileFormat, Frequency, Jurisdiction, LocationPeriodOrJurisdiction, MediaType, MediaTypeOrExtent, PeriodOfTime, PhysicalMedium, SizeOrDuration, Standard.*

Já, nas propriedades, foi utilizado como referência a Documentação do DSpace 7.x, a mais recente até então. Observou-se que, dentro das 55 propriedades adotadas do DC, enumera-se 2 Correspondência exata, 1 Mais genérico e 11 Mais específico e 39 Não há correspondência. Além disso, foi contabilizado 12 *Crosswalking* relativo e 2 *Crosswalking* absoluto.

Gráfico 3 - Correspondência de Propriedades do Dublin Core



Fonte: Elaborado pelo autor (2025)

Em resumo:

- ***Crosswalking* Absoluto (Correspondência Exata):** *creator, issued.*
- ***Crosswalking* Relativo:**
    - **Mais específico:** *created, contributor, dateAccepted, dateCopyrighted, dateSubmitted, modified, publisher, rightsHolder, references, source, isFormatOf.*
    - **Mais genérico:** *hasFormat.*
- **Não há correspondência:** *abstract, accessRights, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, available, bibliographicCitation, conformsTo, coverage, date, description, educationLevel, extent, format, hasPart, hasVersion, isPartOf, isReferencedBy, isReplacedBy, isRequiredBy, isVersionOf, language, license, mediator, medium, provenance, relation, replaces, requires, rights, spatial, subject, tableOfContents, temporal, title, type, valid.*

Com base nos resultados obtidos, pode-se compreender que há um nível proporcional de equivalência entre classes de metadados. A porcentagem de correspondência entre as classes DC é de 50%, pois, das 22 classes, 11 possuem

equivalência. A conversão dos padrões DC para o modelo PROV evidencia a possibilidade de correspondência e compatibilidade com os objetivos do W3C de preencher a lacuna entre as comunidades DC e PROV, fornecendo informações valiosas sobre as diferentes características de ambos os modelos de dados e facilitando a adoção do modelo PROV.

No entanto, no que diz respeito às propriedades, a correspondência é ainda menor, cerca de 29,09%, pois, das 55 propriedades analisadas, apenas 16 apresentaram algum nível de equivalência. Assim, considerando um total de 77 elementos do Dublin Core (22 classes + 55 propriedades), houve uma correspondência geral de aproximadamente 35,06%.

5.2 Mapeamento DATACITE METADATA SCHEMA

No quadro 12 será demonstrado como os elementos de metadados do *DataCite Metadata Schema* podem ser mapeados para os elementos correspondentes no padrão PROV. Esse mapeamento considera tantas correspondências diretas (*crosswalking* absoluto), quanto adaptações (*crosswalking* relativo), refletindo as particularidades de cada padrão.

Para este trabalho, foi utilizada a versão mais recente do *DataCite Metadata Schema*, a 4.6, lançada em dezembro de 2024. Este padrão é amplamente utilizado para a descrição, publicação e citação de dados de pesquisa e outros recursos acadêmicos, garantindo interoperabilidade e rastreabilidade em contextos de dados de pesquisa. Após a apresentação do mapeamento, será realizada uma análise dos resultados obtidos.

Finalmente, é importante ressaltar que as melhorias na versão 4.6 foram impulsionadas por solicitações da comunidade acadêmica e de membros do *DataCite*. O *feedback*, de modo geral, permitiu a expansão e o aprimoramento do esquema para atender a uma gama ainda mais diversificada de casos de uso, consolidando sua posição como uma referência fundamental para a descrição de recursos de pesquisa.

Quadro 13 – *Crosswalk DATACITE METADATA SCHEMA* para PROV *term*

DATACITE METADATA SCHEMA	**Definições DATACITE[9]**	**PROV**	**Definições PROV[10]**	**Grau de correspondência**	**Observações**
Identifier	O identificador é uma sequência única que identifica um recurso.	-	-	Não há correspondência	
identifierType	O tipo de Identificador.	-	-	Não há correspondência	
Creator	Os principais pesquisadores envolvidos na produção dos dados, ou os autores da publicação, em ordem de prioridade. Para instrumentos, este é o fabricante ou desenvolvedor do instrumento. Para fornecer múltiplos criadores, repita esta propriedade.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais genérico que	*Crosswalking* relativo.
creatorName	O nome completo do criador.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais específico	*Crosswalking* relativo.
nameType	O tipo de nome.	-	-	Não há correspondência	

[9] https://datacite-metadata-schema.readthedocs.io/_/downloads/en/4.6/pdf/

[10] https://www.w3.org/TR/2013/REC-prov-o-20130430/#wasDerivedFrom

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
givenName	O nome pessoal ou primeiro nome do criador.	-	-	Não há correspondência	
familyName	O sobrenome ou sobrenome do criador.	-	-	Não há correspondência	
nameIdentifier	Identifica exclusivamente uma pessoa física ou jurídica, de acordo com vários esquemas.	-	-	Não há correspondência	
nameIdentifierScheme	O nome do esquema de identificador de nome.	-	-	Não há correspondência	
schemeURI	O URI do esquema de identificador de nome.	-	-	Não há correspondência	
Affiliation	A afiliação organizacional ou institucional do criador.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais específico	*Crosswalking* relativo.
affiliationIdentifier	Identifica exclusivamente a afiliação organizacional do criador.	-	-	Não há correspondência	
affiliationIdentifierScheme	O nome do esquema de identificação de afiliação.	-	-	Não há correspondência	
SchemeURI	O URI do esquema de identificador	-	-	Não há correspondência	

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
	de afiliação.				
Title	Um nome ou título pelo qual um recurso é conhecido. Pode ser o título de um conjunto de dados ou o nome de um pedaço de software ou instrumento.	-	-	Não há correspondência	
titleType	O tipo de título (diferente do título principal).	-	-	Não há correspondência	
Publisher	O nome da entidade que detém, arquiva, publica, imprime, distribui, libera, emite ou produz o recurso. Esta propriedade será usada para formular a citação, então considere a proeminência da função.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais específico	*Crosswalking* relativo.
publisherIdentifier	Identifica exclusivamente o editor, de acordo com vários esquemas.				
publisherIdentifierScheme	O nome do esquema				

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
	de identificação do editor.				
schemeURI	O URI do esquema de identificador do editor.				
PublicationYear	O ano em que os dados foram ou serão disponibilizados publicamente. No caso de recursos como software ou dados dinâmicos, onde pode haver vários lançamentos em um ano, inclua a propriedade Date e subpropriedades (dateType/dateInformation) para fornecer mais informações sobre os detalhes da data de publicação ou lançamento.	generatedAtTime	Geração é a conclusão da produção de uma nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.	Mais genérico que	*Crosswalking* relativo.
Subject	Subject, keyword, classification code, or key phrase describing the resource.	-	-	Não há correspondência	

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
subjectScheme	O nome do esquema do assunto ou código de classificação ou autoridade, se for usado.	-	-	Não há correspondência	
schemeURI	O URI do esquema de identificador de assunto.	-	-	Não há correspondência	
valueURI	O URI do termo do assunto.	-	-	Não há correspondência	
classificationCode	O código de classificação usado para o termo de assunto no esquema de assunto.	-	-	Não há correspondência	
Contributor	A instituição ou pessoa responsável por coletar, gerenciar, distribuir ou contribuir de outra forma para o desenvolvimento do recurso. Para fornecer múltiplos contribuidores, repita esta propriedade.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais genérico que	*Crosswalking* relativo.
contributorType	O tipo de contribuidor do recurso.	-	-	Não há correspondência	

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
contributorName	O nome completo do colaborador.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais específico	*Crosswalking* relativo.
nameType	O tipo de nome.	-	-	Não há correspondência	
givenName	O nome pessoal ou primeiro nome do colaborador.	-	-	Não há correspondência	
Familyname	O sobrenome ou sobrenome do colaborador.	-	-	Não há correspondência	
nameIdentifier	Identifica exclusivamente uma pessoa física ou jurídica, de acordo com vários esquemas.	-	-	Não há correspondência	
nameIdentifierScheme	O nome do esquema de identificador de nome.	-	-	Não há correspondência	
schemeURI	O URI do esquema de identificador de nome.	-	-	Não há correspondência	
Affiliation	A afiliação organizacional ou institucional do colaborador.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais específico	*Crosswalking* relativo.
affiliationIdentifier	Identifica exclusivamente a afiliação	-	-	Não há correspondência	

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
	organizacional do colaborador.				
affiliationIdentifierScheme	O nome do esquema de identificador de afiliação.	-	-	Não há correspondência	
SchemeURI	URI do esquema de identificador de afiliação.	-	-	Não há correspondência	
Date	Datas diferentes relevantes para o trabalho.	-	-	Não há correspondência	
dateType	O tipo de data.	-	-	Não há correspondência	
Accepted	A data em que o editor aceitou o recurso em seu sistema.	generatedAtTime-	Geração é a conclusão da produção de uma nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.	Correspondência exata	*Crosswalking* relativo.
Available	Os dados em que o editor aceitou o recurso em seu sistema.	-	-	Não há correspondência	
Collected	Os dados específicos e	-	-	Não há correspondência	

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
	documentados em que o recurso recebe o status de protegido por direitos autorais, se aplicável.				
Copyrighted	A data específica e documentada em que o recurso recebe o status de protegido por direitos autorais, se aplicável.	generatedAtTime	Geração é a conclusão da produção de uma nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.	Mais específico	*Crosswalking* relativo.
Created	Os dados específicos e documentados em que o recurso recebe o status de protegido por direitos autorais, se aplicável.	generatedAtTime	Geração é a conclusão da produção de uma nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.	Não há correspondência	*Crosswalking* relativo.
Issued	A data em que o recurso é publicado ou	generatedAtTime	Geração é a conclusão da produção de uma	Não há correspondência	*Crosswalking* relativo.

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
	distribuído, por exemplo, para um centro de dados.		nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.		
Submitted	A data em que o recurso é publicado ou distribuído, por exemplo, para um centro de dados.	generatedAtTime	Geração é a conclusão da produção de uma nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.	Não há correspondência	*Crosswalking* relativo.
Updated	A data em que o recurso é publicado ou distribuído, por exemplo, para um centro de dados.	generatedAtTime	Geração é a conclusão da produção de uma nova entidade por uma atividade. Esta entidade não existia antes da geração e se torna disponível para uso após esta geração.	Não há correspondência	*Crosswalking* relativo.

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
dateType (for StartDate/EndDate)	O tipo de data, que pode incluir datas de início e término.	-	-	Não há correspondência	
dateInformation	Informações específicas sobre a data, se apropriado.	-	-	Não há correspondência	
Language	O idioma principal do recurso.	-	-	Não há correspondência	
resourceType	Uma descrição do recurso.	-	-	Não há correspondência	
resourceTypeGeneral	O tipo geral de um recurso.	-	-	Não há correspondência	
alternateIdentifier	Um identificador diferente do identificador primário aplicado ao recurso sendo registrado.	-	-	Não há correspondência	
alternateIdentifierType	O tipo do identificador alternativo.	-	-	Não há correspondência	
relatedIdentifier	Identificadores de recursos relacionados.	-	-	Não há correspondência	
relatedIdentifierType	O tipo do identificador relacionado.	-	-	Não há correspondência	
relationType	Descrição do relacionamento entre o recurso sendo registrado (A) e o recurso relacionado	-	-	Não há correspondência	

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
	(B).				
isReferencedBy	Indica que A é usado como fonte de informação por B.	-	-	Não há correspondência	
references	Indica que B é usado como fonte de informação para A.	wasDerivedFrom	Derivação é uma transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova, ou a construção de uma nova entidade com base em uma entidade pré-existente.	Correspondência exata	*Crosswalking* relativo.
isVersionOf	Indica que A é uma versão de B.	-	-	Não há correspondência	
hasVersion	Indica que A tem uma versão B.	-	-	Não há correspondência	
isVariantFormatOf	Indica que A é uma variante ou formato diferente de B.	alternateOf	Duas entidades alternativas apresentam aspectos da mesma coisa. Esses aspectos podem ser os mesmos	Mais específico	*Crosswalking* relativo.

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
			ou diferentes, e as entidades alternativas podem ou não se sobrepor no tempo.		
isPartOf	Indica que A é uma parte de B; pode ser usado para elementos de uma série.	-	-	Não há correspondência	
hasPart	Indica que A inclui a parte B.	-	-	Não há correspondência	
isObsoletedBy	Indica que A é substituído por B.	-	-	Não há correspondência	
obsoletes	Indica que A substitui B.	-	-	Não há correspondência	
isDerivedFrom	Indica que B é uma fonte na qual A é baseado.	wasDerivedFrom	Derivação é uma transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova, ou a construção de uma nova entidade com base em uma entidade pré-existente.	Mais específico	*Crosswalking* relativo.

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
Other relation types	Outros tipos de relacionamento que não se encaixam nas categorias existentes.	-	-	Não há correspondência	
relatedMetadataScheme	O nome do esquema.	-	-	Não há correspondência	
schemeURI	O URI do esquema de metadados relacionados.	-	-	Não há correspondência	
schemeType	O tipo do esquema de metadados relacionados, vinculado ao schemeURI.	-	-	Não há correspondência	
resourceTypeGeneral	O tipo geral do recurso relacionado.	-	-	Não há correspondência	
Size	Tamanho (por exemplo, bytes, páginas, polegadas) ou duração (extensão) de um recurso.	-	-	Não há correspondência	
Format	O formato técnico do recurso.	-	-	Não há correspondência	
Version	O número da versão do recurso.	-	-	Não há correspondência	
Rights	Qualquer informação de direitos para este recurso.	-	-	Não há correspondência	

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
rightsURI	O URI da licença.	-	-	Não há correspondência	
rightsIdentifier	Uma versão curta e padronizada do nome da licença.	-	-	Não há correspondência	
rightsIdentifierScheme	O nome do esquema.	-	-	Não há correspondência	
schemeURI	O URI do esquema do identificador de direitos.	-	-	Não há correspondência	
Description	Todas as informações adicionais que não se encaixam em nenhuma das outras categorias.	-	-	Não há correspondência	
descriptionType	O tipo da descrição.	-	-	Não há correspondência	
Abstract	Uma breve descrição do recurso e do contexto em que o recurso foi criado.	-	-	Não há correspondência	
Methods	A metodologia empregada para o estudo ou pesquisa.	-	-	Não há correspondência	
TechnicalInformation	Informações técnicas detalhadas que podem estar associadas ao design, implementação,	-	-	Não há correspondência	

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
	operação, uso e/ou manutenção de um processo, sistema ou instrumento.				
TableOfContents	Uma listagem do índice de conteúdos.	-	-	Não há correspondência	
Other	Outras informações de descrição que não se encaixam em uma categoria existente.	-	-	Não há correspondência	
GeoLocation	Região espacial ou local nomeado onde os dados foram coletados ou sobre os quais os dados são focados.	-	-	Não há correspondência	
geoLocationPoint	Um ponto de localização no espaço.	-	-	Não há correspondência	
pointLongitude	A dimensão longitudinal do ponto.	-	-	Não há correspondência	
pointLatitude	A dimensão latitudinal do ponto.	-	-	Não há correspondência	
geoLocationBox	Os limites espaciais de uma caixa.	-	-	Não há correspondência	
westBoundLongitude	A dimensão longitudinal ocidental da caixa.	-	-	Não há correspondência	

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
eastBoundLongitude	A dimensão longitudinal oriental da caixa.	-	-	Não há correspondência	
southBoundLatitude	A dimensão latitudinal sul da caixa.	-	-	Não há correspondência	
northBoundLatitude	A dimensão latitudinal norte da caixa.	-	-	Não há correspondência	
geoLocationPlace	Descrição de uma localização geográfica.	-	-	Não há correspondência	
geoLocationPolygon	Uma área de polígono desenhada, definida por um conjunto de pontos e linhas conectando os pontos em uma cadeia fechada.	-	-	Não há correspondência	
polygonPoint	Um ponto de localização em um polígono.	-	-	Não há correspondência	
pointLongitude	A dimensão longitudinal do ponto.	-	-	Não há correspondência	
pointLatitude	A dimensão latitudinal do ponto.	-	-	Não há correspondência	
inPolygonPoint	Para qualquer área delimitada que seja maior que metade da Terra, defina um ponto (aleatório) dentro.	-	-	Não há correspondência	

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
pointLongitude	A dimensão longitudinal do ponto.	-	-	Não há correspondência	
pointLatitude	A dimensão latitudinal do ponto.	-	-	Não há correspondência	
fundingReference	Informações sobre o suporte financeiro (financiamento) para o recurso sendo registrado.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais específico	*Crosswalking* relativo.
funderName	Nome do provedor de financiamento.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais específico	*Crosswalking* relativo.
funderIdentifier	Identifica exclusivamente uma entidade	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais específico	*Crosswalking* relativo.
funderIdentifierType	O tipo do identificador do financiador.	-	-	Não há correspondência	
SchemeURI	O URI do esquema do identificador do financiador.	-	-	Não há correspondência	
awardNumber	O código atribuído pelo financiador a um prêmio patrocinado (bolsa).	-	-	Não há correspondência	
awardURI	O URI que leva a uma página fornecida pelo financiador para mais informações	-	-	Não há correspondência	

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
	sobre o prêmio (bolsa).				
awardTitle	O título ou nome legível por humanos do prêmio (bolsa).	-	-	Não há correspondência	
RelatedItem	Informações sobre um recurso relacionado ao que está sendo registrado.	-	-	Não há correspondência	
relationType	Descrição do relacionamento entre o recurso sendo registrado (A) e o recurso relacionado (B).	-	-	Não há correspondência	
isReferencedBy	Indica que A é usado como fonte de informação por B.	-	-	Não há correspondência	
references	Indica que B é usado como fonte de informação para A.	wasDerivedFrom	Derivação é uma transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova, ou a construção de uma nova entidade com base em uma	Mais específico	*Crosswalking* relativo.

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
			entidade pré-existente.		
isVersionOf	Indica que A é uma versão de B.	-	-	Não há correspondência	
hasVersion	Indica que A tem uma versão B.	-	-	Não há correspondência	
isVariantFormatOf	Indica que A é uma variante ou formato diferente de B.	alternateOf	Duas entidades alternativas apresentam aspectos da mesma coisa. Esses aspectos podem ser os mesmos ou diferentes, e as entidades alternativas podem ou não se sobrepor no tempo.	Mais específico	*Crosswalking* relativo.
isPartOf	Indica que A é uma parte de B; pode ser usado para elementos de uma série.	-	-	Não há correspondência	
hasPart	Indica que A inclui a parte B.	-	-	Não há correspondência	
isObsoletedBy	Indica que A é substituído por B.	-	-	Não há correspondência	
obsoletes	Indica que A substitui B.	-	-	Não há correspondência	
isDerivedFrom	Indica que B é uma	wasDerivedFrom	Derivação é uma	Mais específico	*Crosswalking* relativo.

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
	fonte na qual A é baseado.		transformação de uma entidade em outra, uma atualização de uma entidade resultando em uma nova, ou a construção de uma nova entidade com base em uma entidade pré-existente.		
Other relation types	Outros tipos de relacionamento que não se encaixam nas categorias existentes.	-	-	Não há correspondência	
relatedItemType	O tipo geral do item relacionado.	-	-	Não há correspondência	
relatedItemIdentifier	O identificador do item relacionado.	-	-	Não há correspondência	
relatedItemIdentifierType	O tipo do identificador do item relacionado.	-	-	Não há correspondência	
Title	Um nome ou título pelo qual um recurso é conhecido. Pode ser o título de um conjunto de dados, o nome de	-	-	Não há correspondência	

DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
	um software ou de um instrumento.				
Type	O tipo geral do recurso.	-	-	Não há correspondência	
Volume	O volume do item relacionado.	-	-	Não há correspondência	
Issue	O número ou nome da edição do item relacionado.	-	-	Não há correspondência	
Number	O número do recurso dentro do item relacionado, por exemplo, número do relatório ou número do artigo.	-	-	Não há correspondência	
Type	O tipo do número do item relacionado, por exemplo, número do relatório ou número do artigo.	-	-	Não há correspondência	
firstPage	A primeira página do recurso dentro do item relacionado, por exemplo, do capítulo, artigo ou artigo de conferência em anais.	-	-	Não há correspondência	
lastPage	A última página do	-	-	Não há correspondên	

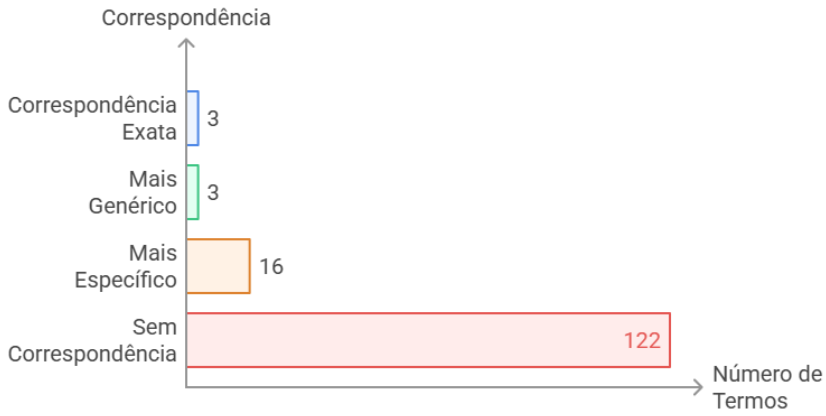
DATACITE METADATA SCHEMA	Definições DATACITE[9]	PROV	Definições PROV[10]	Grau de correspondência	Observações
	recurso dentro do item relacionado, por exemplo, do capítulo, artigo ou artigo de conferência em anais.			cia	
publisher	O nome da entidade que detém, arquiva, publica, imprime, distribui, lança, emite ou produz o recurso.	-	-	Não há correspondência	
edition	A edição ou versão do item relacionado.	-	-	Não há correspondência	
contributor	A instituição ou pessoa responsável por coletar, gerenciar, distribuir ou contribuir de outra forma para o desenvolvimento do recurso.	-	-	Não há correspondência	
contributorType	O tipo de contribuidor do recurso.	-	-	Não há correspondência	
contributorName	O nome completo do contribuidor.	-	-	Não há correspondência	

Fonte: Elaborado pelo autor.

Com base nos dados coletados a partir da realização do *Crosswalk* do padrão *DataCite Metadata Schema* para o modelo sugerido da Família PROV, foi possível

observar que há 144 termos presentes no *DataCite*, e apenas 22 termos possuem equivalência com a PROV, sendo que 19 são *Crosswalking* relativo e 3 *Crosswalking* absoluto, logo, grau de equivalência identificado é: 122 Não há correspondência, 3 Correspondência exata, 3 Mais genérico e 16 Mais específico que.

Gráfico 4 - Correspondência de Propriedades do Dublin Core



Fonte: Elaborado pelo autor (2025)

Em resumo:

- **Crosswalking Absoluto (Correspondência Exata):**
  creator, issued, references.
- **Crosswalking Relativo:**
  - **Mais específico:**
    creatorName, affiliation, publisher, contributor, contributorName, affiliation, accepted, copyrighted, created, submitted, updated, fundingReference, funderName, funderIdentifier, isDerivedFrom, isVariantFormatOf.
  - **Mais genérico:** *Creator, PublicationYear, hasFormat.*
- **Não há correspondência:**
  identifier, identifierType, nameType, givenName, familyName, nameIdentifier, nameIdentifierScheme, schemeURI, title, titleType, subject, subjectScheme, valueURI, classificationCode, date, dateType, available, collected, language, resourceType, resourceTypeGeneral, alternateIdentifier, alternateIdentifierType, relatedIdentifier, relatedIdentifierType, relationType, isReferencedBy, isVersionOf, hasVersion, isPartOf, hasPart, isObsoletedBy, obsoletes, size, format, version, rights, rightsURI, rightsIdentifier, rightsIdentifierScheme, description, des

criptionType, abstract, methods, technicalInformation, tableOfContents, geoLocation, geoLocationPoint, pointLongitude, pointLatitude, geoLocationBox, westBoundLongitude, eastBoundLongitude, southBoundLatitude, northBoundLatitude, geoLocationPlace, geoLocationPolygon, polygonPoint, inPolygonPoint, funderIdentifierType, awardNumber, awardURI, awardTitle, relatedItem, relatedItemType, relatedItemIdentifier, relatedItemIdentifierType, volume, issue, number, firstPage, lastPage, edition, contributorType.

A partir dos resultados obtidos, pode-se compreender que existe um nível razoavelmente baixo de equivalência entre os termos de metadados. A porcentagem de correspondência entre as classes *DataCite* é de 15,28%, já que dos 144 termos, 22 apresentam correspondência, sendo 3 correspondências exatas, 3 mais genéricos e 16 mais específicos. Além disso, 122 termos não possuem correspondência com o modelo PROV. Portanto, a conversão dos padrões *DataCite* para PROV evidencia baixa possibilidade de correspondência com os metadados do W3C.

Dado o baixo nível de correspondência entre os termos de metadados dos padrões *DataCite Metadata Schema* e a Família PROV, é evidente que a integração entre esses dois modelos apresenta desafios significativos. No entanto, pode ser utilizada como uma base para futuras adaptações e melhorias, visando uma maior interoperabilidade entre os padrões de metadados de proveniência.

5.3 Mapeamento DDI - DATA DOCUMENTATION INITIATIVE

Para o mapeamento, fez-se a escolha pelo padrão DDI *Codebook* 2.5, essa escolha se justifica pelo seu amplo uso em repositórios de dados, sendo um padrão consolidado para a documentação de conjuntos de dados simples ou únicos. Como este estudo se concentra na análise de padrões de metadados empregados em repositórios de dados, a utilização do DDI-C é particularmente relevante, pois seus metadados permitem uma descrição detalhada e padronizada dos dados encontrados no contexto dos repositórios de dados de pesquisa.

Para identificar e mapear os elementos de metadados do padrão DDI *Codebook*, foi realizada uma análise detalhada da estrutura XML do padrão, seguindo

as especificações documentadas pela *Data Documentation Initiative*. Cada elemento de metadados foi examinado individualmente, considerando sua descrição, função e relação com os dados descritos. Por exemplo, o elemento *titl*, que representa o título do conjunto de dados, foi identificado como um campo essencial para a documentação, enquanto o elemento *AuthEnty*, que indica a entidade responsável pela autoria dos dados, foi categorizado como um metadado crítico para a atribuição de autoria e responsabilidade.

Sua estrutura baseada em XML oferece flexibilidade para o mapeamento e interoperabilidade, permitindo uma análise mais precisa das correspondências com outros padrões, como a Família PROV, dentro do escopo específico de repositórios de dados.

Quadro 14 – *Crosswalk* DDI - DATA DOCUMENTATION INITIATIVE para PROV term

DATA DOCUMENTATION INITIATIVE	**Definições DDI[11]**	**PROV**	**Definições PROV[12]**	**Grau de correspondência**	**Observações**
abstract	Um resumo não formatado que descreve o propósito, natureza e escopo da coleta de dados, características especiais de seu conteúdo, principais áreas temáticas cobertas e as perguntas que os pesquisadores tentaram responder ao conduzir o estudo.	-	-	Não há correspondência	
accsPlac	Local onde a coleção de dados está	-	-	Não há correspondência	

[11] https://docs.ddialliance.org/DDI-Codebook/2.5/xmlschema/schemas/codebook_xsd/schema-overview.html#a1
[12] https://www.w3.org/TR/2013/REC-prov-o-20130430/#wasDerivedFrom

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	atualmente armazenada.				
actMin	Resumo das ações tomadas para minimizar a perda de dados.	-	-	Não há correspondência	
algorithmSpecification	Especificação do algoritmo utilizado.	-	-	Não há correspondência	
algorithmVersion	Versão do algoritmo utilizado.	-	-	Não há correspondência	
altTitl	Título alternativo pelo qual o trabalho é comumente referido ou uma abreviatura do título.	-	-	Não há correspondência	
anlyInfo	Informações sobre a avaliação dos dados.	-	-	Não há correspondência	
anlysUnit	Informações sobre quem ou o que a variável/nCube descreve.	-	-	Não há correspondência	
anlyUnit	Unidade básica de análise ou observação que o arquivo descreve: indivíduos, famílias/domicílios, grupos, instituições/organizações, unidades administrativas, etc.	-	-	Não há correspondência	
attribute	Identifica um atributo dentro dos elementos identificados pelo seletor ou	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	specificElements em que o vocabulário controlado é usado.				
AuthEnty	Pessoa, entidade corporativa ou agência responsável pelo conteúdo substantivo e intelectual do trabalho.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais genérico que	*Crosswalking* relativo.
authorizationStatement	Texto da autorização.	-	-	Não há correspondência	
authorizingAgency	Nome da agência ou agente que autorizou o estudo.	-	-	Não há correspondência	
avlStatus	Declaração sobre a disponibilidade da coleção.	-	-	Não há correspondência	
backward	Contém uma referência aos IDs de possíveis perguntas anteriores.	-	-	Não há correspondência	
biblCit	Referência bibliográfica completa contendo todos os elementos padrão de uma citação que pode ser usada para citar o trabalho.	-	-	Não há correspondência	
boundPoly	Permite a criação de múltiplos polígonos para descrever de maneira mais detalhada a área	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	geográfica coberta pelo conjunto de dados.				
caseQnty	Número de casos ou observações.	-	-	Não há correspondência	
catgry	Descrição de uma resposta específica.	-	-	Não há correspondência	
catgryGrp	Descrição de categorias de resposta que podem ser agrupadas.	-	-	Não há correspondência	
catLevel	Usado para descrever os níveis da hierarquia de categorias.	-	-	Não há correspondência	
catStat	Pode incluir frequências, porcentagens ou resultados de tabulação cruzada.	-	-	Não há correspondência	
catValu	A resposta explícita.	-	-	Não há correspondência	
citation	Codifica as informações bibliográficas para o trabalho no nível especificado.	-	-	Não há correspondência	
citReq	Texto da exigência de que uma coleção de dados seja citada corretamente em artigos ou outras publicações baseadas na análise dos dados.	-	-	Não há correspondência	
cleanOps	Métodos usados para "limpar" a coleção de	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	dados, como verificação de consistência, verificação de códigos inválidos, etc.				
codeBook	Descrição do codebook.	-	-	Não há correspondência	
codeListAgencyName	Nome da agência que mantém a lista de códigos.	-	-	Não há correspondência	
codeListID	Identifica a lista de códigos da qual o valor é tirado.	-	-	Não há correspondência	
codeListName	Identifica a lista de códigos da qual o valor é tirado com um nome legível por humanos.	-	-	Não há correspondência	
codeListSchemeURN	Identifica o esquema da lista de códigos usando um URN.	-	-	Não há correspondência	
codeListURN	Identifica a lista de códigos da qual o valor é tirado com um URN.	-	-	Não há correspondência	
codeListVersionID	Versão da lista de códigos.	-	-	Não há correspondência	
codingInstructions	Descreve instruções específicas de codificação usadas no processamento, limpeza, avaliação ou tabulação de dados.	-	-	Não há correspondência	
codInstr	Instruções especiais para aqueles que	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	converteram informações de uma forma para outra para uma variável específica.				
cohort	Usado quando o nCube contém um número limitado de categorias de uma variável específica, em oposição à gama completa de categorias.	-	-	Não há correspondência	
collDate	Contém a(s) data(s) em que os dados foram coletados.	-	-	Não há correspondência	
collectorTraining	Descreve o treinamento fornecido aos coletores de dados, incluindo treinamento de entrevistadores, testes de processo, conformidade com padrões, etc.	-	-	Não há correspondência	
collMode	Método usado para coletar os dados; características da instrumentação.	-	-	Não há correspondência	
collSitu	Descrição de aspectos notáveis da situação de coleta de dados.	-	-	Não há correspondência	
collSize	Resume o	-	-	Não há	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	número de arquivos físicos que existem em uma coleção, registrando o número de arquivos que contêm dados e observando se a coleção contém documentação legível por máquina e/ou outros arquivos e informações suplementares, como dicionários de dados, declarações de definição de dados ou instrumentos de coleta de dados.			correspondência	
colspec	Especificação de coluna.	-	-	Não há correspondência	
command	Fornece o código de comando para a instrução de codificação.	-	-	Não há correspondência	
complete	Indica a relação dos dados coletados com a quantidade de dados codificados e armazenados na coleção de dados.	-	-	Não há correspondência	
complianceDescription	Descrição da conformidade com padrões.	-	-	Não há correspondência	
concept	O assunto geral ao qual o elemento pai pode ser visto	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	como pertencente.				
conditions	Indica qualquer informação adicional que ajudará o usuário a entender as condições de acesso e uso da coleção de dados.	-	-	Não há correspondência	
confDec	Usado para determinar se a assinatura de uma declaração de confidencialidade é necessária para acessar um recurso.	-	-	Não há correspondência	
ConOps	Métodos para facilitar o controle de dados realizados pelo investigador principal ou pelo arquivo de dados.	-	-	Não há correspondência	
contact	Nomes e endereços dos indivíduos responsáveis pelo trabalho.	-	-	Não há correspondência	
controlledVocabUsed	Fornece um valor de código, bem como uma referência à lista de códigos da qual o valor é tirado.	-	-	Não há correspondência	
copyright	Declaração de direitos autorais para o trabalho no nível	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	apropriado.				
CubeCoord	Elemento vazio contendo apenas os atributos listados.	-	-	Não há correspondência	
custodian	Identifica a agência ou indivíduo responsável por criar ou manter o quadro de amostragem.	-	-	Não há correspondência	
dataAccs	Descreve as condições de acesso e termos de uso para a coleção de dados.	-	-	Não há correspondência	
dataAppr	Outras questões relacionadas à avaliação de dados.	-	-	Não há correspondência	
dataChck	Indica, no nível do arquivo, os tipos de verificações e operações realizadas no arquivo de dados.	-	-	Não há correspondência	
dataColl	Informações sobre a metodologia empregada em uma coleta de dados.	-	-	Não há correspondência	
dataCollector	Entidade (indivíduo, agência ou instituição) responsável por administrar o questionário ou entrevista ou compilar os dados.	-	-	Não há correspondência	
dataDscr	Descrição das variáveis.	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
dataFingerprint	Permite atribuir um valor de hash (impressão digital) aos dados ou ao arquivo de dados.	-	-	Não há correspondência	
dataItem	Identifica um local de armazenamento físico para uma entrada de dados individual, servindo como um link entre o local físico e a descrição do conteúdo lógico de cada item de dados.	-	-	Não há correspondência	
dataKind	O tipo de dados incluídos no arquivo: dados de pesquisa, dados de censo/enumeração, dados agregados, dados clínicos, dados de evento/transação, código-fonte de programa, texto legível por máquina, dados de registros administrativos, dados experimentais, teste psicológico, dados textuais, documentos codificados, diários de orçamento de tempo, dados	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	de observação/classificações, dados produzidos por processos, etc.				
dataMsng	Informações gerais sobre dados ausentes.	-	-	Não há correspondência	
dataProcessing	Descreve vários procedimentos de processamento de dados não capturados em outro lugar na documentação, como topcoding, recodificação, supressão, tabulação, etc.	-	-	Não há correspondência	
dataSrc	Lista os livros, artigos, séries e/ou arquivos de dados legíveis por máquina que serviram como fontes da coleção de dados.	-	-	Não há correspondência	
defntn	Razão pela qual o grupo foi constituído dessa maneira.	-	-	Não há correspondência	
depDate	Data em que o trabalho foi depositado no arquivo que originalmente o recebeu.	-	-	Não há correspondência	
depositr	Nome da pessoa (ou instituição) que forneceu este trabalho ao	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	arquivo que o armazena.				
deposReq	Informações sobre a responsabilidade do usuário em informar os arquivos sobre o uso dos dados, fornecendo citações ao trabalho publicado ou fornecendo cópias dos manuscritos.	-	-	Não há correspondência	
derivation	Usado apenas no caso de uma variável derivada, fornece uma descrição de como a derivação foi realizada e o comando usado para gerar a variável derivada, bem como uma especificação das outras variáveis no estudo usadas para gerar a derivação.	-	-	Não há correspondência	
description	Descrição de uma atividade de desenvolvimento.	-	-	Não há correspondência	
developmentActivity	Descreve o processo de desenvolvimento do estudo como uma série de atividades de desenvolvimento.	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
deviat	Informações indicando correspondências e discrepâncias entre as unidades amostradas (obtidas) e estatísticas disponíveis para a população (idade, proporção de sexo, estado civil, etc.) como um todo.	-	-	Não há correspondência	
digitalFingerprintValue	Valor da impressão digital digital.	-	-	Não há correspondência	
dimensns	Dimensões do arquivo geral.	-	-	Não há correspondência	
disclaimer	Informações sobre a responsabilidade pelo uso da coleção de dados.	-	-	Não há correspondência	
distDate	Data em que o trabalho foi disponibilizado para distribuição/apresentação.	-	-	Não há correspondência	
distrbtr	Organização designada pelo autor ou produtor para gerar cópias do trabalho específico, incluindo quaisquer edições ou revisões necessárias.	wasAttributedTo	*Attribution* é a atribuição de uma entidade a um agente.	Mais genérico	*Crosswalking* relativo.
distStmt	Declaração de distribuição para o trabalho no nível	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	apropriado.				
div	Elemento de formatação: marca uma subdivisão em um texto.	-	-	Não há correspondência	
dmns	Define uma variável como uma dimensão do nCube e deve ser repetido para descrever cada uma das dimensões do cubo.	-	-	Não há correspondência	
docDscr	Consiste em informações bibliográficas que descrevem o documento DDI-compliant como um todo.	-	-	Não há correspondência	
docSrc	Citação para o documento de origem.	-	-	Não há correspondência	
docStatus	Indica se a documentação está sendo apresentada/distribuída antes de ser finalizada.	-	-	Não há correspondência	
drvcmd	O comando real usado para gerar a variável derivada.	-	-	Não há correspondência	
drvdesc	Descrição textual da maneira como esta variável foi derivada.	-	-	Não há correspondência	
eastBL	A coordenada mais a leste que delimita a extensão geográfica do conjunto de dados.	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
embargo	Fornece informações sobre variáveis/nCubes que não estão atualmente disponíveis devido a políticas estabelecidas pelos pesquisadores principais e/ou produtores de dados.	-	-	Não há correspondência	
emph	Elemento de formatação: marca palavras ou frases que são enfatizadas para efeito retórico.	-	-	Não há correspondência	
entry	Entrada de tabela.	-	-	Não há correspondência	
EstSmpErr	Medida de quão precisamente se pode estimar um valor populacional a partir de uma determinada amostra.	-	-	Não há correspondência	
evaluationProcess	Descreve o processo de avaliação seguido.	-	-	Não há correspondência	
evaluator	Identifica pessoas ou organizações envolvidas na avaliação.	-	-	Não há correspondência	
exPostEvaluation	Descreve procedimentos de avaliação não abordados nos processos de avaliação	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	de dados.				
ExtLink	Permite que os codificadores forneçam links de qualquer elemento arbitrário contendo ExtLink como um subelemento para recursos eletrônicos fora do codebook.	-	-	Não há correspondência	
fileCitation	Fornece uma opção de citação bibliográfica completa para cada arquivo de dados descrito em fileDscr.	-	-	Não há correspondência	
fileCont	Resumo ou descrição do arquivo.	-	-	Não há correspondência	
fileDscr	Informações sobre o(s) arquivo(s) de dados que compõem uma coleção.	-	-	Não há correspondência	
fileName	Contém um título curto que será usado para distinguir um arquivo/parte específico de outros arquivos/partes na coleção de dados.	-	-	Não há correspondência	
filePlac	Indica se o arquivo foi produzido em um arquivo ou produzido em outro lugar.	-	-	Não há correspondência	
fileQnty	Número total	-	-	Não há	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	de arquivos físicos associados a uma coleção.			correspondência	
fileStrc	Tipo de estrutura de arquivo.	-	-	Não há correspondência	
fileTxt	Fornece informações descritivas sobre o arquivo de dados.	-	-	Não há correspondência	
fileType	Tipos de arquivos de dados incluem dados brutos (ASCII, EBCDIC, etc.) e arquivos dependentes de software, como conjuntos de dados SAS, arquivos de exportação SPSS, etc	-	-	Não há correspondência	
format	Formato físico do arquivo de dados: formato de comprimento de registro lógico, formato de imagem de cartão (ou seja, dados com múltiplos registros por caso), formato delimitado, formato livre, etc.	-	-	Não há correspondência	
forward	Contém uma referência aos IDs de possíveis perguntas seguintes.	-	-	Não há correspondência	
frameUnit	Fornece	-	-	Não há	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	informações sobre a unidade do quadro de amostragem.			correspondência	
frequenc	Frequência com que os dados foram coletados.	-	-	Não há correspondência	
fundAg	Fonte(s) de financiamento para a produção do trabalho.	-	-	Não há correspondência	
geoBndBox	Descrição geométrica fundamental para qualquer conjunto de dados que modela geografia.	-	-	Não há correspondência	
geogCover	Informações sobre a cobertura geográfica dos dados.	-	-	Não há correspondência	
geogUnit	Nível mais baixo de agregação geográfica coberto pelos dados.	-	-	Não há correspondência	
geoMap	Usado para apontar, usando um atributo "URI", para um mapa externo que exibe a geografia em questão.	-	-	Não há correspondência	
grantNo	Número do contrato/bolsa do projeto que patrocinou o esforço.	-	-	Não há correspondência	
gringLat	Latitude (coordenada y) de um ponto.	-	-	Não há correspondência	
gringLon	Longitude	-	-	Não há	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	(coordenada x) de um ponto.			correspondência	
guide	Lista de termos e definições usados na documentação.	-	-	Não há correspondência	
head	Elemento de formatação: marca um cabeçalho para uma divisão, lista, etc.	-	-	Não há correspondência	
hi	Elemento de formatação: marca uma palavra ou frase como graficamente distinta do texto circundante, sem fazer nenhuma afirmação sobre as razões.	-	-	Não há correspondência	
holdings	Informações sobre os acervos físicos ou eletrônicos do trabalho citado.	-	-	Não há correspondência	
IDNo	String ou número único (número do produtor ou do arquivo).	-	-	Não há correspondência	
imputation	Processo pelo qual se estimam valores ausentes para itens que um entrevistado não forneceu.	-	-	Não há correspondência	
instrumentDevelopment	Descreve qualquer trabalho de	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	desenvolvimento no instrumento de coleta de dados.				
invalrng	Valores para uma variável específica que representam dados ausentes, respostas não aplicáveis, etc.	-	-	Não há correspondência	
item	Contraparte de Range; usado para codificar valores individuais.	-	-	Não há correspondência	
itm	Elemento de formatação: marca entradas (itens) em uma lista.	-	-	Não há correspondência	
ivuInstr	Instruções específicas para o indivíduo que conduz uma entrevista.	-	-	Não há correspondência	
key	Permite uma listagem dos valores e rótulos das categorias.	-	-	Não há correspondência	
keyword	Palavras ou frases que descrevem aspectos salientes do conteúdo de uma coleção de dados.	-	-	Não há correspondência	
label	Elemento de formatação: contém o rótulo associado a um item em uma lista; em glossários,	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	marca o termo sendo definido.				
labl	Uma breve descrição do elemento pai.	-	-	Não há correspondência	
Link	Permite que os codificadores forneçam links de qualquer elemento arbitrário contendo Link como um subelemento para outros elementos no codebook.	-	-	Não há correspondência	
list	Elemento de formatação: contém qualquer sequência de itens (entradas) organizados como uma lista.	-	-	Não há correspondência	
location	Elemento vazio contendo apenas os atributos listados.	-	-	Não há correspondência	
locMap	Mapeia entradas de dados individuais para um ou mais locais de armazenamento físico.	-	-	Não há correspondência	
logRecL	Comprimento do registro lógico, ou seja, número de caracteres de dados no registro.	-	-	Não há correspondência	
measure	Indica as características de medição do	-	-	Não há correspondência	

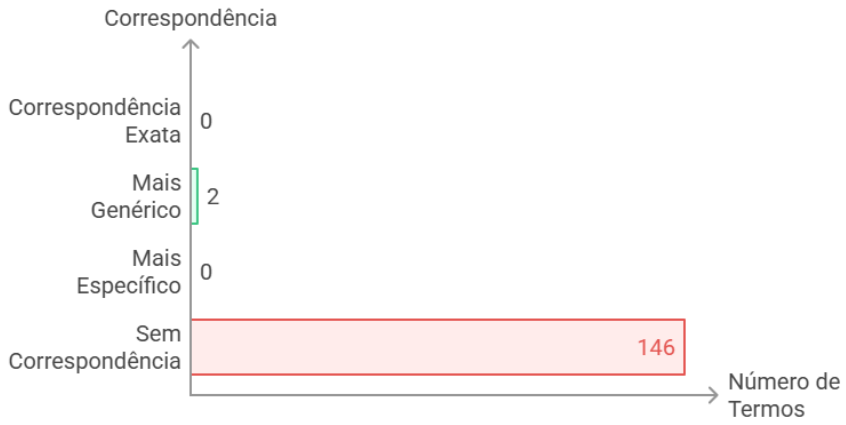
DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	conteúdo da célula: tipo de agregação usada, unidade de medida e escala de medida.				
method	Descreve a metodologia e o processamento envolvidos em uma coleta de dados.	-	-	Não há correspondência	
mi	Contém a menor unidade no mrow que carrega significado.	-	-	Não há correspondência	
mrow	Contém a expressão de apresentação mi.	-	-	Não há correspondência	
nation	Indica o(s) país(es) coberto(s) no arquivo.	-	-	Não há correspondência	
nCube	Descreve a estrutura lógica de um array n-dimensional, em que cada coordenada se cruza com todas as outras dimensões em um único ponto.	-	-	Não há correspondência	
nCubeGrp	Um grupo de nCubes que pode compartilhar um assunto comum, surgir da interpretação de uma única pergunta ou	-	-	Não há correspondência	

DATA DOCUMENTATION INITIATIVE	Definições DDI[11]	PROV	Definições PROV[12]	Grau de correspondência	Observações
	ser vinculado por algum outro fator.				
northBL	A coordenada mais ao norte que delimita a extensão geográfica do conjunto de dados.	-	-	Não há correspondência	
notes	Para informações de esclarecimento /anotação sobre o elemento pai.	-	-	Não há correspondência	
origArch	Arquivo de onde a coleção de dados foi obtida; o arquivo de origem.	-	-	Não há correspondência	
otherMat	Permite a inclusão de outros materiais relacionados ao estudo, conforme identificado e rotulado pelos usuários do DTD/Schema (codificadores) .	-	-	Não há correspondência	

Fonte: Elaborado pelo autor.

Com base nos dados coletados a partir da realização do *Crosswalk* entre o padrão DDI *Codebook* e o modelo sugerido da Família PROV, foi possível identificar que o DDI *Codebook* possui 148 termos de metadados. Desses, apenas 2 apresentaram algum grau de correspondência com os termos da PROV, sendo 2 termos Mais genérico. Dessa forma, o grau de equivalência entre os dois modelos reflete um cenário de baixíssima correspondência geral.

Gráfico 5 - Correspondência de Propriedades do Dublin Core



Fonte: Elaborado pelo autor (2025)

Em resumo:

- **Crosswalking relativo:**

  Mais genérico: *AuthEnty, Distrbtr.*

- **Não há correspondência:**

  abstract, accsPlac, actMin, algorithmSpecification, algorithmVersion, altTitl, anlyInfo, anlysUnit, anlyUnit, attribute, authorizationStatement, authorizingAgency, avlStatus, backward, biblCit, boundPoly, caseQnty, catgry, catgryGrp, catLevel, catStat, catValu, citation, citReq, cleanOps, codeBook, codeListAgencyName, codeListID, codeListName, codeListSchemeURN, codeListURN, codeListVersionID, codingInstructions, codInstr, cohort, collDate, collectorTraining, collMode, collSitu, collSize, colspec, command, complete, complianceDescription, concept, conditions, confDec, ConOps, contact, controlledVocabUsed, copyright, CubeCoord, custodian, dataAccs, dataAppr, dataChck, dataColl, dataCollector, dataDscr, dataFingerprint, dataItem, dataKind, dataMsng, dataProcessing, dataSrc, defntn, depDate, depositr, deposReq, derivation, description, developmentActivity, deviat, digitalFingerprintValue, dimensns, disclaimer, distDate, distStmt, div, dmns, docDscr, docSrc, docStatus, drvcmd, drvdesc, eastBL, embargo, emph, entry, EstSmpErr,

evaluationProcess, evaluator, exPostEvaluation, ExtLink, fileCitation, fileCont, fileDscr, fileName, filePlac, fileQnty, fileStrc, fileTxt, fileType, format, forward, frameUnit, frequenc, fundAg, geoBndBox, geogCover, geogUnit, geoMap, grantNo, gringLat, gringLon, guide, head, hi, holdings, IDNo, imputation, instrumentDevelopment, invalrng, item, itm, ivuInstr, key, keyword, label, labl, Link, list, location, locMap, logRecL, measure, method, mi, mrow, nation, nCube, nCubeGrp, northBL, notes, origArch, otherMat.

A correspondência entre os termos do DDI *Codebook* e a Família PROV é de aproximadamente 1,35%, considerando que, dos 148 termos analisados, apenas 2 apresentaram correspondência. Ambos os termos identificados possuem uma relação de maior especificidade em um *crosswalk* relativo. Apesar dessa conexão, a integração entre o DDI *Codebook* e a Família PROV continua sendo um desafio, visto que a grande maioria dos termos, 146 de 148, não possuem correspondência exata ou especifica com os termos da PROV.

Os resultados obtidos indicam que o padrão DDI *Codebook*, embora seja amplamente utilizado para descrever dados nas ciências sociais e comportamentais, apresenta limitações de interoperabilidade quando comparado aos objetivos e estrutura da Família PROV. O termo identificado como correspondência exata representa um ponto de convergência entre os modelos, enquanto os termos mais genéricos que evidenciam uma relação mais relativa e dependente de interpretações e ajustes no mapeamento.

Esses resultados sugerem que, apesar de sua utilidade para descrever e gerenciar dados, o DDI *Codebook* requer adaptações e extensões para alcançar maior compatibilidade com modelos que seguem os princípios do W3C, como a Família PROV.

6 BOAS PRÁTICAS DE METADADOS DE PROVENIÊNCIA EM REPOSITÓRIOS DE DADOS DE PESQUISA

A gestão eficiente de dados de pesquisa é indispensável para o avanço científico e a promoção da inovação, justamente por promover a transparência informacional. A documentação de boas práticas[13], "*Data on the Web Best Practices* W3C", proposta pela W3C no início de 2017, por meio de um grupo de especialistas, disponibilizou um documento com 14 tópicos que estabelece um conjunto de boas práticas a serem seguidas na estruturação da oferta e do consumo de dados.

Trabalhos como de Torino e Vidotti (2021), destacam as boas práticas para dados na web, fortalecendo a boa prática no uso de padrões metadados e incentivando a comunidade a adotar técnicas que assegurem a qualidade e a interoperabilidade dos dados.

Além disso, é importante entender como a proveniência dos dados contribui para a confiabilidade e a transparência das informações. A documentação apresentada pela W3C se preocupa com a proveniência dos dados no tópico "8.4 *Data Provenance*", e diz que a proveniência é uma das formas que os consumidores dispõem para julgar a qualidade de um conjunto de dados. Ela também pontua que, ao entender a origem e história de um dado, torna-se um auxílio a determinar se o dado é confiável e fornece um contexto para interpretação dos dados.

Já, contudo, no sumário de boas práticas apresentado pela W3C, a proveniência é tratada no item 5 "*Provide data provenance information*", e indicam a necessidade do conhecimento do conjunto de dados pelos usuários, enquanto os agentes de software poderão processar automaticamente as informações de proveniência.

Conforme as orientações recomendadas, no "*Data on the Web Best Practices* W3C" para a publicação de dados, é imprescindível assegurar que os metadados associados ao conjunto de dados "[...] incluam as informações de proveniência sobre o conjunto de dados em um formato legível por humanos. Através de um aplicativo de computador que pode processar automaticamente as informações de proveniência sobre o conjunto de dados." (LÓSCIO; BURLE; CALEGARI, 2017, não paginado, tradução nossa). Os efeitos positivos dessa prática incluem a reutilização dos dados,

[13] https://www.w3.org/Translations/DWBP-pt-BR/

uma melhor compreensão de seu contexto e maior confiança nas informações fornecidas.

Como exemplo ilustrativo, o quadro 14 apresenta um modelo de forma a descrever e estruturar metadados de proveniência, destacando elementos essenciais como a origem dos dados e agentes envolvidos.

Quadro 15 – Exemplo de informações de proveniência

```
:stops-2015-05-05
    a dcat:Dataset, prov:Entity ;
    dct:title "Bus stops of MyCity" ;
    dcat:keyword "transport", "mobility", "bus" ;
    dct:issued "2015-05-05"^^xsd:date ;
    dcat:contactPoint <http://data.mycity.example.com/transport/contact> ;
    dct:temporal <http://reference.data.gov.uk/id/year/2015> ;
    dct:spatial <http://sws.geonames.org/3399415> ;
    dct:publisher :transport-agency-mycity ;
    dct:accrualPeriodicity <http://purl.org/linked-data/sdmx/2009/code#freq-A> ;
    dct:language <http://id.loc.gov/vocabulary/iso639-1/en> ;
    dct:creator :adrian
    .

:adrian
    a foaf:Person, prov:Agent ;
    foaf:givenName "Adrian" ;
    foaf:mbox <mailto:adrian@mycitytransport.org> ;
    prov:actedOnBehalfOf :transport-agency-mycity
    .
:transport-agency-mycity
    a foaf:Organization, prov:Agent ;
    foaf:name "Transport Agency of Mycity"
    .
```

Fonte: *World Wide Web Consortium* (2017, não paginado)

O modelo ilustra metadados legíveis por máquinas para o conjunto de dados das paradas de ônibus, incluindo informações de proveniência. As propriedades *dct:creator, dct:publisher* e *dct:issued* são utilizadas para descrever a origem do conjunto de dados. A propriedade *prov:actedOnBehalfOf* indica que Adrian atuou em nome do Departamento de Transportes *MyCity.*

No quadro 15, apresenta-se o padrão de metadados *Dublin Core* em um dos registros que diz respeito sobre ‘Metadados de espécimes do museu do Reino Unido e distância de procrustes entre as asas anteriores esquerda e direita, para cinco

espécies de abelhas, do repositório *Environmental Information Data Centre'* [14] *(EIDC)* que contempla dados de pesquisa sobre ecologia e hidrologia.

Quadro 16 – Padrão de metadado *Dublin Core*

```
dct:title "UK museum specimen metadata and procrustes distance between left and
right forewings, for five bumblebee species" ;
dct:identifier "https://catalogue.ceh.ac.uk/id/2696535e-564a-4c6a-877e-
515996fa97a1","https://doi.org/10.5285/2696535e-564a-4c6a-877e-515996fa97a1" ;
dct:publisher
<https://ror.org/04xw4m193> ;
dct:subject
<http://onto.nerc.ac.uk/CEHMD/topic/4>,"Bumblebee","Museums","Specimen","Win
g shape","Fluctuating
asymmetry","Procustes","Distance","Stress","Centuary","UK","insect declines" ;
dct:type dcmitype:Dataset ;
dct:available "2022-07-25"^^xsd:date ;
dcat:landingPage <https://catalogue.ceh.ac.uk/id/2696535e-564a-4c6a-877e-
515996fa97a1>, <https://doi.org/10.5285/2696535e-564a-4c6a-877e-515996fa97a1> ;
dct:license <https://spdx.org/licenses/OGL-UK-3.0.ttl>
;
```

Fonte: Extraído e Adaptado de (ARCE et al. 2022)
Disponível em: https://catalogue.ceh.ac.uk/documents/2696535e-564a-4c6a-877e-515996fa97a1
Acesso em: 18 jan. 2025.

O registro do quadro 15 apresentado, descreve informações sobre os dados de uma pesquisa, permitindo identificar a presença de metadados como: *dct:title*, que corresponde ao título; *dct:identifier*, que oferece a descrição sobre o conteúdo; *dct:publisher*, relacionado às especificações de publicação; e *dct:type*, que define a natureza do recuso; *dct:available*, demostra a data da publicação e *dct:license*, metadado que concede permissão para fazer algo em relação a um recurso

Além disso, os metadados estão organizados de acordo com o modelo de dados *Resource Description Framework* (RDF) e são representados utilizando a sintaxe *Turtle* (*Terse* RDF *Triple Language*).

Para garantir a proveniência dos dados e melhorar a descrição do registro utilizando elementos do modelo PROV, é possível identificar e recomendar a inclusão de elementos que não foram contemplados no registro atual, para atender a interoperabilidade semântica entre o modelo DC para PROV, que há compatibilidade, ficaria da seguinte forma:

[14] https://eidc.ac.uk/

Quadro 17 – Sugestão de adequação do registro em *Dublin Core*

```
@prefix dct: <http://purl.org/dc/terms/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<https://catalogue.ceh.ac.uk/id/2696535e-564a-4c6a-877e-515996fa97a1>
   dct:title "UK museum specimen metadata and procrustes distance between left
and right forewings, for five bumblebee species" ;
   dct:identifier "https://catalogue.ceh.ac.uk/id/2696535e-564a-4c6a-877e-
515996fa97a1", "https://doi.org/10.5285/2696535e-564a-4c6a-877e-515996fa97a1" ;
   dct:publisher <https://ror.org/04xw4m193> ;
   dct:subject <http://onto.nerc.ac.uk/CEHMD/topic/4>, "Bumblebee", "Museums",
"Specimen", "Wing shape", "Fluctuating asymmetry", "Procustes", "Distance",
"Stress", "Centuary", "UK", "insect declines" ;
   dct:type dcmitype:Dataset ;
   dct:available "2022-07-25"^^xsd:date ;
   dcat:landingPage <https://catalogue.ceh.ac.uk/id/2696535e-564a-4c6a-877e-
515996fa97a1>, <https://doi.org/10.5285/2696535e-564a-4c6a-877e-515996fa97a1> ;
   dct:license <https://spdx.org/licenses/OGL-UK-3.0.ttl> ;
   prov:wasGeneratedBy <https://example.org/activity/wing-measurement> ;
   prov:wasDerivedFrom <https://example.org/museum-collection/12345> ;
   prov:wasAttributedTo <https://ror.org/04xw4m193> .

<https://example.org/activity/wing-measurement>
   prov:used <https://example.org/tool/morphometric-software> ;
   prov:startedAtTime "2021-01-01T00:00:00Z"^^xsd:dateTime ;
   prov:endedAtTime "2021-12-31T23:59:59Z"^^xsd:dateTime ;
   prov:wasInformedBy <https://example.org/previous-study/67890> .
```

Fonte: Elaborado pelo autor (2025)

A inclusão dos elementos PROV, destacado em cinza no registro, permite uma descrição mais completa e robusta da proveniência dos dados, atendendo a requisitos de rastreabilidade, transparência e interoperabilidade. Alguns exemplos concretos que pode ser observado no quadro 17 é:

- *prov:wasGeneratedBy:* Descrever o que gerou a entidade.
- *prov:generatedAtTime:* Momento ou hora em que uma atividade teve início.
- *prov:wasDerivedFrom:* Descreve a entidade preexistente da qual uma derivação foi originada.

Os elementos como *prov:wasGeneratedBy, prov:generatedAtTime* e *prov:wasDerivedFrom* permitem uma descrição detalhada das entidades, atividades e agentes envolvidos no ciclo de vida dos dados, garantindo que a origem e as transformações sofridas sejam claramente documentadas.

O padrão de metadados *Datacite Metadata Schema*, contém conjuntos de elementos de metadados que é empregado para descrever, publicar e referenciar dados de pesquisa e outros materiais acadêmicos. Seu objetivo principal é facilitar a citação e a descoberta de dados, complementando, e não substituindo, os esquemas de metadados específicos de cada área do conhecimento. Dessa forma, funciona como uma base mínima que pode ser ampliada com metadados mais detalhados e voltados para disciplinas específicas.

No quadro 17, é possível ver a estrutura de descrição utilizada no padrão *Datacite*, que foi retirado do repositório *4TU.ResearchData*[15], esse registro diz respeito a dados do sensor de tornozelo de pacientes com doença de Parkinson em condições de vida semi-livres para detecção de congelamento da marcha.

Quadro 18 – Padrão de metadado *Datacite Metadata Schema*

```
<resource xmlns="http://datacite.org/schema/kernel-4"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://datacite.org/schema/kernel-4
http://schema.datacite.org/meta/kernel-4.4/metadata.xsd">
<identifier identifierType="DOI">10.4121/40e06061-f441-43b5-9235-
006829206509.v1</identifier>
    <creators>
      <creator>
        <creatorName nameType="Personal">Juan Daniel Delgado-
Terán</creatorName>
        <givenName>Juan Daniel</givenName>
        <familyName>Delgado-Terán</familyName>
        <nameIdentifier nameIdentifierScheme="https://orcid.org/">0000-0001-9072-
0404</nameIdentifier>
        </creator>
    </creators>
<titles>
     <title>Ankle Sensor Data from Parkinson's Disease Patients in Semi-Free Living
Conditions for Freezing of Gait Detection</title>
</titles>
<publisher>4TU.ResearchData</publisher>
<publicationYear>2025</publicationYear>
<resourceType resourceTypeGeneral="Dataset">Dataset</resourceType>
```

Fonte: Extraído e Adaptado de (DELGADO-TERÁN, 2025)
Disponível em: https://data.4tu.nl/datasets/40e06061-f441-43b5-9235-006829206509/1
Acesso em: 19 jan. 2025.

Conforme o quadro 18, é possível identificar os padrões *Datacite Metadata Schema*, como "*identifierType*" que exprime o tipo do identificador único;

[15] https://data.4tu.nl/

“*creatorName*” quer dizer o nome completo do criador, “*givenName*” identifica o primeiro nome atribuído ao autor entre outros metadados.

Para adequar a descrição de um registro utilizando elementos do *Datacite Metadata Schema* para PROV, é viável analisar e sugerir a adição de elementos que não foram considerados no registro existente, de maneira semelhante a demonstração anterior, é possível atender à necessidade de interoperabilidade com a família PROV da seguinte forma:

Quadro 19 – Sugestão de adequação do registro em *Datacite Metada Schema*

```
<resource xmlns="http://datacite.org/schema/kernel-4"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://datacite.org/schema/kernel-4
http://schema.datacite.org/meta/kernel-4.4/metadata.xsd">
   <identifier identifierType="DOI">10.4121/40e06061-f441-43b5-9235-
006829206509.v1</identifier>
   <creators>
      <creator>
         <creatorName nameType="Personal">Juan Daniel Delgado-
Terán</creatorName>
         <givenName>Juan Daniel</givenName>
         <familyName>Delgado-Terán</familyName>
         <nameIdentifier nameIdentifierScheme="https://orcid.org/">0000-0001-9072-
0404</nameIdentifier>
      </creator>
   </creators>
   <titles>
      <title>Ankle Sensor Data from Parkinson's Disease Patients in Semi-Free Living
Conditions for Freezing of Gait Detection</title>
   </titles>
   <publisher>4TU.ResearchData</publisher>
   <publicationYear>2025</publicationYear>
<resourceType resourceTypeGeneral="Dataset">Dataset</resourceType>
<contributors>
      <contributor contributorType="DataCurator">
         <contributorName>Maria Silva</contributorName>
         <nameIdentifier nameIdentifierScheme="ORCID">0000-0002-1234-
5678</nameIdentifier>
      </contributor>
   </contributors>
   <dates>
      <date dateType="Collected">2023-03-15</date>
      <date dateType="Available">2025-01-01</date>
   </dates>
   <relatedIdentifiers>
      <relatedIdentifier relatedIdentifierType="DOI"
relationType="IsCitedBy">10.1016/j.example.2024.123456</relatedIdentifier>
   </relatedIdentifiers>
   <rights rightsURI="https://creativecommons.org/licenses/by/4.0/">Creative
```

```
Commons Attribution 4.0 International</rights>
   <fundingReferences>
      <fundingReference>
         <funderName>National Science Foundation</funderName>
         <awardNumber>123456</awardNumber>
      </fundingReference>
   </fundingReferences>
   <geoLocations>
      <geoLocation>
         <geoLocationPlace>Amsterdam, Netherlands</geoLocationPlace>
      </geoLocation>
   </geoLocations>
</resource>
```

Fonte: Elaborado pelo autor (2025)

Alguns exemplos concretos, destacados em cinza, podem ser observados no Quadro 19, os quais correspondem a família PROV:

- *prov:wasAttributedTo:* Associar criadores e contribuidores aos agentes responsáveis.
- *prov:generatedAtTime*: Associar a data de coleta ao momento de geração dos dados.
- *prov:wasDerivedFrom:* Associar identificadores relacionados à proveniência dos dados.

A incorporação de elementos extras ao *Datacite*, aliada à integração com PROV, possibilitaria uma descrição mais detalhada e consistente do recurso

Agora, no padrão de metadados DDI, apresenta-se o registro do repositório *Indepth Data Repository*[16], um dos registros que diz respeito a Bangladesh, Gana, África do Sul, Vietname - A evolução da transição demográfica e sanitária em quatro países de baixa e média renda.

Quadro 20 – Padrão de metadado DDI

```
<codeBook xmlns="http://www.icpsr.umich.edu/DDI"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="1.2.2"
ID="INDEPTH.GH001.Transitions.v1" xml-lang="en"
xsi:schemaLocation="http://www.icpsr.umich.edu/DDI
http://www.icpsr.umich.edu/DDI/Version1-2-2.xsd">
   <docDscr> <citation> <titlStmt>
      <titl> The evolving demographic and health transition in four low- and middle-
income countries </titl>
      <IDNo> DDI.INDEPTH.GH001.Transitions.v2 </IDNo>
   </titlStmt>
<prodStmt>
```

[16] http://www.indepth-ishare.org/

```
    <producer abbr="iS2TT" affiliation="INDEPTH Network" role="Documentation of
the study"> iSHARE2 Technical Team </producer>
    <producer abbr="int.indepth" affiliation="INDEPTH Network" role="agency">
INDEPTH Network </producer>
    <producer abbr="KHe" affiliation="INDEPTH Network" role="Documentation
author"> Kobus Herbst </producer>
    <software version="4.0.9" date="2013-04-23"> Nesstar Publisher </software>
</prodStmt> <verStmt>
```

Fonte: Extraído e Adaptado de (HOULE et al. 2016)
Disponível em: https://www.indepth-ishare.org/index.php/catalog/87/study-description#page=export-metadata&tab=study-desc
Acesso em: 18 jan. 2025.

No quadro 20, o registro apresenta a descrição de uma pesquisa que analisa o declínio na mortalidade e as mudanças na distribuição das causas de morte ao longo do tempo em quatro populações de países em desenvolvimento na África (África do Sul e Gana) e Ásia (Vietnã e Bangladesh). Esses metadados estão organizados em um formato estruturado utilizando *Extensible Markup Language* (XML), da mesma forma que foram extraídos.

Utilizando dados de vigilância demográfica e análise de histórico de eventos, o trabalho evidencia padrões de mortalidade e a transição epidemiológica, destacando a necessidade de sistemas de saúde adaptados para enfrentar desafios simultâneos de doenças transmissíveis e o aumento de condições não transmissíveis que demandam cuidados prolongados, especialmente em populações com alta prevalência de HIV, (HOULE et al. 2016).

Nesse contexto, foram identificados metadados como “*titl*” (título) atribuído à pesquisa, ao “*IDNo*” (número de identificação), à “*affiliation*” (afiliação),” date” (data de publicação, entre outros.

Do mesmo modo, é possível atender a demanda de interoperabilidade com a família PROV, da seguinte forma:

Quadro 21 – Sugestão de adequação do registro em DDI

```
<codeBook xmlns="http://www.icpsr.umich.edu/DDI"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="1.2.2"
ID="INDEPTH.GH001.Transitions.v1" xml-lang="en"
xsi:schemaLocation="http://www.icpsr.umich.edu/DDI
http://www.icpsr.umich.edu/DDI/Version1-2-2.xsd">
  <docDscr>
    <citation>
      <titlStmt>
        <titl>The evolving demographic and health transition in four low- and
middle-income countries</titl>
        <IDNo>DDI.INDEPTH.GH001.Transitions.v2</IDNo>
```

```
        </titlStmt>
        <prodStmt>
          <producer abbr="iS2TT" affiliation="INDEPTH Network"
role="Documentation of the study">iSHARE2 Technical Team</producer>
          <producer abbr="int.indepth" affiliation="INDEPTH Network"
role="agency">INDEPTH Network</producer>
          <producer abbr="KHe" affiliation="INDEPTH Network"
role="Documentation author">Kobus Herbst</producer>
          <software version="4.0.9" date="2013-04-23"> Nesstar Publisher
</software>
</prodStmt> <verStmt>
        </prodStmt>
        <verStmt>
          <version>2.0</version>
        </verStmt>
      </citation>
    </docDscr>
    <studyScope>
      <coverage>
        <geogCover>Bangladesh, Ghana, South Africa, Vietnam</geogCover>
        <popCover>Adults and children in demographic surveillance
sites</popCover>
      </coverage>
      <objective>To analyze mortality decline and changes in cause-of-death
distribution over time</objective>
    </studyScope>
    <dataColl>
      <timeMeth>Longitudinal</timeMeth>
      <dataCollector>INDEPTH Network</dataCollector>
      <collMode>Face-to-face interviews</collMode>
      <collDate date="2010-01-01/2015-12-31">2010-2015</collDate>
    </dataColl>
    <dataProc>
      <cleanOps>Data cleaning and validation were performed using Nesstar
Publisher</cleanOps>
      <anlyOps>Event history analysis was conducted to study mortality
patterns</anlyOps>
    </dataProc>
    <useStmt>
      <confDec>Data are available under restricted access due to privacy
concerns</confDec>
      <contact>Contact INDEPTH Network for access requests</contact>
    </useStmt>
    <relPubl>
      <citation>
        <titl>Mortality patterns in low- and middle-income countries</titl>
        <IDNo>DOI:10.1016/j.example.2016.123456</IDNo>
      </citation>
    </relPubl>
    <funding>
      <funder>Bill & Melinda Gates Foundation</funder>
      <grantNo>123456</grantNo>
    </funding>
</codeBook>
```

Fonte: Elaborado pelo autor (2025)

Alguns exemplos, destacados em cinza, podem ser observados no Quadro 21, os quais correspondem a família PROV:

- *prov:wasAttributedTo*: Associar produtores e afiliações aos agentes responsáveis.
- *prov:generatedAtTime:* Associar a data de coleta ao momento de geração dos dados.
- *prov:wasInformedBy*: Associar publicações relacionadas à proveniência dos dados.

O padrão DDI desempenha um papel fundamental no processo de pesquisa, abrangendo etapas como coleta, normalização, análise, compartilhamento e arquivamento de dados. Por adotar um esquema em XML, o DDI permite que os dados sejam vinculados a outras fontes de informação.

A partir da análise dos três padrões de metadados, *Dublin Core*, *DataCite* e DDI, que foram coletados de repositórios de dados de pesquisa, foi possível verificar a viabilidade da interoperabilidade semântica proposta pelo PROV. Essa abordagem possibilita a ampliação da descrição e a validação desta pesquisa, uma vez que, ao integrar os diferentes tipos de metadados, garante-se a consistência e a interoperabilidade entre os dados. A adoção das recomendações do PROV para cada um desses padrões permite uma descrição mais interoperável, essencial para a validação e o rastreamento da proveniência dos dados ao longo de seu ciclo de vida.

Por meio dos modelos apresentados, podemos verificar que as boas práticas aliadas aos metadados de proveniência, contribuem para uma maior clareza e credibilidade na publicação de dados nos repositórios de dados. Ao incluir informações como quem criou, publicou e emitiu o conjunto de dados, é possível estabelecer uma cadeia de responsabilidade clara.

Além disso, o uso de propriedades como *prov:actedOnBehalfOf* permite especificar relações de representação, indicando quando uma pessoa ou entidade age em nome de outra. Metadados como esse, facilitam a compreensão e o uso dos dados por pessoas e sistemas automatizados, promovendo maior interoperabilidade e reutilização.

De acordo com Lóscio, Burle e Calegari (2017), a principal dificuldade

relacionada aos metadados está na necessidade de definir o contexto dos dados dentro de um sistema de informação. Sem essa descrição adequada, a descoberta e o reaproveitamento dos dados tornam-se limitados, frequentemente ficando restritos ao próprio fornecedor.

No artigo intitulado Boas práticas para dados na web: análise do portal Dados Abertos Capes, as autoras Torino e Vidotti (2021), argumentam nas considerações finais:

> Considerando ainda os metadados, é imprescindível que a proveniência dos dados seja adequadamente registrada. A confiabilidade, a veracidade, a compreensão e o reuso estão dependentes dessa prática, capaz de assegurar a integridade e a autenticidade do dado, garantindo que ele não tenha sido adulterado e evidenciando a confiabilidade, o que demonstra direta relação com a sua qualidade e preservação.

O registro adequado de metadados de proveniência proporciona que os dados não tenham sido adulterados e facilita sua localização, permitindo que possam ser reaproveitados de maneira consistente e confiável. Dessa forma, metadados de proveniência associados as boas práticas, proporcionam a preservação dos dados, uma vez que possibilitam sua validação e o reuso em diferentes situações.

Conclui-se então, que a gestão eficiente de metadados de proveniência é fundamental para assegurar a confiabilidade, a transparência e a reutilização dos dados de pesquisa. Conforme destacado nas boas práticas propostas pelo W3C em *Data on the Web Best Practices*, a proveniência dos dados é um elemento essencial para que os consumidores possam avaliar a qualidade e a confiabilidade de um conjunto de dados. Nesse contexto, a adoção de padrões de metadados que descrevem a origem, o histórico e as transformações dos dados é crucial para garantir a rastreabilidade e a interoperabilidade.

Os padrões *prov:wasAttributedTo, prov:generatedAtTime e prov:wasInformedBy,* definidos pelo modelo PROV (PROV-O), desempenham um papel central na documentação da proveniência. Esses elementos permitem descrever de forma clara e estruturada:

1. ***prov:wasAttributedTo***: Associa agentes (pessoas, organizações ou sistemas) às entidades (dados ou recursos), indicando quem é responsável pela criação, publicação ou modificação dos dados. Esse padrão é essencial para

estabelecer uma cadeia de responsabilidade, garantindo que os usuários possam identificar os agentes envolvidos no ciclo de vida dos dados. Por exemplo, no padrão *Dublin Core*, o elemento *dct:creator* pode ser mapeado para *prov:wasAttributedTo*, permitindo associar criadores aos dados de forma interoperável.

2. ***prov:generatedAtTime***: Documenta o momento em que os dados foram gerados ou modificados. Esse padrão é fundamental para contextualizar temporalmente os dados, permitindo que os usuários entendam a atualidade e a relevância das informações. Em padrões como *DataCite*, a data de publicação (*publicationYear*) pode ser complementada com *prov:generatedAtTime* para indicar o momento exato da geração dos dados.
3. **prov:wasInformedBy**: Descreve a relação entre entidades, indicando que um conjunto de dados foi influenciado ou derivado de outro. Esse padrão é especialmente útil para documentar a proveniência em casos de reutilização ou integração de dados de diferentes fontes. Por exemplo, no padrão DDI, a relação entre estudos ou publicações pode ser explicitada com *prov:wasInformedBy*, garantindo a rastreabilidade das fontes originais.

Quadro 22 – Modelo de mapeamento entre padrões de metadados

Elementos PROV	**Descrição**	**Exemplo de Mapeamento**	**Finalidade**
prov:wasAttributedTo	Associa agentes (criadores, editores) a entidades (dados).	**dct:creator** (Dublin Core) → **prov:wasAttributedTo**	Atribuir autoria/responsabilidade de forma interoperável.
prov:generatedAtTime	Registra o momento exato de geração/modificação dos dados.	**datacite:publicationYear** (DataCite) → **prov:generatedAtTime**	Contextualizar temporalmente os dados para atualidade e rastreabilidade.
prov:wasInformedBy	Indica dependência ou derivação entre entidades (ex.: reuso de dados).	**ddi:isBasedOn** (DDI) → **prov:wasInformedBy**	Documentar fluxos de dados e origens para transparência.

Fonte: Elaborado pelo autor (2025)

No quadro 22, pode-se ver, em suma, como a Família PROV pode colaborar com a interoperabilidade semântica entre os padrões de metadados. A integração desses padrões aos esquemas de metadados existentes, como *Dublin Core, DataCite* e DDI, permite uma descrição mais robusta e interoperável da proveniência. Conforme demonstrado nos quadros 17, 19 e 21, a inclusão de elementos do PROV nos registros de metadados amplia a capacidade de rastreamento e validação dos dados, atendendo às recomendações das boas práticas propostas pelo W3C.

Além disso, a adoção desses padrões contribui para a interoperabilidade semântica, uma vez que eles fornecem um vocabulário comum para descrever a proveniência em diferentes contextos e domínios. Isso facilita a integração de dados de múltiplas fontes e promove a reutilização consistente e confiável dos dados.

7 CONSIDERAÇÕES FINAIS

Este estudo teve como objetivo principal avaliar como os metadados podem auxiliar na compreensão e gestão da proveniência de dados nos repositórios de dados, com foco na aderência dos padrões de metadados de proveniência, baseados na Família PROV, aos principais padrões utilizados na descrição de informações em repositórios de dados. A investigação buscou também identificar como tais padrões podem contribuir para assegurar a autenticidade e veracidade das informações, através dos metadados de proveniência. Essa análise foi fundamentada em uma abordagem teórica e exploratória, utilizando métodos científicos e o mapeamento *Crosswalk* para potencializar a interoperabilidade semântica entre diferentes padrões de metadados.

Antes de apresentar as porcentagens de correspondência entre os padrões de metadados analisados (*Dublin Core, DataCite Metadata Schema* e DDI *Codebook*) e os elementos da Família PROV, é importante destacar que os valores percentuais são calculados com base no total de elementos ou propriedades de cada padrão específico, ou seja, as porcentagens refletem a proporção de metadados de cada padrão que apresentam equivalência com a Família PROV, considerando o universo de metadados do próprio padrão.

Dessa forma, a comparação entre os três padrões não deve ser interpretada como uma equivalência direta entre eles, mas sim como uma análise individual de cada um em relação à PROV. Essa abordagem permite compreender como cada padrão se alinha aos princípios de proveniência, respeitando suas particularidades e estruturas próprias.

Os resultados obtidos indicam diferentes níveis de correspondência entre os padrões analisados e os elementos da Família PROV. No caso do *Dublin Core*, dos 22 elementos investigados, metade apresentou equivalência com a Família PROV, sendo três elementos classificados como *Crosswalking* absoluto (correspondência exata) e oito como *Crosswalking* relativo (sete mais específicos e um mais genérico). Em relação ao grau de equivalência, observou-se que três termos possuem correspondência exata, um é mais genérico, sete são mais específicos e 11 não possuem correspondência.

Nas propriedades do *Dublin Core*, considerando as 55 propriedades analisadas com base na documentação do DSpace 7.x, foram identificadas duas correspondências exatas, uma mais genérica e 11 mais específicas. Além disso, 39 propriedades não apresentaram correspondência com a Família PROV. Ao todo, 14 propriedades foram classificadas como *Crosswalking* relativo e 2 como *Crosswalking* absoluto. Dessa forma, a correspondência geral entre as classes e propriedades do *Dublin Core* com a PROV é de aproximadamente 35,06%, com 50% de correspondência entre as classes e 29,09% entre as propriedades.

Com base nos dados coletados a partir da realização do *Crosswalk* do padrão *DataCite Metadata Schema* para o modelo sugerido da Família PROV, foi possível observar que há 144 termos presentes no *DataCite*, e apenas 22 termos possuem equivalência com a PROV, sendo que 19 são *Crosswalking* relativo e três *Crosswalking* absoluto. O grau de equivalência identificado é: 122 termos sem correspondência, três correspondências exatas, três mais genéricos e 16 mais específicos.

A partir dos resultados obtidos, pode-se compreender que existe um nível razoavelmente baixo de equivalência entre os termos de metadados. A porcentagem de correspondência entre os termos do *DataCite* é de 15,28%, já que dos 144 termos, 22 apresentam correspondência, sendo três correspondências exatas, três mais genéricos e 16 mais específicos. Além disso, 122 termos não possuem correspondência com o modelo PROV. Portanto, a conversão dos padrões *DataCite* para PROV evidencia baixa possibilidade de correspondência com os metadados do W3C.

Com base nos dados coletados a partir da realização do *Crosswalk* entre o padrão DDI *Codebook* e o modelo sugerido da Família PROV, foi possível identificar que o DDI *Codebook* possui 148 termos de metadados. Desses, apenas dois apresentaram algum grau de correspondência com os termos da PROV, sendo dois termos mais genéricos. Dessa forma, o grau de equivalência entre os dois modelos reflete um cenário de baixíssima correspondência geral.

A correspondência entre os termos do DDI *Codebook e* a Família PROV é de

aproximadamente 1,35%, considerando que, dos 148 termos analisados, apenas dois apresentaram correspondência. Ambos os termos identificados possuem uma relação de maior especificidade em um *Crosswalking* relativo. Apesar dessa conexão, a integração entre o DDI *Codebook* e a Família PROV continua sendo um desafio, visto que a grande maioria dos termos, 146 de 148, não possuem correspondência exata ou específica com os termos da PROV.

Os resultados gerais demonstram que, enquanto o *Dublin Core* apresenta um nível moderado de correspondência com a PROV, o *DataCite Metadata Schema* e o DDI *Codebook* enfrentam desafios mais significativos para garantir a interoperabilidade semântica. Essas análises fornecem uma visão detalhada sobre as lacunas existentes e podem orientar estratégias futuras para aprimorar a integração desses padrões com a Família PROV.

A pesquisa apresenta contribuições significativas para a área de Ciência da Informação e Ciência da Computação. Ao explorar a relação entre metadados e proveniência em repositórios de dados, o estudo oferece perspectivas valiosas sobre como os padrões analisados podem ser adequados para melhorar a interoperabilidade e a rastreabilidade das informações. A análise dos mapeamentos realizados também proporciona um ponto de partida para desenvolvedores e gestores de repositórios que desejam implementar metadados de proveniência em seus sistemas.

Além disso, o trabalho sugere caminhos para futuras investigações, incluindo a análise de outros padrões de metadados que não foram abordados neste estudo, bem como a avaliação de contextos específicos de aplicação da Família PROV. Por exemplo, explorar como a PROV pode ser aplicada em outros repositórios ou em padrões de metadados específicos, utilizados em áreas de pesquisa como padrões de metadados geoespaciais ou imagéticos.

Embora a compatibilidade entre os padrões analisados e a Família PROV seja limitada em alguns casos, os resultados demonstram o potencial de integração e adaptação. Essa constatação ressalta a importância de iniciativas colaborativas entre comunidades acadêmicas e desenvolvedores de padrões para superar barreiras à

interoperabilidade e promover, ainda mais, a confiabilidade das informações prestadas em ambientes digitais.

Dessa forma, é fundamental que os profissionais da tecnologia da informação busquem constantemente atualizar suas competências, garantindo a capacidade de gerenciar adequadamente os dados de pesquisa, atender às necessidades dos pesquisadores interessados em compartilhar e disseminar seus dados, e responder de forma eficaz às demandas dos usuários por busca e acesso a essas informações. Essa atuação contribui diretamente para o fortalecimento da sociedade, promovendo o uso responsável e estratégico dos dados disponíveis, otimizando tempo que é essencial para os pesquisadores e usuários.

Em síntese, é possível responder que: os padrões de metadados utilizados para a descrição de dados em repositórios de pesquisa apresentam potencial para assegurar a rastreabilidade e a autenticidade das informações ao longo do tempo, mas com limitações. A análise realizada evidenciou que padrões como *Dublin Core*, *DataCite Metadata Schema* e DDI *Codebook* possuem graus variados de correspondência com os elementos da Família PROV, que é referência em modelos de proveniência. Apesar de alguns padrões demonstrarem um nível razoável de compatibilidade, como o *Dublin Core*, outros apresentam baixa aderência, como o *DataCite* e o DDI *Codebook*. Isso indica que, embora exista viabilidade para a integração e o fortalecimento da rastreabilidade e autenticidade, adaptações são necessárias para melhorar a interoperabilidade entre padrões e atender plenamente aos requisitos de confiabilidade em ambientes digitais.

Por fim, destaca-se que o estudo reforça a relevância de uma abordagem interdisciplinar para o avanço do campo dos metadados e da proveniência digital. A integração de conhecimentos da Ciência da Informação e da Ciência da Computação é essencial para atender às demandas de autenticidade, confiabilidade e rastreabilidade em sistemas de informação contemporâneos.

REFERÊNCIAS

ALBUQUERQUE, A. C.; SOUTO, D. V. B. Acerca do princípio da proveniência: apontamentos conceituais. **Ágora**, v. 23, n. 46, p. 14-44, 2013. Disponível em: http://hdl.handle.net/20.500.11959/brapci/13256. Acesso em: 13 jun. 2023.

ALVES, R. C. V. **Metadados como elementos do processo de catalogação.** 2010. 132 f. Tese (Doutorado em Ciência da Informação)–Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010.alves e santos 2013.

ALVES, R. C. V.; SANTOS, P. L. V. A. da C. **Metadados no domínio bibliográfico**. Rio de Janeiro: Intertexto, 2013.

ALVES, R. C. V. **Metadados como elementos do processo de catalogação**. 2010. 132 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010. Disponível em: http://repositorio.unesp.br/handle/11449/103361. Acesso em: 22 jul. 2024.

ARAKAKI, F. A. Metadados e modelo prov: perspectivas dos dados de proveniência em contextos digitais. **Informação & Informação**, v. 25, n. 3, p. 187-211, 2020. DOI: 10.5433/1981-8920.2020v25n3p187 Acesso em: 13 jun. 2023.

ARAKAKI, F. A; SANTOS, P. L. V. A. C. Proveniência e contexto digital: contribuições da ciência da informação. **Palabra Clave (La Plata), La Plata**, v. 10, n. 2, 1 abr. 2021. Universidad Nacional de La Plata. DOI: http://dx.doi.org/10.24215/18539912e124.

ARAKAKI, F. A. **Metadados administrativos e a proveniência dos dados: modelo baseado na família PROV.** Tese (Doutorado) – Universidade Estadual Paulista (UNESP), Faculdade de Filosofia e Ciências, 2019.

ARAKAKI, F. A.; GALEFFI, L. F; et al. BIBFRAME: tendência para a representação bibliográfica na web. **RBBD. Revista Brasileira de Biblioteconomia e Documentação**, v. 13, n. 0, p. 2231–2249, 23 dez. 2017. Disponível em: <https://rbbd.febab.org.br/rbbd/article/view/995>. Acesso em: 28 jun. 2023.

ARCE, A.; CANTWELL-JONES, A.; TANSLEY, M.; BARNES, I.; BRACE, S.; MULLIN, M.; NOTTON, D.; OLLERTON, J.; EATOUGH, E.; RHODES, M.W.; BIAN, X.; HOGAN, J.; HUNTER, T.; JACKSON, S.; WHIFFIN, A.; BLAGODEROV, V.; BROAD, G.; JUDD, S.; KOKKINI, P.; LIVERMORE, L.; DIXIT, M.K.; PEARSE, W.D.; GILL, R. UK museum specimen metadata and procrustes distance between left and right forewings, for five bumblebee species. **NERC EDS Environmental Information Data Centre**, 2022. Disponível em: https://doi.org/10.5285/2696535e-564a-4c6a-877e-515996fa97a1. Acesso em: 18 jan. 2025.

ARQUIVO NACIONAL (BRASIL). **Dicionário brasileiro de terminologia arquivística.** Rio de Janeiro: Arquivo Nacional, 2015. Disponível em: http://www.arquivonacional.gov.br/images/pdf/Dicion_Term_Arquiv.pdf. Acesso em: 05 abr. 2024.

AZAMBUJA, Luis Alberto Barbosa. **Proposta de inspeção de usabilidade de um repositório de dados de pesquisa brasileiro**. 2019. Dissertação (Mestrado) – Universidade Federal de Santa Catarina, Florianópolis, 2019.

BACA, M. **Introduction to Metadata.** 3rd ed. Los Angeles: Getty Research Institute, 2016.

BALL, A. **Review of data management lifecycle models**. Bath: University of Bath, 2012. Disponível em: https://purehost.bath.ac.uk/ws/portalfiles/portal/206543/redm1rep120110ab10.pdf. Acesso em: 24 dez. 2024.

BAPTISTA, A. A. **"Repositórios institucionais: democratizando o acesso ao conhecimento".** Salvador: EDUFBA p. 71-90. Disponível em: https://repositorium.sdum.uminho.pt/handle/1822/11517. Acessado em: 03 out. 2024

BELLOTTO, H. L. A diplomática como chave da teoria arquivística. **Archeion Online,** v. 3, n. 2, 2015. Disponível em: http://hdl.handle.net/20.500.11959/brapci/14906. Acesso em: 13 jun. 2023.

BODÊ, E. C. **Preservação de coleções de digitais**: o papel dos formatos. 2008. 153f. Dissertação (Mestrado em Ciência da Informação) – Universidade de Brasília, Brasília, 2008.

BOERES, S. A. A.; ARELLANO, M. A. M. Políticas e estratégias de preservação de documentos digitais. *In*: ENCONTRO DA CIÊNCIA DA INFORMAÇÃO, 6., 2005, Salvador. **Anais** […]. Salvador: 2005 Disponível em: http//www.cinform.ufba.br. Acesso em 15 set. 2024.

BONETTI, L. G.; SILVA, T. G. M. da; GABRIEL JUNIOR, R. F.; SOUZA, M. G. de; RODRIGUES, H. F.; ARAKAKI, A. C. S. Níveis de FAIRness nos Repositórios de Dados Aleia e Deposita Dados. **Ciência da Informação**, v. 53, n. 3, 29 out. 2024. Disponível em: https://revista.ibict.br/ciinf/article/view/7214. Acesso em: 1 abr. 2025.

BRASIL. Lei Nº 12.527, de 18 de novembro de 2011. **Diário Oficial da União**: seção 1, Brasília, DF, ano 148, n. 221-A, 18 nov. 2011. Disponível em: https://legislacao.presidencia.gov.br/atos/?tipo=LEI&numero=12527&ano=2011&ato=dc1UTUU1UMVpWT65a Acesso em: 13 jun. 2023

BRASE, J. European initiative to facilitate access to research data. ***D-Lib Magazine***, v. 15, 2009. Disponível em: https://dx.doi.org/10.1045/may2009-inbrief. Acesso em: 12 dez. 2024.

BROOKS, David. The Philosophy of data. **The New York Times**, New York, 4 Feb. 2013. Disponível em: https://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html. Acesso em: 20 fev. 2024.

BUFREM, Leilah Santiago; SILVA, Fábio Mascarenhas e; SOBRAL, Natanael Vitor;

CORREIA, Anna Elizabeth Galvão Coutinho. Produção Internacional Sobre Ciência Orientada a Dados: análise dos termos data science e e-science na scopus e na web of science. **Informação & Informação**, [S.L.], v. 21, n. 2, p. 40, 20 dez. 2016. Universidade Estadual de Londrina. http://dx.doi.org/10.5433/1981-8920.2016v21n2p40.

CASTRO, F. F.; SIMIONATO, A. C. **Revisitando ontologia e metadados à luz dos ambientes informacionais digitais.** Perspectivas em Ciência da Informação, [*s. l.*], v. 25, n. 4, p. 3-23, dez. 2020. DOI: http://dx.doi.org/10.1590/1981-5344/3329.

COSTA, M. M; CUNHA, M. B. O bibliotecário no tratamento de dados oriundos da e-science: considerações iniciais. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 19, n. 3, p. 189-206, 2014. Disponível em: https://doi.org/10.1590/1981-5344/1900. Acesso em: 04 out. 2024.

COSTA, M. P.; BRAGA, T. Repositórios de dados de pesquisa no mundo. **Cadernos BAD (Portugal)**, v., n., 2016.

CONEGLIAN, C. S.; SEGUNDO, J. E. S**. Materialização da Web Semântica: um modelo de construção dinâmica de consultas baseados em mapeamento de ontologias.** Perspectivas em Ciência da Informação, v. 23, n. 2, p. 33-49, 2018. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/22557. Acesso em: 28 dez. 2024.

CURTY, R. G.; GAMA, F. A. **Conjugando diplomática e xml: aproximação possível no contexto da proveniência de documentos jurídicos digitais.** Arquivística.net, v. 3, n. 2, 2007. Disponível em: http://hdl.handle.net/20.500.11959/brapci/50512. Acesso em: 13 jun. 2023.

CURTY, R. G. O paradigma da publicação de dados e suas diferentes abordagens. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília. **Anais** [**...**]. Marília: UNESP, 2017. p. 1-20.

GRANT, Rebeca. Recordkeeping and research data management: a review of perspectives. **Records Management Journal**, Dublin, v. 27, n. 2, p. 159-174, 2017. Disponível em: http://hdl.handle.net/10197/8769. Acesso em: 10 out. 2024.

CHADEGANI, Arezoo Aghaei et al. A comparison between two main academic literature collections: Web of Science and Scopus databases. **Asian Social Science**, Toronto, v. 9, n. 5, p. 18-26, 2013. Disponível em: <https://arxiv.org/ftp/arxiv/ papers/1305/1305.0377.pdf>. Acesso em: 23 jun. 2016.

CHAN, L. M.; ZENG, M. L**.** Metadata interoperability and standardization–a study of methodology part I. **D-Lib magazine**, v. 12, n. 6, p. 3, 2006. Disponível em: https://www.dlib.org/dlib/june06/chan/06chan.html#Zeng-Xiao. Acesso em 15 maio 2024.

DATACITE. DataCite Metadata Schema*.* Disponível em: https://schema.datacite.org/.

Acesso em: 12 dez. 2024.

DELGADO-TERÁN, J. D. *Ankle Sensor Data from Parkinson's Disease Patients in Semi-Free Living Conditions for Freezing of Gait Detection*. **4TU.ResearchData**, 2025. DOI: 10.4121/40e06061-f441-43b5-9235-006829206509.v1. Disponível em: < https://data.4tu.nl/datasets/40e06061-f441-43b5-9235-006829206509/1 >. Acesso em: 18 jan. 2025.

DONDI, C.; LEFFERTS, M.; DELFT, M. V. Pesquisa de proveniência e o consortium of european research libraries. **Ponto de Acesso**, v. 16, n. 3, p. 186-208, 2022. DOI: 10.9771/rpa.v16i3.52306 Acesso em: 13 jun. 2023.

DUVAL, E. *et al*. Metadata Principles and Practicalities. **D-Lib Magazine,** [S.l.], v. 8, n. 4, p. 1-10, 2002. Disponível em: https://www.dlib.org/dlib/april02/weibel/04weibel.html. Acesso em: 22 jul. 2024.

FERREIRA, Norma Sandra de Almeida. As Pesquisas Denominadas “Estado Da Arte”. **Educação & Sociedade**, v. 23, n. 79, ago. 2002. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-73302002000300013&lng=pt&nrm=iso&tlng=pt>. Acesso em: 02 dez. 2024.

FORCE11. **The FAIR Data Principles**. La Jolla, CA: Force11, 2016. Disponível em: <https://www.force11.org/group/fairgroup/fairprinciples>. Acesso em: 17 ago. 2018.

FREUND, G. P.; SEMBAY, M. J.; MACEDO, D. D. J. Proveniência de dados e segurança da informação: relações interdisciplinares no domínio da ciência da informação. **Revista Ibero-Americana de Ciência da Informação**, v. 12 No 3, n. 3, p. 807-825, 2019. DOI: 10.26512/rici.v12.n3.2019.21203 Acesso em: 13 jun. 2023.

GAVA, Tânia Barbosa Salles; FLORES, Daniel; ALEIXO, Diana Vilas Boas Souto; CRISTOVÃO, Henrique Monteiro; FERRARI, Luciana Itida; MORAES, Margarete Farias de. Dados de pesquisa na Arquivologia: uma reflexão. **Em Questão**, [S.L.], v. 30, n. 1, p. 1-29, 2024. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/1808-5245.30.135857. Disponível em: https://doi.org/10.1590/1808-5245.30.135857. Acesso em: 06 out. 2024.

GEZELTER, D. What, exactly, is Open Science? **The OpenScience Project**, 2009. Disponível em: <http://www.openscience.org/blog/?p=269>. Acesso em: 18 jan. 2025.

GIL, Y.; MILES, S. **PROV Model Primer**. Disponível em: <https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/>. Acesso em: 18 jun. 2024.

GUPTA, S.; MÜLLER-BIRN, C. A study of e-Research and its relation with research data life cycle: a literature perspective. **Benchmarking**: an international journal, [England], v. 25, n. 6, p.1656-1680, 2018. Disponível em: https://doi.org/10.1108/BIJ-02-2017-0030. Acesso em: 24 dez. 2024.

GRAY, Jim. Jim gray on escience: a transformed scientific method. *In*: HEY, Tony;

Tansley, Stewart; TOLLE, Kristin (ed.). **The Fourth Data-Intensive Scientific Discovery Paradigm**. Washington: Microsoft research, 2009. p. 17- 31.

GRÁCIO, J. C. A. **Metadados para a descrição de recursos da Internet: o padrão Dublin**
Core, aplicações e a questão da interoperabilidade. 2002. 127 f. Dissertação (mestrado) –
Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2002. Disponível em:
http://hdl.handle.net/11449/93722. Acesso em: 21 jul. 2024.

HOULE, B; BAWAH, A; CLARK, S. Bangladesh, Ghana, South Africa, Vietnam - The evolving demographic and health transition in four low- and middle-income countries. **INDEPTH Data Repository,** 2016. Disponível em: <https://www.indepth-ishare.org/index.php/catalog/87/study-description>. Acesso em: 18 jan. 2025.

LEMIEUX, V.; INPROVENANCE GROUP. **Proveniência: passado, presente e futuro em perspectiva interdisciplinar e multidisciplinar**. In: LEMIEUX, Victoria (Ed.). Construindo confiança na informação. Berna: Springer International Publishing, 2016. p. 3-45.

LEBO, T.; SAHOO, S.; MCCGUINNESS, D. (ed.). **PROV-O**: the PROV Ontology. 2013. Disponível em: http://www.w3.org/TR/2013/REC-prov-o-20130430/. Acesso em: 05 out. 2024.

LIU, J. **Metadata and its applications in the Digital Library**: approaches and practices. London: Libraies Unliited, 2007.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. **Data on the Web Best Practices**. Disponível em: <https://www.w3.org/TR/dwbp/>. Acesso em: 17 dez. 2024.

MARTINS, D. L.; SILVA, M. F.; SANTAREM SEGUNDO, J. E.; SIQUEIRA, J. Repositório digital com o software livre Tainacan: revisão da ferramenta e exemplo de implantação na área cultural com a Revista Filme Cultura. **Encontro Nacional de Pesquisa em Ciência da Informação**, Marília (SP), v. 18, 2017. Disponível em:
https://brapci.inf.br/index.php/res/download/125134. Acesso em: 02 out. 2024.

MARQUES FELIPE, C. B.; DOS SANTOS, R. F. Avaliação de metadados em repositórios de dados de pesquisa sobre biodiversidade. **Em Questão**, Porto Alegre, v. 28, n. 3, p. 117591, 2022. DOI: 10.19132/1808-5245283.117591. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/117591. Acesso em: 13 jan. 2025.

MÉNDEZ RODRIGUEZ, E. **Metadatos y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales.** Espanha: Treas, 2002.

MILLER, S. J. **Metadata for digital collections: a how-to-do-it manual.** New York: Neal-Schuman, 2011.

MOREAU, L. e GROTH, P. (2013). **Provenance: an introduction to PROV.** Synthesis lectures on the semantic web: theory and technology, 3(4), 1-129. DOI: https://10.2200/s00528ed1v01y201308wbe007.

MOREAU, L; MISSIER, P. **PROV-DM: The PROV Data Model**. (2013). Disponível em: <https://www.w3.org/TR/2013/REC-prov-dm-20130430/>. Acesso em: 25 dez. 2023

NATIONAL INFORMATION STANDARDS ORGANIZATION. Issues in crosswalking content metadata standards. **Information standards quarterly**. v. 11, n. 1, p. 01–16, 1999.

NORTE, M. B. **Glossário de termos técnicos em Ciência da Informação**: inglês/português. São Paulo: Cultura Acadêmica; Marília (SP): Oficina Universitária, 2010. Disponível em: https://www.marilia.unesp.br/Home/Publicacoes/glossario.pdf. Acesso em: 02 out. 2024.

PETRITSCH, B. **Metadata for research data in practice**. Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare, v. 70, p. 200–207, 2017.

POMERANTZ, J. **Metadata**. Cambridge, Massachusetts ; London, England: The MIT Press, 2015. (The MIT Press essential knowledge series).

RESEARCH DATA REPOSITORIES INFORMATION. 2021. Disponível em: www.re3data.org. Acesso em: 18 de mar. 2024.

RILEY, J. **Understanding Metadata: what is metadata, and what is it for?**. [S.l.]: National Information Standards Organization (NISO), 2004. Disponível em: <http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf>. Acesso em: 17 de mar. 2024.

SANCHEZ, F. A., DA SILVA, N. B. P., & VECHIATO, F. L. (2019). **Padrões de metadados para representação e organização da informação em repositórios de dados de pesquisa.** Informação &Amp; Tecnologia, 5(1), 37–51. https://doi.org/10.22478/ufpb.2358-3908.2018v5n1.38350

TORINO, E.; VIDOTTI, S. A. B. G. Boas práticas para dados na web: análise do portal Dados Abertos Cape: s. **Informação & Sociedade**, *[S. l.]*, v. 31, p. 1–25, 2021. DOI: 10.22478/ufpb.1809-4783.2021v31n1.50790. Disponível em: https://periodicos.ufpb.br/index.php/ies/article/view/50790. Acesso em: 17 dez. 2024.

VELLUCCI, S. L. Metadata. **Annual review of information science and technology (ARIST)**, v. 33, p. 187–222, 1998.

SANKOH, O. A. et al. **INDEPTH Network Cause-Specific Mortality - Release 2014.** 2014.
Dados de pesquisa. Disponível em: <http://www.indepthishare.

org/index.php/catalog/48/study-description>. Acesso em: 24 dez. 2024.

SAYÃO, L. F.; SALES, L. F. **Algumas considerações sobre os repositórios digitais de dados de pesquisa.** Informação & Informação, Londrina, PR, v.21, n.2, p.90–115, 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/27939>. Acesso em: 02 out. 2024.

SAYÃO, Luís Fernando; SALES, Luana Farias. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. **Informação e Sociedade**, João Pessoa, v. 22, n. 3, p. 179-191, 2012. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/12224/8586. Acesso em: 20 out. 2024.

SAYÃO, Luís Fernando; SALES, Luana Farias. Afinal, o que é dado de pesquisa? **BIBLOS**, Rio Grande, v. 34, n. 2, p. 1-20, 2020. Disponível em: https://doi.org/10.14295/biblos.v34i2.11875. Acesso em: 10 out. 2024.

SEMELER, Alexandre Ribas; PINTO, Adilson Luiz. Os diferentes conceitos de dados de pesquisa na abordagem da biblioteconomia e dados. **Ciência da Informação**, Brasília, v. 48, n. 1, p. 113-129, 2019. Disponível em: https://doi.org/10.18225/ci.inf.v48i1.4461. Acesso em: 10 out. 2024.

SILVA, R. E. da. **As tecnologias da web semântica no domínio bibliográfico**. 134 f. 2013.
Dissertação (Mestrado em Ciência da Informação) – Universidade Estadual Paulista "Júlio de
Mesquita Filho", Faculdade de Filosofia e Ciências, Marília, 2013. Disponível em: http://repositorio.unesp.br/handle/11449/93653. Acesso em: 21 jul. 2024.

SILVA, S. L. N. **Mapeamento entre padrões de metadados: um estudo do dublin core e do bibframe**. 2022. 59 f. Monografia - Curso de Biblioteconomia, Faculdade de Ciência da Informação da Universidade de Brasília, Brasília, 2022.

SILVA, Felipe Ivo da; ARAKAKI, Felipe Augusto. Análise da produção científica sobre metadados de proveniência. **Anais do 9º Encontro Brasileiro de Bibliometria e Cientometria - Ebbc**, Brasília, v. 9, p. 1-4, 23 jul. 2024. Instituto Brasileiro de Informação em Ciência e Tecnologia. http://dx.doi.org/10.22477/ix.ebbc.238. Disponível em: https://ebbc.inf.br/ojs/index.php/ebbc/article/view/238. Acesso em: 06 out. 2024.

SILVA, Amanda Marissa Soares*; et al*. Curadoria digital e arquivologia: olhares sobre o documento arquivístico digital. **Revista Ibero-Americana de Ciência da Informação,** Brasília, v. 14, p. 567-582, 2021. Disponível em: https://doi.org/10.26512/rici.v14.n2.2021.37558. Acesso em: 06 out. 2024.

SIMMHAN, Yogesh L.; PLALE, Beth; GANNON, Dennis. A survey of data provenance techniques. **Computer Science Department, Indiana University, Bloomington IN**, v. 47405, p. 69, 2005.

SIMIONATO, A. C. **Modelagem conceitual DILAM: princípios descritivos de arquivos, bibliotecas e museus para o recurso imagético digital.** 2015. 200 f. Tese (Doutorado em Ciência da Informação) - Universidade Estadual Paulista “Júlio de Mesquita Filho”, Marília, 2015.

ST. PIERRE, M.; LAPLANT, W. P. **Issues in crosswalking content metadata standards.** Baltimore: NISO, 1998. Disponível em: <http://www.niso.org/publications/white_papers/crosswalk>. Acesso em: 19 fev. 2013.

TOGNOLI, N. B. e GUIMARÃES, J. A. C. **Provenance as a knowledge organization principle.** 2019. Knowledge organization, 46(7), 558-68. Disponível em: http://www.isko.org/cyclo/provenance Acesso em: 11 jun. 2023.

TOMOYOSE, K. **Vocabulários para a publicação de dados de pesquisa nos princípios *Linked Data*: aspectos do *Data Catalog Vocabulary* (DCAT)**. Dissertação de Mestrado – Programa de Pós-graduação em Ciência da Informação Tomoyose, Universidade Federal de São Carlos, 2021, 185 f.

THE ROYAL SOCIETY. **Science as an open enterprise**. Londres: The Royal Society Science Policy Centre, 2012. 105p. Disponíıvel em: <https://royalsociety.org/˜/media/policy/projects/sape/2012-06-20-saoe.pdf>. Acesso em: 05 out. 2024.

VARDIGAN, M. The DDI Matures: 1997 to the Present. **IASSIST Quarterly**, Chapel Hill, v. 37, n. 1-4, p. 45-50, 2013. Disponível em: https://doi.org/10.29173/iq501. Acesso em: 24 dez. 2024.

VIANA, C. Lúcia de M.; MÁRDERO A., Miguel Á.; SHINTAKU, M. **Repositórios institucionais em ciência e tecnologia: uma experiência de customização do DSpace.** 2005.

ZENG, M. L.; QIN, J. **Metadata.** 2. ed. New York: Neal-Schuman Publishers, 2016.

HAYNES, David. **Metadata for Information Management and Retrieval:** Understanding metadata and its use. [S.l.]: Facet Publishing, 2018.

JOUDREY, Daniel N.; TAYLOR, Arlene G.; WISSER, Katherine M. **The organization of information.** Fourth edition ed. Santa Barbara, California: Libraries Unlimited, 2018. (Library and information science text series).

JOSSERAND, C. 2016 Les Données de provenance des collections des bibliothéques. Mémoire d’étude. [Em linha]. (Jan. 2016). [Consult. 1 fev. 2020]. Disponível em: https://tinyurl.com/y7wajvk9.

WALPORT, M.; BREST, P. Sharing research data to improve public health. **The Lancet**, v. 377, n. 9765, p. 537–539, 18 fev. 2011.

WEIBEL, S. **The Dublin core: a simple content description model for electronic resources.** Bulletin of the American Society for Information Science, p.9-11, Oct./Nov. 1997.

WILKINSON, M. et al. The FAIR Guiding Principles for scientific data management and stewardship. **Revista Nature**, California, v. 3, n. 1, p. 1-9, 2016. Disponível em: https://www.nature.com/articles/sdata201618. Acesso em: 18 jan. 2025

WORLD WIDE WEB CONSORTIUM. **PROV Model Primer**. 2013. Disponível em: https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/ Acesso em: 11 jun. 2023.