Ciência da Informação *Express*
ISSN 2675-8717
https://doi.org/10.60144/v6i.2025.136

FAMÍLIA PROV E REPOSITÓRIO DE DADOS NO CONTEXTO DO RE3DATA

PROV Family and Data Repository in the context of Re3data

Felipe Ivo da Silva
Mestrando em Ciência da Informação. Universidade Federal de São Carlos, São Carlos, São Paulo, Brasil.
felipe_ivodasilva@hotmail.com
https://orcid.org/0009-0005-1379-4692

Felipe Augusto Arakaki
Doutor e mestre em Ciência da Informação. Universidade de Brasília, Brasília, Distrito Federal, Brasil.
felipe.arakaki@unb.br
https://orcid.org/0000-0002-3983-2563

RESUMO

Objetivo: investigar a adoção dos padrões de metadados da Família PROV em repositórios de dados de pesquisa listados no *Research Data Repositories Information*. **Metodologia:** para busca e recuperação de repositórios de dados, utilizou-se a base Re3Data e foram analisados 3.302 repositórios, dos quais apenas 8 adotam esses padrões. **Resultados:** o baixo uso revela desafios na gestão de dados, pois a proveniência é fundamental para rastreamento, integridade e reutilização dos dados ao longo do tempo. **Conclusão:** a pesquisa ressalta a necessidade de ampliar a adoção dos padrões PROV. O uso desses padrões contribui para uma gestão mais eficaz dos dados científicos, fortalecendo a transparência e a confiabilidade da ciência. A adoção desses metadados pode aprimorar a reprodutibilidade da pesquisa, garantindo que os dados sejam compreendidos e reutilizados corretamente. Assim, é essencial que políticas institucionais incentivem a implementação desses padrões nos repositórios de dados científicos.

Palavras-chave: Metadado. Proveniência. Repositório de dados. PROV. PROV-O.

ABSTRACT

Objective *this study investigates the adoption of PROV Family metadata standards in research data repositories listed in the Research Data Repositories Information.* ***Method:*** *the Re3Data database was used to search and retrieve data repositories, and 3,302 repositories were analyzed, of which only 8 adopt these standards.* ***It resulted****: The low use reveals challenges in data management, since provenance is essential for tracking, integrity, and reuse of data over time.* ***Conclusions:*** *The research highlights the need to expand the adoption of PROV standards. The use of these standards contributes to more effective management of scientific data, strengthening the transparency and reliability of science. The adoption of these metadata can improve research reproducibility, ensuring that data are understood and reused correctly. Therefore, it is essential that institutional policies encourage the implementation of these standards in scientific data repositories.*

Keywords: *Metadata. Provenance. Data repository. PROV.*

1 INTRODUÇÃO

A informação desempenha um papel fundamental no desenvolvimento global e, com o advento da tecnologia, a quantidade de informações digitais se proliferou significativamente. Segundo Brandt *et al.* (2019), com esse aumento exponencial de informação, a discussão sobre metadados ganhou destaque nas áreas de informação e tecnologia. À luz disso, o estudo da Proveniência se faz indispensável para a caracterização e confiabilidade da informação. Haynes (2018, p. 134, tradução nossa) ressalta que a disponibilização de informações sobre a proveniência é essencial para comprovar a autenticidade de um registro, garantindo que ele não tenha sido alterado e que a evidência apresentada seja confiável, já que permite a validação e a credibilidade dos recursos informacionais, como os metadados, especialmente no contexto digital.

Ademais, a proveniência é um termo que compreende diversas áreas de estudo, como ciência de dados, biblioteconomia e gestão de informações. Nesse sentido, o conceito de proveniência adotada neste artigo é aquela descrita por Moreau e Groth (2013), que caracteriza um conjunto de informações sobre indivíduos, instituições, entidades e ações ligadas à criação, influência ou disseminação de dados ou objetos específicos.

Arakaki e Santos (2021) destacaram que o termo proveniência é utilizado para identificar o indivíduo ou a entidade responsável por gerar, armazenar e gerenciar informações e recursos em vários domínios. Assim sendo, esse processo de identificação beneficia-se dos metadados, que servem como uma solução que permite aos usuários avaliarem informações sobre um recurso informacional específico com mais precisão. Como resultado, é possível garantir a qualidade e a veracidade dos dados.

Assim, os repositórios de dados surgem como ferramentas de apoio à execução dos processos necessários para a gestão de dados de pesquisa (Sayão; Sales, 2016) e desempenham um papel fundamental na ciência aberta, permitindo que pesquisadores depositem seus dados para que possam ser acessados e

reutilizados por outros. No entanto, nem todos os repositórios adotam metadados de proveniência, o que pode afetar a qualidade e a usabilidade dos dados.

Em essência, os metadados podem ser considerados uma aplicação prática que diz respeito à descrição de objetos, ao desenvolvimento de bancos de dados e ao registro digital de transações. Joudrey, Taylor e Wisser (2018) também enfatizam que os metadados podem incluir informações descritivas sobre o contexto, a qualidade, a condição ou as características dos dados.

Esta definição implica que os metadados não incluam apenas informações descritivas, como as encontradas em ferramentas tradicionais de descoberta de recursos, mas também informações necessárias para o gerenciamento, uso e preservação do recurso de informação (por exemplo, dados de localização, informações de exibição on-line, informações de propriedade, dados de condição). Nesse sentido, os metadados são parte do processo de descrição e podem ser guiados por princípios de catalogação.

Sendo assim, a definição e a padronização dos metadados de recursos de informação estabelecem a criação de ferramentas que garantem a identificação e a preservação dos recursos de informação, bem como a busca, o acesso, o uso e a reutilização de informações. Então, os metadados de proveniência se apresentam como uma solução para os usuários avaliarem com maior precisão a escolha de um recurso de informação específico, garantindo a qualidade e a precisão dos dados. Portanto, os metadados são vistos como uma aplicação prática relacionada à catalogação, à indexação, ao desenvolvimento de banco de dados e ao registro de transações digitais.

Embora os metadados sejam amplamente discutidos na Ciência da Informação, a adoção da proveniência ainda é limitada nos repositórios de dados de pesquisa pois, como afirma Arakaki (2019), “o estudo sobre a proveniência, mostrou que a temática no Brasil é incipiente e carece de pesquisas teóricas e iniciativas de cunho prático/profissional”. A lacuna na literatura reside na ausência de estudos que mapeiem sistematicamente a adoção dos padrões da Família PROV e seus impactos na confiabilidade e reutilização dos dados. Essa pesquisa busca preencher

essa lacuna ao identificar os repositórios que utilizam esses padrões, analisando desafios e implicações para a gestão de dados, contribuindo para a Ciência Aberta e para o aprimoramento da curadoria digital.

Diante desse contexto, o objetivo deste trabalho é verificar quais repositórios de dados utilizam os padrões de metadados da Família PROV, segundo o *Research Data Repositories Information* (R3DATA). Portanto, essa investigação é necessária para entender não apenas a necessidade de adesão a esses padrões, mas também a importância da proveniência em repositórios de dados de pesquisa, particularmente no contexto da Ciência Aberta.

Por fim, por meio dessa análise, pretende-se destacar a importância do uso de padrões de metadados de proveniência, em especial a Família PROV, e entender como esses padrões contribuem positivamente para a vanguarda da informação, garantindo a qualidade e a precisão dos dados e fortalecendo sua relevância na tecnologia.

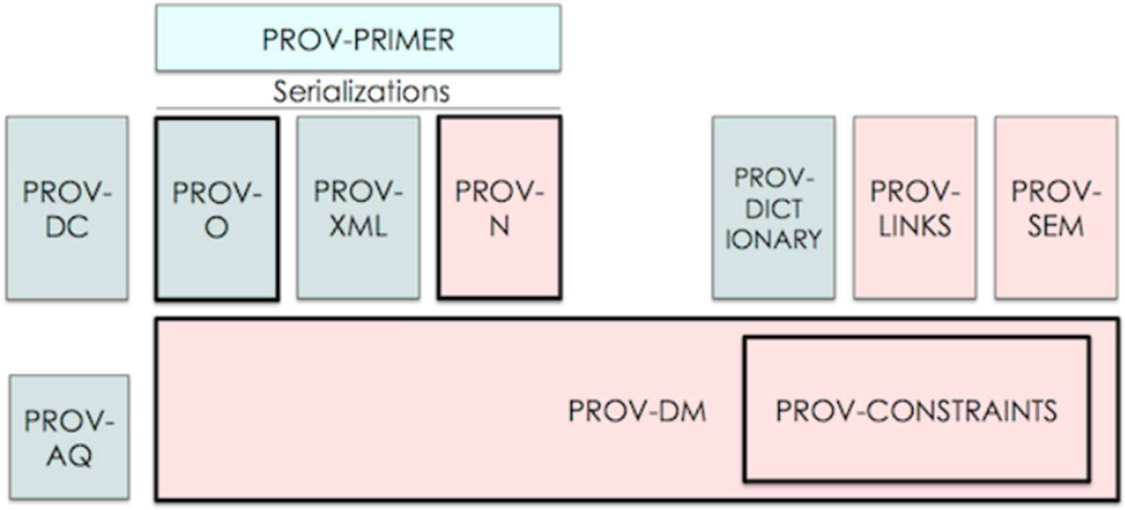
2 FAMÍLIA PROV

Em 2013, o World Wide Web Consortium (W3C) formou um grupo de trabalho para desenvolver um conjunto de especificações chamado Família PROV, que define um modelo para garantir a interoperabilidade de dados na web. Essa família é composta por quatro principais recomendações: 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 dessas diretrizes, foram publicadas oito notas complementares que oferecem orientações e informações adicionais, facilitando o entendimento e a aplicação do modelo PROV.

A Figura 1 ilustra a relação entre os documentos que compõem a Família PROV, destacando a estrutura e as conexões entre elas.

Ciência da Informação Express
ISSN 2675-8717

Figura 1 - Família PROV



Fonte: Groth e Moreau (2013, não paginado).

De acordo com Groth e Moreau (2013), a família de documentos PROV define um modelo, serializações e outras definições de suporte que permitem a troca de informações de proveniência em ambientes heterogêneos, como a Web. Assim, o documento W3C sobre a família PROV fornece uma visão geral não normativa, além de estabelecer diretrizes e sugestões para o seu uso.

Assim, a família PROV, desenvolvida pelo W3C, oferece um conjunto de especificações que promovem a interoperabilidade e a troca de informações de proveniência na Web, assegurando que metadados abrangem não apenas descrições contextuais, mas também informações essenciais para o gerenciamento, uso e preservação dos dados. Ao permitir o rastreamento detalhado de informações sobre a origem, os padrões de metadados PROV fornecidos pela W3C auxiliam na promoção da garantia dos dados disponibilizados em repositórios de dados.

Para facilitar a adoção do modelo PROV, o W3C realizou um mapeamento com o *Dublin Core*, de modo a permitir que o padrão de metadados de proveniência

seja mais facilmente incorporado em repositórios de dados que já adotam o *Dublin Core* como metadado principal.[1]

A partir do Quadro 1, é possível observar um exemplo de aplicação da propriedade prov.wasGeneratedBy da Família PROV. O exemplo utiliza prefixos, como RDF e XSD, para definir *namespaces,* e estabelece relações entre entidades e atividades. Esse modelo ilustra como a propriedade prov.wasGeneratedBy pode ser utilizada para indicar que uma entidade foi gerada por uma determinada atividade, seguindo as especificações da W3C para a Família PROV.

Quadro 1 – Exemplo da propriedade: prov:wasGeneratedBy.

```
1 @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
2 @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
3 @prefix owl: <http://www.w3.org/2002/07/owl#> .
4 @prefix prov: <http://www.w3.org/ns/prov#> .
5 @prefix : <http://example.com/> .
6 :bar_chart
7    a prov:Entity;
8    prov:wasGeneratedBy :illustrating;
9 :illustrating a prov:Activity .
```

Fonte: Elaborado pelo W3C[2].

No Quadro 2, é apresentado um exemplo de aplicação da propriedade prov.wasDerivedFrom da Família PROV. O exemplo ilustra como a propriedade prov.wasDerivedFrom pode ser empregada para representar a relação de derivação entre entidades, seguindo as especificações da W3C para a Família PROV.

Quadro 2 – Exemplo da propriedade: prov:wasDerivedFrom.

```
1 @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
2 @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
3 @prefix owl: <http://www.w3.org/2002/07/owl#> .
4 @prefix prov: <http://www.w3.org/ns/prov#> .
5 @prefix ex: <http://example.com/vocab#> .
6 @prefix : <http://example.com/> .
7 :bar_chart
8    a prov:Entity, ex:Barchart;
9    prov:wasDerivedFrom :aggregatedByRegions;
10 .
```

[1] https://www.w3.org/TR/2012/WD-prov-dc-20121211/
[2] https://www.w3.org/TR/prov-o/#wasGeneratedBy

```
11 :aggregatedByRegions
12    a prov:Entity, ex:Dataset;
13 .
```

Fonte: Elaborado pelo W3C[3].

Os modelos apresentados nos quadros 1 e 2, propostos pela W3C, têm como objetivo orientar a aplicação e o uso mais eficaz dos padrões de metadados. Esses modelos destacam características essenciais para o processo de representação de recursos informacionais, contribuindo para a sua implementação. Ambos os modelos enfatizam a importância da representação clara e estruturada das relações entre entidades, atividades e agentes, seguindo as especificações da W3C.

De acordo com o Re3data (2024), o *Dublin Core* é o padrão de metadado mais utilizado em repositórios de dados, são cerca de 619 repositórios. Além disso, a compatibilidade com o PROV facilita a troca de informações detalhadas sobre a proveniência dos dados nesses ambientes, maximizando o alcance e a eficácia da aplicação dos metadados de proveniência.

Por fim, os metadados padronizados e bem definidos são fundamentais para a criação de ferramentas de catalogação que aprimoram a busca, o acesso, o uso e a reutilização de informações, consolidando sua relevância na tecnologia.

3 REPOSITÓRIOS DE DADOS DE PESQUISA

A intensa geração de dados e informações no cenário acadêmico-científico impulsiona a criação de ambientes digitais informacionais, como os Repositórios de Dados de Pesquisa, que são “[...] ferramenta necessária para armazenar e gerenciar os dados utilizados e produzidos durante uma pesquisa” (Vidotti *et al*, 2017, p. 8).

No contexto da Ciência Aberta, o *Research Data Repositories Information,* mais conhecido como Re3data, é um registro global de repositórios de dados de pesquisa que centraliza os dados de pesquisa de diversas áreas do conhecimento, disponíveis em vários formatos e acessíveis a todos.

[3] https://www.w3.org/TR/prov-o/#wasDerivedFrom

Este diretório global, iniciado na Alemanha em 2012, facilita a consulta de dados de pesquisa, uma vez que inclui repositórios que permitem o armazenamento permanente e o acesso a conjuntos de dados para pesquisadores, órgãos de financiamento, editores e instituições acadêmicas. Cabe aqui citar que entre os parceiros fundadores do Re3data, estão a Escola de Biblioteconomia e Ciência da Informação de Berlim, o Helmholtz Open Science Office do Centro Alemão de Pesquisa em Geociências GFZ, a Biblioteca KIT do Instituto de Tecnologia de Karlsruhe (KIT) e as Bibliotecas da Universidade Purdue (Re3data, 2024).

O diretório oferece múltiplas opções de busca, como país, tema e tipo de dado. Por isso, o Re3data cobre uma ampla gama de áreas do conhecimento, promovendo o compartilhamento, o acesso e a maior visibilidade dos dados de pesquisa. Cada resultado recuperado no diretório inclui uma descrição ou resumo com informações sobre as pesquisas correspondentes, as métricas de diferentes diretórios, além de detalhes como país de origem, tema ou palavra-chave e tipo de documento.

Para serem reutilizados, os dados de pesquisa devem ser preservados e facilmente acessíveis para localização e recuperação. De acordo com Hedstrom (1997), a preservação digital é um processo que abrange “planejamento, alocação de recursos e aplicação de métodos e tecnologias de preservação necessários para garantir que as informações digitais de valor contínuo permaneçam acessíveis e utilizáveis”. Nesse sentido, os metadados de proveniência se configuram como parte das atividades necessárias para garantir que os objetos digitais possam ser localizados, reproduzidos e compreendidos ao longo do tempo.

Diante disso, é possível afirmar que os repositórios de dados são fundamentais para o compartilhamento de dados, pois sua estrutura é projetada para esse propósito. Assim, a descrição detalhada e a recuperação dos dados armazenados são operações fundamentais nesses ambientes, uma vez que, além de fornecer o sistema, é necessário implementar uma gestão eficaz que permita o uso e reuso dos dados.

No âmbito da proveniência, essa gestão também assegura a preservação dos dados a longo prazo e facilita o acesso para aqueles que buscam reutilizá-los de forma eficiente, fornecendo informações indispensáveis para garantia informacional.

4 PROCEDIMENTOS METODOLÓGICOS

Este estudo caracteriza-se como uma revisão qualitativa exploratória e descritiva da literatura. Assim, a escolha deste tipo de pesquisa se deu por ser uma abordagem que permite contextualizar o problema da catalogação de conjuntos de metadados de proveniência com destaque para a família PROV. Adicionalmente, essa abordagem contribui para a construção de análises que oferecem subsídios para a formulação do referencial teórico a ser utilizado na pesquisa empreendida (Alves-Mazzotti, 2002).

Dessa forma, a análise dos dados é predominantemente qualitativa exploratória, sendo realizada a partir da categorização e da síntese das informações obtidas. Para alcançar o objetivo proposto, utilizou-se como instrumento metodológico o diretório *Registry of Research Data Repositories* - Re3data, a fim de identificar quais repositórios utilizam metadados de proveniência da família PROV. Assim, este método consiste em identificar padrões e categorizar as instituições responsáveis, assim como o nome do repositório e o tema de estudo predominante dentro do repositório.

Já que observa e compreende os diversos aspectos teóricos inerentes ao fenômeno estudado, é também uma pesquisa de natureza exploratória, realizando-se a partir das informações obtidas na revisão de literatura (Gil, 2017), além de descrever o uso de metadados de proveniência em Repositórios de Dados.

Para o levantamento dos dados, utilizou-se a seleção do Re3data a fim de mapear os repositórios de dados que fazem uso dos metadados de proveniência proposto pela *Word Web Consortium* (W3C).

5 ANÁLISE DOS RESULTADOS

Nesta seção, serão apresentadas as informações coletadas do Re3Data que demonstram quais são os repositórios de dados de pesquisa que utilizam a família PROV para descrição de suas coleções. Assim sendo, para garantir a qualidade dos dados coletados, foi utilizado os filtros de busca do Re3Data para a pesquisa de padrões de metadados em repositórios de dados – foram identificados 29 padrões de metadados em 3.302 repositórios. Este mapeamento focou exclusivamente na identificação de repositórios que adotam os padrões da Família PROV, sem investigar a existência de outros padrões de metadados que possam desempenhar funções semelhantes ou compatíveis com a proveniência. Então, observa-se que destes repositórios pode haver o uso de mais de um padrão de metadados, destacando-se que apenas oito repositórios utilizam a família PROV. [4]

A categorização apresentada no Quadro 3 foi realizada com base na análise dos repositórios de dados de pesquisa identificados no Re3Data que utilizam a família PROV para a descrição de suas coleções. Para isso, foram considerados três aspectos principais: a instituição responsável pelo repositório, sua localização geográfica e a área temática à qual está vinculado.

Essa categorização permite compreender a distribuição dos repositórios que adotam a proveniência como parte de seus metadados, evidenciando sua aplicabilidade em diferentes domínios do conhecimento.

Quadro 3 – Ficha de caracterização dos repositórios de dados que utilizam a PROV.

Nome do repositório	Instituição responsável - País	Tema/Área
Rotterdam Ophthalmic Data Repository	Rotterdam Ophthalmic Institute – Holanda	Medicamento e Ciências da Vida
DANDI	Dartmouth College – Estados Unidos	Medicamento, Neurociências e Ciências da Vida
MorphoSource	Duke University, Trinity College of Arts & Sciences – Estados Unidos	Geral

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

Jülich DATA	Forschungszentrum Jülich – Alemanha	Ciências da Vida, Ciências Naturais e Ciências da Engenharia
DesignSafe-CI Data Depot Repository	National Science Foundation – Estados Unidos	Geral
TROLLing	CLARIN-ERIC – União Europeia	Linguística, Humanidades e Ciências Humanas e Sociais
NSF Arctic Data Center	DataONE – Estados Unidos	Geral
DataverseNO	Inland Norway University of Applied Sciences – Noruega	Geral

Fonte: elaborado pelos autores (2024).[5]

O primeiro ponto que precisa ser destacado é que, embora os oito repositórios selecionados sejam mantidos por instituições de diferentes nacionalidades, como Estados Unidos, Holanda, Noruega e Alemanha, grande parte de seu conteúdo está em inglês. Outro aspecto, é que da perspectiva de descrição dos conteúdos, há metadados em todos os repositórios que identificam a origem do conteúdo. Além disso, nos registros, são observadas informações como título, autor ou responsabilidade, data e local de criação ou coleta, formato, tamanho, permissões e requisitos de uso e sistema.

Assim, foi observado que os padrões de metadados utilizados nos oito repositórios foram fornecidos pela *Digital Curation Centre* (DCC), um centro global, líder em curadoria de informações digitais, que se concentra no desenvolvimento de capacidade, competência e habilidades para gerenciar dados de pesquisa.

Essa instituição promove a tutela na descrição informacional dos repositórios que fazem uso dos metadados da Família PROV. Rusbridge, *et al*. (2005, p. 5, tradução nossa) demonstram a importância da implementação acerca da proveniência e qualidade dos dados:

> O valor da curadoria digital é diminuído se a evidência, quanto à origem e integridade dos dados, for perdida ou desconhecida. O trabalho atual do DCC nesta área inclui o desenvolvimento de modelos formais para dar suporte à descrição do estado do problema

[5] https://www.re3data.org/search?query=PROV

> e para estender nossa compreensão dos problemas que surgem quando os dados são copiados de um banco de dados para outro ou ao rastrear de onde os dados vieram que não foram documentados adequadamente (sua proveniência não é adequadamente clara). À medida que esses modelos formais surgem, eles ajudarão aqueles que desenvolvem ferramentas de software e padrões para rastrear, trocar e gerenciar a procedência de dados conforme eles são transferidos entre bancos de dados.

Desta maneira, o levantamento realizado revelou que apenas oito repositórios de dados de pesquisa afirmam que utilizam os padrões de metadados da Família PROV, conforme indicado pelo *Research Data Repositories Information*. De acordo com o Re3Data, todos esses repositórios estão vinculados a uma instituição, o *Digital Curation Centre* (DCC), que é responsável pela implementação dos esquemas de padrões de metadados da Família PROV nesses repositórios.

Apesar do DCC e do Re3data indicarem o uso dos padrões de metadados da família PROV nos oito repositórios, não foram localizados manuais que exemplificassem sua utilização e implementação. No entanto, as listas de perfis de aplicação de padrões de metadados do DCC, que indicam o uso da PROV, direcionam aos manuais originais da W3C[6].

A análise abrangeu os repositórios listados no Quadro 3, possibilitando a identificação dos repositórios TROLLing e DataverseNO que utilizam padrões PROV, especificamente os padrões prov:wasGeneratedBy e prov:wasDerivedFrom.

No Quadro 4, apresenta-se o mapeamento da presença ou ausência de padrões de metadados, bem como o formato dos arquivos analisados. Foram investigados os padrões de metadados dos dez primeiros documentos publicados em 2024, selecionados com base na atualidade de suas publicações.

Quadro 4 – Mapeamento dos repositórios para verificar a existência da PROV.

Nome do Repositório	Documentos em 01/01/2024 até 15/11/2024	Documentos analisados	Padrões PROV	Formato do Arquivo
Rotterdam Ophthalmic Data	10	10	Não foi possível identificar o uso de quaisquer padrões de	-

[6] https://www.dcc.ac.uk/resources/metadata-standards/prov

Repository			metadados.	
DANDI	131	10	0	Json
MorphoSource	141	10	O repositório possui metadados estruturados internamente, mas não os exibe explicitamente ao usuário.[7]	-
Jülich DATA	99	10	0	HTML
DesignSafe-CI Data Depot Repository	273	10	0	XML
TROLLing	27	10	10 documentos prov:wasGeneratedBy prov:wasDerivedFrom.	HTML
NSF Arctic Data Center	189	10	0	XML e HTML
DataverseNO	179	10	4 documentos prov:wasGeneratedBy prov:wasDerivedFrom.	HTML

Fonte: elaborado pelos autores (2024).

Diante do exposto, na amostragem observada, foram identificados quatro repositórios que não utilizavam padrões de metadados PROV na descrição documental; um repositório que possui metadados estruturados internamente, mas não os exibe explicitamente ao usuário, limitando a transparência e a verificação de padrões como a Família PROV; e um repositório que não foi possível localizar qualquer perfil de aplicação de padrões de metadados, o que aponta para a necessidade de maior transparência na disponibilização de metadados nos repositórios analisados.

As duas propriedades do PROV identificadas nos repositórios, prov:wasGeneratedBy e prov:wasDerivedFrom, tem como definição: 1) prov:wasGeneratedBy, gerado pela conclusão da produção de uma nova entidade por uma atividade; e 2) prov:wasDerivedFrom, derivado de uma transformação de uma entidade em outra, o que pode ser ainda uma atualização, ou a construção de uma nova entidade baseada em uma entidade pré-existente (Lebo *et al*., 2013).

[7] https://www.morphosource.org/terms/ms#metadata

Além disso, o repositório *Rotterdam Ophthalmic Data Repository* não disponibilizou nenhum tipo de formato de dados nos dez primeiros documentos analisados que permitisse a extração de qualquer tipo de metadado, impossibilitando a verificação do uso de padrões da Família PROV e de outros padrões. Já o repositório *MorphoSource*, embora não disponibilize formatos de dados visíveis ao usuário que permitam a extração direta de metadados nos primeiros documentos analisados, possui metadados estruturados internamente. Esses metadados, no entanto, não são exibidos de forma explícita na interface do usuário.

A ausência de informações estruturadas nesses dois repositórios representa uma lacuna na transparência dos dados, dificultando a avaliação da rastreabilidade e da interoperabilidade dos metadados.

Nos repositórios que apresentaram "0" padrões PROV, foram identificados outros padrões de metadados, mas sem uso do *namespace* explícito com a proveniência da Família PROV.

De modo geral, a análise deste cenário aponta para uma adversidade significativa: embora existam outras fontes que catalogam repositórios de dados de pesquisa, como o OpenDOAR[8] e DataCite[9], a quantidade de repositórios que adotam os metadados PROV ainda é significativamente baixa. Isso é preocupante, considerando a importância dos padrões de metadados de proveniência para identificação, preservação, descoberta, acesso, utilização e reutilização de informações (Moreau *et al*., 2011).

Por isso, a baixa adoção desses padrões indica uma lacuna na prática de gestão de dados que pode comprometer a integridade e a utilidade dos dados de pesquisa a longo prazo. A disseminação e a implementação mais amplas dos metadados PROV são importantes para melhorar a confiabilidade dos repositórios de dados e assegurar que a comunidade científica possa confiar nos dados disponibilizados (Moreau e Missier, 2013).

[8] https://v2.sherpa.ac.uk/opendoar/
[9] https://datacite.org/

6 CONSIDERAÇÕES FINAIS

Do ponto de vista da Ciência Aberta, o Re3data destaca-se como um diretório online de repositórios de dados, facilitando o acesso a dados de pesquisa em diversas áreas do conhecimento. Essas iniciativas desempenham um papel essencial na comunicação científica, ampliando a visibilidade das pesquisas e maximizando seu impacto na sociedade.

Os padrões de metadados são fundamentais para os Repositórios de Dados de Pesquisa, pois permitem uma descrição padronizada e estruturada das informações (Sanchez *et al*., 2018). No caso específico da Família PROV, seu uso possibilita o registro da proveniência dos dados, fornecendo um contexto detalhado sobre sua origem e transformações, o que fortalece a confiabilidade e a acessibilidade dessas informações. Além disso, os metadados de proveniência favorecem a interoperabilidade entre sistemas, documentando e referenciando todas as etapas envolvidas na produção e reutilização dos dados.

Entretanto, esta pesquisa apresenta algumas limitações. A análise foi baseada exclusivamente nos repositórios listados no Re3data, o que pode não abranger toda a diversidade de repositórios existentes. Além disso, a metodologia adotada permitiu mapear a adoção dos padrões PROV, mas não aprofundou os motivos específicos que levam à baixa adesão. Estudos futuros poderiam explorar esses desafios, identificando barreiras institucionais, técnicas e culturais que dificultam a implementação dos metadados de proveniência.

Os resultados obtidos são parcialmente convergentes com as expectativas do estudo, pois confirmam que a adoção dos padrões da Família PROV em repositórios de dados de pesquisa ainda é limitada, estando alinhada com a literatura que aponta desafios na implementação de metadados de proveniência. No entanto, os achados também contrastam com algumas expectativas, uma vez que, apesar da crescente valorização da Ciência Aberta identificou-se um número reduzido de repositórios que utilizam a Família PROV.

Como implicações teóricas e aplicadas, este estudo reforça a importância da Família PROV na gestão dos dados de pesquisa e sugere que sua adoção deve ser incentivada por meio de políticas institucionais e iniciativas educacionais. Além disso, novas investigações podem examinar estratégias para ampliar essa adoção, bem como avaliar o impacto do uso dos metadados de proveniência na reprodutibilidade e integridade dos dados científicos, além da possibilidade de ampliar a análise para um número maior de documentos dentro dos repositórios já examinados, permitindo uma investigação mais abrangente sobre a presença e a utilização dos padrões da Família PROV.

Por fim, destaca-se a necessidade de um esforço contínuo da comunidade científica para integrar e promover os padrões de proveniência nos repositórios de dados para garantia da proveniência e integridade dos dados. O suporte institucional, como o fornecido pelo DCC, aliado à conscientização sobre os benefícios desses padrões, é essencial para garantir uma gestão eficaz dos dados e fortalecer a transparência e confiabilidade da ciência (Davidson *et al*., 2017).

REFERÊNCIAS

ARAKAKI, Felipe Augusto; SANTOS, Plácida Leopoldina Ventura Amorim da Costa. Proveniência e contexto digital: contribuições da ciência da informação**. Palabra Clave (La Plata)**, La Plata, v. 10, n. 2, 1 abr. 2021. 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.

ALVES-MAZZOTTI, A. J. A "revisão bibliográfica" em teses e dissertações: meus tipos inesquecíveis - o retorno. In: BIANCHETTI, L.; MACHADO, A. M. N. (org.). **A bússola do escrever:** desafios e estratégias na orientação de teses e dissertações. Florianópolis: Editora da UFSC, 2002. p. 25-44.

BRANDT, Mariana Baptista; VIDOTTI, Silvana Aparecida Borsetti Gregório; SANTOS, Plácida Leopoldina Ventura Amorim da Costa; ZAFALON, Zaira Regina. Catalogação de metadados: descrição de metadados de negócio a partir dos princípios e objetivos bibliográficos. **Perspectivas em Ciência da Informação**, Belo

Horizonte, v. 24, n. 3, p. 3-18, 2019. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/22627. Acesso em: 3 fev. 2025.

DAVIDSON, S. B.; KHANNA, S.; MILO, T.; ROY, S.; TANNEN, V. Provenance views for module privacy. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2011, Athens, Greece. **Proceedings...** New York: ACM, 2011. p. 557–568. DOI: 10.1145/1989284.1989305. Disponível em: https://dl.acm.org/doi/10.1145/1989284.1989305. Acesso em: 11 mar. 2025.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 6. ed. São Paulo: Atlas, 2017.

GO FAIR. FAIR principles. [2016]. Disponível em: https://www.go-fair.org/fair-principles/.

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

HEDSTROM, M. **Digital preservation:** a time bomb for digital libraries**.** Computers and the humanities, Países Baixos, V31, mai 1997.

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).

LEBO, Timothy; SAHOO, Satya; MCGUINNESS, Deborah. **PROV-O: The PROV Ontology**. Disponível em: https://www.w3.org/TR/2013/REC-prov-o-20130430/. Acesso em: 16 dez. 2024.

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

MOREAU, L. *et al*. The Open Provenance Model core specification. **Future Generation Computer System***s*, Amsterdã, v. 27, n. 6, p. 743-756, 2011. Disponível em: https://doi.org/10.1016/j.future.2010.07.005. Acesso em: 11 mar. 2025.

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

RE3DATA. **About - re3data.org**. Disponível em: https://www.re3data.org/about. Acesso em: 13 nov. 2024.

RE3DATA. **Metadata Standards - re3data.org Metrics**. Disponível em: https://www.re3data.org/metrics/metadataStandards. Acesso em: 7 nov. 2024.

RUSBRIDGE, C. et al. O Centro de Curadoria Digital: uma visão para a curadoria digital. In: SIMPÓSIO INTERNACIONAL IEEE SOBRE SISTEMAS E TECNOLOGIA DE ARMAZENAMENTO EM MASSA, 2005. **Anais** [...]. [S.l.]: IEEE, 2005. DOI: 10.1109/LGDI.2005.1612461. Acesso em: 11 mar. 2025.

SANCHEZ, F. A.; DA SILVA, N. B. P.; VECHIATO, F. L. Padrões de metadados para representação e organização da informação em repositórios de dados de pesquisa. **Informação & Tecnologia,** João Pessoa: UFPB, v. 5, n. 1, p. 37–51, 2019. DOI: 10.22478/ufpb.2358-3908.2018v5n1.38350. Disponível em: https://periodicos.ufpb.br/index.php/itec/article/view/38350. Acesso em: 11 mar. 2025.

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. DOI: https://doi.org/10.5433/1981-8920.2016v21n2p90Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/27939. Acesso em: 11 mar. 2025.

SILVA, N. R. Diretórios de dados abertos de pesquisas: o caso do Re3Data. **Ciência da Informação Express**, Lavras, v. 2, n. 5, p. 1-4, 3 maio 2021. **DOI:** https://doi.org/10.60144/v2i.2021.84. Disponível em: https://cienciadainformacaoexpress.ufla.br/index.php/revista/article/view/84. Acesso em: 11 mar. 2025.

VIDOTTI, S. A. B. G *et al.* Repositório De Dados De Pesquisa Para Grupo De PesquisA: um projeto piloto*. In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017. **Anais**...Marília - SP. Anais... Marília - SP: PPGCI, UNESP. Disponível em: http://enancib.marilia.unesp.br/index.php/xviiienancib/ENANCIB/paper/view/388. Acesso em: 11 mar. 2025.

Ciência da Informação Express
ISSN 2675-8717

Notas e créditos do artigo

• **Reconhecimentos:** Não se aplica.

• **Financiamento:** Não se aplica.

• **Conflitos de interesse:** Não se aplica.

• **Aprovação ética:** Não se aplica.

• **Disponibilidade de dados e material:** Os conjuntos de dados analisados durante o presente estudo estão disponíveis no Re3data, disponível no link: https://www.re3data.org/metrics/metadataStandards https://www.re3data.org/search?query=PROV

• **Manuscrito publicado como *preprint*:** o manuscrito foi originalmente publicado como trabalho completo no Seminário Nacional de Catalogação e Tecnologia (SNCat), em 2024. Posteriormente, passou por nova avaliação *double-blindpeer review*, além de receber ajustes e atualizações de conteúdo.

.

•Contribuições dos autores:

Contribuição	Silva, F. I.	Arakaki, F. A.
Concepçãodo estudo	X	X
Conceitualização	X	X
Metodologia	X	
Coleta de dados / investigação	X	
Curadoria de dados	X	
Análisedos dados	X	
Discussão dos resultados	X	X
Visualização (gráficos, tabelas e outros)	X	X
Rascunho original	X	
Revisão e edição final	X	X
Supervisão e administração		X

• Divulgação Científica
Autoriza *post* no Ciexpress: divulgação científica?
(X) Sim.
( ) Não

• Publicador
Universidade Federal de Lavras (UFLA).
As ideias expressas neste artigo são de responsabilidade de sua autoria, não representando, necessariamente, a opinião dos editores ou da universidade.

Editor do canal de comunicação e divulgação científica Ciência da Informação Express
Nivaldo Calixto Ribeiro, Universidade Federal de Lavras (UFLA).

Revisor de linguística
Dos autores.

Revisor de referências
Dos autores.

• Histórico

Recebido em: 29/05/2023	Aceito em: 06/07/2023	Publicado em: