METADADOS DE PROVENIÊNCIA E PRINCIPIOS FAIR EM REPOSITÓRIOS DE DADOS

Felipe Ivo da Silva
Universidade Federal de São Carlos (UFSCar), Brasil | felipe_ivodasilva@hotmail.com
https://orcid.org/0009-0005-1379-4692

Felipe Augusto Arakaki
Universidade de Brasília (UnB), Brasil | felipe.arakaki@unb.br
https://orcid.org/0000-0002-3983-2563

DOI: *10.22477/xiv.biredial.411*
EJE TEMÁICO: *Datos abiertos*

RESUMEN

A pesquisa aborda a importância dos metadados de proveniência aliados aos princípios FAIR na curadoria digital de repositórios de dados. O problema reside na necessidade de garantir a confiabilidade, rastreabilidade e reutilização dos dados em um cenário de produção massiva de informações. Justifica-se pela crescente demanda por transparência e interoperabilidade na ciência aberta. O objetivo é analisar como a integração entre metadados da Família PROV e os princípios FAIR fortalece a gestão e a preservação dos dados científicos. Adotou-se uma abordagem teórico-exploratória, com levantamento bibliográfico e revisão sistemática de literatura, priorizando fontes da Ciência da Informação e da Computação. Os resultados demonstram que repositórios que utilizarem metadados de proveniência e aplicam os princípios FAIR podem apresentar maior confiabilidade e potencial de reuso dos dados. O estudo destaca um modelo prático com metadados PROV e a organização dos princípios FAIR. Conclui-se que tais práticas ampliam a transparência, facilitam colaborações interdisciplinares e otimizam a curadoria digital. A pesquisa sugere que a adoção ampla dessas estratégias é essencial para o fortalecimento da ciência aberta e o avanço do conhecimento em repositórios de dados.

Palabras-clave: Metadados, Proveniência, Principios FAIR e Repositórios de dados.

ABSTRACT

The research addresses the importance of provenance metadata combined with the FAIR principles in the digital curation of data repositories. The problem lies in the need to ensure data reliability, traceability, and reusability in a scenario of massive information production. It is justified by the growing demand for transparency and interoperability in open science. The objective is to analyze how the integration between the PROV Family metadata and the FAIR principles strengthens the management and preservation of scientific data. A theoretical-explora-

tory approach was adopted, with a bibliographic survey and systematic literature review, prioritizing sources from Information Science and Computer Science. The results demonstrate that repositories that utilize provenance metadata and apply the FAIR principles can exhibit greater reliability and potential for data reuse. The study highlights a practical model using PROV metadata and the organization of the FAIR principles. It is concluded that such practices expand transparency, facilitate interdisciplinary collaborations, and optimize digital curation. The research suggests that the broad adoption of these strategies is essential for strengthening open science and advancing knowledge in data repositories.

Keywords: Metadata, Provenance, FAIR Principles and Data Repositories.

INTRODUÇÃO

A transformação digital tem impulsionado uma produção massiva de dados em diversos setores da sociedade, gerados tanto por indivíduos quanto por instituições. Esse cenário exige uma gestão cada vez mais eficiente e organizada das informações, pautada em abordagens orientadas por dados. Nesse contexto, surge a necessidade de estruturar a preservação de dados em repositórios de dados, de forma que sejam confiáveis, rastreáveis e reutilizáveis, garantindo sua integridade e valor ao longo do tempo.

A curadoria digital desempenha um papel fundamental nesse processo, como posto por Silva et al. (2021), que a curadoria digital pode ser entendida 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 envolvendo atividades de gestão, preservação e disseminação de dados para assegurar seu acesso contínuo e eficiente. Um dos aspectos essenciais dessa curadoria é a proveniência de metadados, que registra a origem, as transformações e o histórico de uso das informações, contribuindo para sua autenticidade e confiabilidade. A documentação adequada da proveniência, por meio de metadados específicos, como os da Família PROV, permite que pesquisadores compreendam a trajetória dos dados, facilitando sua interpretação e reutilização em novos estudos.

Além disso, os princípios FAIR (*Findable, Accessible, Interoperable, Reusable*) surgem como diretrizes fundamentais para aprimorar a gestão de dados em repositórios científicos FORCE11 (2016). Esses princípios visam tornar os dados mais fáceis de localizar, acessíveis em formatos adequados, interoperáveis entre diferentes sistemas e reutilizáveis em diversas pesquisas. A combinação entre metadados de proveniência e a aplicação dos critérios FAIR fortalece a transparência e a eficiência dos repositórios, alinhando-se aos objetivos da ciência aberta, que preza pela disponibilização ampla e colaborativa do conhecimento.

Dessa forma, este trabalho busca discutir a importância dos metadados de proveniência alinhada aos princípios FAIR em repositórios de dados, destacando seu papel na preservação, recuperação e interoperabilidade das informações científicas. Ao analisar essas práticas, pretende-se contribuir para a reflexão sobre estratégias que ampliem a confiabilidade e o impacto dos dados de pesquisa no ambiente acadêmico e além.

MÉTODOS

Este estudo adota uma abordagem teórico-exploratória, com o objetivo de analisar a relação entre metadados de proveniência e os princípios FAIR em repositórios de dados, considerando sua relevância para a curadoria digital e a ciência aberta. Para isso, foi conduzido um levantamento bibliográfico, complementado por uma revisão sistemática de literatura, a fim de consolidar uma base conceitual sólida e identificar tendências recentes no tema.

A pesquisa parte de uma perspectiva interdisciplinar, refletindo a natureza multifacetada da gestão de metadados em ambientes digitais. Como destacam Gava et al. (2024), a proveniência dos dados é essencial para garantir sua confiabilidade e reutilização, exigindo abordagens metodológicas que considerem tanto aspectos técnicos quanto gerenciais.

A seleção das fontes priorizou artigos científicos indexados em bases como *Scopus*, *Web of Science* e *Google Scholar*, com ênfase em publicações em torno da Ciência da Informação e Ciência da Computação, garantindo a atualidade das discussões. Conforme Gil (2019), pesquisas exploratórias permitem delimitar o objeto de estudo e formular novas interpretações a partir da síntese crítica do conhecimento existente. Além disso, seguindo Marconi e Lakatos (2003), a revisão de literatura é fundamental para mapear contribuições teóricas, identificar lacunas e estabelecer conexões entre diferentes perspectivas sobre metadados de proveniência e FAIR.

RESULTADOS

Os repositórios de dados têm se consolidado como infraestruturas fundamentais para a gestão do conhecimento científico na era digital. Esses repositórios incorporam funcionalidades avançadas de preservação e disseminação de dados, garantindo não apenas o armazenamento, mas também a acessibilidade e o reuso da informação ao longo do tempo (Martins et al., 2017). Essa capacidade é particularmente relevante no contexto da *e-Science*, onde o volume e a complexidade dos dados exigem abordagens sofisticadas de curadoria digital (Gray, 2009).

Um dos principais achados desta pesquisa refere-se ao papel crítico dos metadados de proveniência na garantia da confiabilidade dos dados científicos. Como destacam Silva et al. (2021), esses metadados, especialmente os baseados na Família PROV, permitem rastrear toda a trajetória dos dados, desde sua geração até eventuais transformações. Essa documentação detalhada é fundamental para validar pesquisas que utilizam dados secundários, além de facilitar sua interpretação correta por comunidades científicas diversas. A proveniência adequada também atende a um dos princípios centrais da ciência aberta, qual seja, a transparência metodológica (Gezelter, 2009).

Para exemplificar a aplicação prática, o Quadro 1 apresenta um modelo estruturado para organização de metadados de proveniência, da Família PROV, com ênfase nos componentes fundamentais que incluem a fonte original dos dados e os agentes participantes do processo.

Quadro 1 – Exemplo de metadados 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 apresenta metadados estruturados em formato legível por sistemas computacionais para o banco de dados das paradas de transporte coletivo, incorporando detalhes de proveniência. Os campos *dct:creator*, *dct:publisher* e *dct:issued,* documentam respectivamente a autoria, a entidade publicadora e a data de disponibilização do conjunto de dados. Já o elemento *prov:actedOnBehalfOf* especifica que o usuário que realizou as operações como representante autorizado da Secretaria Municipal de Mobilidade de *MyCity*.

A pesquisa demonstra que, com o uso dos princípios FAIR alinhada aos usos de metadados de proveniência, nos repositórios de dados, demonstra um aumento efetivo na garantia

dos dados. A aplicação desses princípios mostra-se particularmente relevante na classificação dos repositórios - sejam institucionais, temáticos ou multidisciplinares (Sayão e Sales, 2020). Repositórios que adotam essa abordagem demonstram maior capacidade de integrar dados de diferentes fontes e disciplinas, potencializando colaborações científicas e o avanço do conhecimento.

Para exemplificar os princípios FAIR, no Quadro 2, pode-se observar os requisitos norteadores que fundamentam a abordagem dos princípios FAIR.

Quadro 2 – Apresentação dos princípios 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 persistente; •F2 - os dados estão descritos com meta dados ricos; •F3 - os metadados estão registrados ou indexados em um recurso pesquisável; •F4 - os metadados especificam o identificador de dados.	•A1 - os metadados são recuperáveis pelo seu identificador usando um protocolo de comunicação pa dronizado; – A1.1 - o protocolo é aberto, gratuito e universalmente implementável; – A1.2 - o protocolo permite um procedi mento de autenticação e autorização, quando necessário; •A2 - os metadados estão acessíveis, mesmo quando os dados não estão mais disponíveis.	•I1 - os metadados usam uma linguagem formal, acessível, compartilhada e amplamente 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.	•R1 - os metadados tem uma pluralidade de atributos precisos e relevantes; – R1.1 - os metadados são 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 rele vantes ao domínio.

Fonte: Silva, 2025; baseado em FORCE11, 2016.

Em síntese, os resultados desta pesquisa evidenciam que a combinação entre metadados de proveniência e a adoção consistente dos princípios FAIR representa o caminho mais promissor para o desenvolvimento de repositórios de dados verdadeiramente eficazes. Essas práticas não apenas auxiliam na preservação técnica dos dados, mas também ampliam seu potencial de reuso, fomentam novas descobertas científicas e fortalecem os ideais da ciência aberta. A superação dos desafios identificados exigirá esforços coordenados entre pesquisadores, instituições científicas e desenvolvedores de infraestrutura digital, em um movimento que já se mostra fundamental para o futuro da pesquisa em todas as áreas do conhecimento.

CONCLUSÕES

Este estudo evidenciou o papel fundamental dos metadados de proveniência e dos princípios FAIR na gestão eficaz de repositórios de dados. A análise demonstrou que a documentação da proveniência, em especial da Família PROV, constitui um pilar essencial para garantir a confiabilidade, reprodutibilidade e reutilização dos dados, em consonância com os ideais da ciência aberta.

A investigação revelou que a aplicação consistente dos princípios FAIR eleva significativamente a funcionalidade dos repositórios, transformando-os em plataformas dinâmicas que não apenas armazenam, mas também potencializam o valor dos dados. Essas práticas facilitam a descoberta de informações, fomentam colaborações interdisciplinares e otimizam recursos de pesquisa, evitando duplicações.

Contudo, identificamos desafios persistentes que demandam atenção: a heterogeneidade de padrões entre diferentes áreas do conhecimento, as complexidades técnicas inerentes ao manejo de dados diversificados e a necessidade de ampliar a adoção dessas boas práticas na comunidade acadêmica. Estas limitações apontam para direções promissoras em pesquisas futuras, incluindo estudos empíricos que avaliem implementações concretas.

O estudo em questão estabelece bases sólidas para compreender como os metadados de proveniência e a rigorosa aplicação dos princípios FAIR pode revolucionar a gestão de dados nos repositórios de dados. A materialização desse potencial exigirá esforços coordenados entre pesquisadores, instituições e desenvolvedores, mas promete construir um ambiente de conhecimento mais transparente, interoperável e impactante para o avanço científico.

BIBLIOGRAFÍA

FORCE11. (2016). *The FAIR data principles*. https://www.force11.org/group/fairgroup/fairprinciples

Gava, T. B. S., Flores, D., Aleixo, D. V. B. S., Cristovão, H. M., Ferrari, L. I., & Moraes, M. F. (2024). Dados de pesquisa na Arquivologia: uma reflexão. *Em Questão*, 30(1), 1-29. https://doi.org/10.1590/1808- 5245.30.135857

Gil, A. C. (2019). *Como elaborar projetos de pesquisa* (6a ed.). Atlas.

Gray, J. (2009). Jim Gray on eScience: A transformed scientific method. *In* T. Hey, S. Tansley, & K. Tolle (Eds.), *The fourth paradigm:* Data-intensive scientific discovery (pp. 17-31). Microsoft Research.

Gezelter, D. (2009). *What, exactly, is open science?* The OpenScience Project. http://www.openscience.org/blog/?p=269

Marconi, M. A., & Lakatos, E. M. (2003). *Fundamentos de metodologia científica* (5a ed.). Atlas. https://ria.ufrn.br/123456789/1239

Martins, D. L., Silva, M. F., Santarem Segundo, J. E., & Siqueira, J. (2017). 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*, 18. https://brapci.inf.br/index.php/res/download/125134

Sayão, L. F., & Sales, L. F. (2020). Afinal, o que é dado de pesquisa? *Biblos,* 34(2), 1-20. https://doi.org/10.14295/biblos.v34i2.11875

Silva, A. M. S., Silva, F. C., Santos, P. M., & Oliveira, R. F. (2021). Curadoria digital e arquivologia: olhares sobre o documento arquivístico digital. *Revista Ibero-Americana de Ciência da Informação,* 14(2), 567- 582. https://doi.org/10.26512/rici.v14.n2.2021.37558

Silva, F. I. (2025). *Proveniência de dados e metadados em repositórios de dados de pesquisa* [Dissertação de mestrado, Universidade Federal de São Carlos]. Repositório Institucional UFSCar. https://repositorio.ufscar.br/items/8076ed80-45a0-4641-9ac6-9ee3986f18aa

World Wide Web Consortium. (2013). *PROV model primer.* https://www.w3.org/TR/2013/NOTE-prov-primer 20130430/

ANEXO 1

RESUMEN BIOGRÁFICO DE LOS AUTORES

Felipe Ivo da Silva

Doutorado em andamento em Ciência da Computação, Universidade Federal de São Carlos, (UFSCar). Mestre em Ciência da Informação pela Universidade Federal de São Carlos, (UFSCar). Pós Graduado em Data Science pela Universidade Anhembi Morumbi - São Paulo (2024). Graduado em em Banco de Dados pelo Centro Universitário Estácio Ribeirão Preto (2022). Pesquisador bolsista do Instituto Brasileiro de Informação em Ciência e Tecnologia, IBICT, do Ministério da Ciência, Tecnologia e Inovação (MCTI). Membro do grupo de pesquisa Dados e Metadados (GPDM), Laboratório de Organização e tratamento da informação e Laboratório de estudos e práticas em Organização da Informação e Tecnologias (LOITec). Tem interesse de pesquisa nos temas: Dados, Metadados, Proveniência, Repositórios digitais, Preservação digital, Banco de Dados.

Felipe Augusto Arakaki

Docente do curso de Biblioteconomia da Universidade de Brasília (UnB) e do Programa de Pós-Graduação em Ciência da Informação (PPGCIN) da UNB. Doutor e mestre em Ciência da Informação pela Universidade Estadual Paulista "Júlio Mesquita Filho" - UNESP/Marília e bacharel em Biblioteconomia pela Universidade Estadual Paulista "Júlio de Mesquita Filho" - Campus de Marília. Vice-líder do Grupo de Pesquisa "Dados e Metadados e integrante do Grupo de Pesquisa "Novas Tecnologias em Informação". Coordenador do Grupo de Trabalho de Calogação da FEBAB. Áreas de interesse incluem a Ciência da Informação, principalmente nos temas: Representação e organização da informação, Catalogação, Metadados, Interoperabilidade, Padrões de Metadados, Dublin Core, BIBFRAME, Schema.org, Web Semântica, Linked Data e proveniência dos dados (PROV).