Em Busca de uma Infraestrutura Aberta para Experimentos de Aprendizado Profundo: Integração da `DLProv` e Dataverse*

Débora Pina[1], Liliane Kunstmann[3], Marta Mattoso[1], Marcos Lage[2], Daniel de Oliveira[2]

[1] Universidade Federal do Rio de Janeiro (PESC/COPPE/UFRJ)

[2]Instituto de Computação – Universidade Federal Fluminense (UFF)

[3]Mendelics Análise Genômica

{dbpina,marta}@cos.ufrj.br, {mlage,danielcmo}@ic.uff.br,

liliane.kunstmann@mendelics.com.br

Abstract. *Open Science demands auditability and reuse, which brings challenges to data management in deep learning (DL) training dataflows. To address this, it is important to capture provenance data, such as executed transformations and the execution environment. However, capturing alone is not enough. This information must also be made available. The paper proposes integrating the DLProv tool, which collects provenance data in DL, with the Dataverse repository. The approach automates the publication of data, models, and metadata. Its feasibility was demonstrated through the training of a CNN based on the AlexNet architecture.*

Resumo. *A Ciência Aberta exige auditoria e reúso, o que traz desafios à gestão de dados em dataflows de treinamento de Aprendizado Profundo (AP). Para isso, é fundamental capturar dados de proveniência, como transformações executadas e ambiente de execução. Contudo, apenas capturar não basta, é preciso disponibilizar essas informações. O artigo propõe integrar a ferramenta DLProv, que coleta dados de proveniência em AP, ao repositório Dataverse. A proposta automatiza a publicação de dados, modelos e metadados. A viabilidade foi demonstrada com o treinamento de uma CNN baseada na AlexNet.*

1. Introdução

Nos últimos anos, cresceu o interesse por práticas da chamada Ciência Aberta [Waskita et al. 2023], cujo objetivo é tornar os resultados de pesquisas acessíveis, transparentes e reprodutíveis. Em muitas instituições e agências de fomento, a publicação aberta de dados, código-fonte e metadados já é mandatória ou fortemente incentivada. O Governo Federal, por exemplo, lançou o 6º Plano de Ação em Governo Aberto[1], com diretrizes para fomentar práticas de ciência aberta. Tais iniciativas estão alinhadas com os princípios FAIR [Wilkinson 2016] (*Findable, Accessible, Interoperable, Reusable*), que orientam a gestão de dados para permitir seu reúso por outros pesquisadores.

No entanto, a implementação efetiva da Ciência Aberta enfrenta desafios. Para [Demchenko et al. 2012], o maior obstáculo é a gerência dos metadados científicos, que vão desde atributos simples, *e.g.*, título, licença e palavras-chave, até complexos caminhos de

*Os autores gostariam de agradecer pelo apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código de Financiamento 001, do CNPq e da FAPERJ.

[1]https://www.gov.br/cgu/pt-br/governo-aberto/noticias/2024/01/6-plano-de-acao-nacional-ogp-final.pdf

derivação dos dados, representados por grafos de proveniência [Herschel et al. 2017]. Esses desafios são comuns em domínios como Biologia, mas vêm se intensificando também na área de Inteligência Artificial (IA), especialmente no contexto do Aprendizado Profundo (AP) [Ravi et al. 2022].

Treinamentos de modelos de AP são *dataflows* compostos por etapas de pré-processamento, treinamento, avaliação e seleção de modelos, executadas com diferentes *frameworks*, tanto localmente quanto em ambientes distribuídos. Os modelos resultantes podem ser aplicados em domínios sensíveis, *e.g.*, reconhecimento de padrões em imagens dermatológicas [Blanco et al. 2020]. Dado o potencial impacto ético dessas aplicações, é essencial que esses modelos sejam acessíveis para auditoria e explicação. Por exemplo, se um modelo falha sistematicamente em uma identificação, é necessário investigar se há viés.

Para a auditoria e explicação desses modelos, é fundamental o acesso aos dados de proveniência associados aos dados e modelos treinados, *i.e.*, os registros detalhados das transformações aplicadas, parâmetros utilizados e ambiente computacional das execuções dos *dataflows* de treinamento [Pina et al. 2025]. Os dados de proveniência podem fornecer o contexto necessário para interpretar resultados, identificar vieses e replicar experimentos em diferentes cenários. A captura de dados de proveniência em *dataflows* de AP já é explorada por diversas abordagens [Pina et al. 2025, Schlegel and Sattler 2023]. Uma delas é a `DLProv` [Pina et al. 2025], um serviço compatível com o padrão W3C PROV [Moreau and Groth 2013], que permite consultas sobre decisões tomadas durante o treinamento e representa as relações entre artefatos, mesmo em ambientes distribuídos.

No entanto, a captura da proveniência por si só não garante a auditoria e a explicação de um modelo de AP. É igualmente necessário que os dados brutos e intermediários, modelos, parâmetros e hiperparâmetros estejam disponíveis para consulta. Soluções como a `DLProv` armazenam o caminho os dados e modelos utilizados ou produzidos pelo *dataflow* de treinamento estão armazenados, sem gerenciar seu armazenamento e publicação. Atualmente, toda responsabilidade pelo armazenamento e disponibilização dos dados depende do usuário que executa o *dataflow* de treinamento. Ainda, mesmo em relação aos dados de proveniência capturados pela `DLProv`, o acesso também depende de autorização para submissão de consultas no banco de dados de proveniência. Por outro lado, repositórios de dados públicos como o Dataverse [Crosas 2011] permitem a publicação de dados e metadados de pesquisa estruturados em comunidades hierárquicas chamadas *dataverses*. O Dataverse oferece API REST, suporte a ontologias e atribuição de DOI aos conjuntos de dados. Tais funcionalidades se mostram complementares a soluções como a `DLProv`, no entanto, não encontramos na literatura abordagens que integram grafos de proveniência com repositórios FAIR.

Este artigo propõe integrar os dados de proveniência capturados pela `DLProv` à capacidade de armazenamento e publicação de dados do Dataverse, com o objetivo de fomentar a reprodutibilidade. A proposta visa a publicação de modelos, dados e metadados de proveniência de *dataflows* de AP em conformidade com os princípios FAIR. Nessa proposta, a DLProv monitora a execução do *dataflow* de AP e, a cada transformação executada, os dados e metadados no padrão PROV são automaticamente enviados a um *dataverse*. A publicação dos dados de proveniência no padrão W3C PROV auxilia a interoperabilidade com outras ferramentas de análise, a visualização gráfica dos *dataflows* e a verificação de sua reprodutibilidade. A abordagem foi avaliada com a execução de um *dataflows* de treinamento de uma Rede Neural Convolucional utilizando a arquitetura AlexNet e o *dataset Oxford Flowers* [Nilsback and Zisserman 2006]. Os resultados evidenciaram a viabilidade e potencial da integração proposta.

2. Referencial Teórico

`DLProv` [Pina et al. 2025]. É uma ferramenta que oferece rastreabilidade em *dataflows* de treinamento de AP tratando dados de proveniência como prioridade. DLProv oferece serviços de proveniência para capturar e exportar proveniência prospectiva (*p-prov*) e proveniência retrospectiva (*r-prov*) de forma independente de *framework* de AP, por meio de instrumentação de *scripts*. A `DLProv` adota um modelo de dados que segue o padrão W3C PROV [Moreau and Groth 2013], é extensível, e representa de forma explícita as diferentes etapas dos *dataflows* de treinamento de AP, *e.g.*, a divisão de conjuntos de dados, o treinamento dos modelos de AP e sua avaliação, bem como os relacionamentos existentes entre essas etapas. Durante o treinamento de modelos de AP, os dados de proveniência são persistidos em um banco de dados, podendo ser consultados em tempo real. Ao final do treinamento, a `DLProv` permite a geração de documentos de proveniência com base nos dados capturados e persistidos. Esses documentos, que podem ser gerados em formatos como JSON e PROV-N, incluem entidades, atividades e agentes, que representam dados, transformações realizadas durante a execução do *dataflow*, e os atores responsáveis por essas transformações. Uma vez gerados, os documentos de proveniência podem ser armazenados em bancos de dados de grafos (*e.g.*, Neo4J). Com isso, usuários podem realizar análises que exploram o caminho de derivação de dados em *dataflows* de treinamento de AP, identificam dependências entre etapas, rastreiam a origem de resultados e associam esses resultados a pré-processamentos específicos ou variações nos hiperparâmetros, facilitando a reprodutibilidade e o refinamento dos experimentos.

Dataverse [Crosas 2011]. É uma plataforma de repositório digital de dados de pesquisa desenvolvida por Harvard e que tem como objetivo oferecer uma infraestrutura escalável e interoperável para o armazenamento, publicação e citação de dados científicos, alinhado com os princípios FAIR. O diferencial do Dataverse é sua estrutura hierárquica e modular. Toda organização dentro do repositório se encontra em torno do conceito de comunidade (*dataverse*), que funciona como um contêiner de dados e metadados. Dentro de cada comunidade, é possível criar novas comunidades, criando uma hierarquia. Além de novas comunidades, o usuário pode ciar conjuntos de dados (*datasets*), que de fato armazenam os dados científicos. Essa estrutura facilita a administração do repositório, pois permite que diferentes comunidades possam ter diferentes permissões de acesso, publicação e colaboração. Uma das grandes vantagens do uso do Dataverse é que ele atribui automaticamente identificadores persistentes (DOI) a cada conjunto de dados publicado no repositório. Além disso, cada conjunto de dados pode ser versionado. O Dataverse oferece também apoio a um grande conjunto de metadados para descrever os conjuntos de dados, com campos específicos para diferentes domínios como Astronomia e Biologia. Esses metadados são compatíveis com padrões como Dublin Core, *etc*, o que viabilizam a interoperabilidade com catálogos e serviços externos. A plataforma também disponibiliza uma API RESTful que permite que sejam realizados *uploads*, atualizações, consultas a metadados e publicação de conjuntos de dados, tudo via *script*.

3. Trabalhos Relacionados

Para enfrentar os desafios impostos pela Ciência Aberta no contexto do AP, [Dalgali and Crowston 2019] apresentam um estudo que busca compreender os motivos, formas e perfis dos agentes que compartilham (ou deixam de compartilhar) modelos de AP com vistas ao reúso. Os autores identificam que a motivação mais recorrente para o compartilhamento está associada à promoção da colaboração no desenvolvimento de novos modelos. No mesmo contexto, [Li et al. 2022] discutem cenários em que o compartilhamento dos dados não é viável, e propõem uma estratégia denominada MSS, que permite a

disseminação dos modelos treinados juntamente com metadados descritivos sobre os dados utilizados no treinamento. Dessa forma, mesmo sem acesso direto aos dados originais, outros pesquisadores podem compreender, adaptar e aprimorar os modelos compartilhados. Complementarmente, [Flemisch et al. 2024] discutem um conjunto abrangente de desafios relacionados à implementação da Ciência Aberta e destacam o uso do Dataverse como uma alternativa promissora para o compartilhamento escalável de dados científicos, em consonância com os princípios FAIR.

[Schackart III et al. 2024] abordam os desafios de Ciência Aberta para AP para dados biológicos. No trabalho, os autores apresentam um estudo sobre um inventário de fontes de dados biológicos da literatura científica. O trabalho objetiva a reprodutibilidade e a classifica em três níveis relacionados à facilidade de reprodução. Também se observa no trabalho que a disponibilização dos dados é importante para reprodutibilidade porém não é suficiente e que é preciso também de padronização de código e automação. [Kocak et al. 2023] realizam uma revisão sistemática de AP para aplicações de Radiologia e Medicina Nuclear, utilizando artigos do PubMed. O trabalho avalia a disponibilidade de dados, modelos, código e *software*. De modo geral, foi detectado um esforço ao longo do tempo para disponibilização de artefatos produzidos, porém, ainda muito baixo, o que pode inviabilizar a replicação de experimentos. Há uma grande heterogeneidade nos repositórios, metadados e padrões de dados disponibilizados. Por isso [Borges et al. 2021] propõem uma plataforma de metabusca semântica para repositórios de dados de ciência aberta. Essa metabusca envolve múltiplos passos: coleta de registros de metadados de múltiplos repositórios, a padronização desses registros em um modelo de metadados próprio, e a realização de buscas semânticas utilizando ontologias de domínio para enriquecer as consultas dos usuários.

4. Integração da `DLProv` com o Dataverse

Com o objetivo de viabilizar a publicação de dados, modelos e metadados de proveniência gerados por *dataflows* de treinamento de modelos de AP, esta seção apresenta a proposta de integração entre a ferramenta de captura de dados de proveniência `DLProv` e o Dataverse[2]. A principal motivação dessa integração é permitir que, assim que a `DLProv` capture informações de proveniência durante a execução dos *dataflows* de treinamento de modelos de AP, esses dados possam ser automaticamente organizados e publicados no Dataverse. Para isso, a abordagem proposta é responsável por estruturar comunidades e conjuntos de dados na plataforma, possibilitando que os dados científicos fiquem disponíveis publicamente, promovendo seu reúso e validação por outros usuários.

A arquitetura da solução é apresentada na Figura 1 e se encontra organizada em cinco camadas: (i) Treinamento; (ii) Dados; (iii) Integração; (iv) Análise; e (v) Publicação. É importante ressaltar que as camadas de (i) a (iv) já existem na `DLProv` original, com exceção do componente *Eavesdrop*. A execução se inicia na *Camada de Treinamento*, onde o *dataflow* de treinamento do modelo de AP é iniciado com a definição da arquitetura da rede neural e de seus hiperparâmetros pelo usuário. Durante o treinamento, os dados brutos, os pesos e os modelos são processados por uma *Biblioteca de Treinamento* e identificados em tempo real pelo *Capturador de Proveniência*. Esses dados seguem para a *Camada de Dados*, onde são armazenados em um sistema de arquivos e seus metadados organizados por um *Banco de Proveniência*. A *Camada de Integração* é responsável por processar a proveniência do pré-processamento (caso o usuário importe dados de proveniência capturados em ferramentas externas) e integrá-la com

[2]`https://github.com/UFFeScience/dlprov_dataverse`

a proveniência do treinamento do modelo de AP persistida no *Banco de Proveniência* por meio de um *Mapeador de Proveniência*. Em seguida, a *Camada de Análise* permite ao usuário gerar os documentos de proveniência através do *Exportador de Proveniência* além de visualizar os grafos de proveniência gerados, por meio de um *Visualizador de Proveniência*.

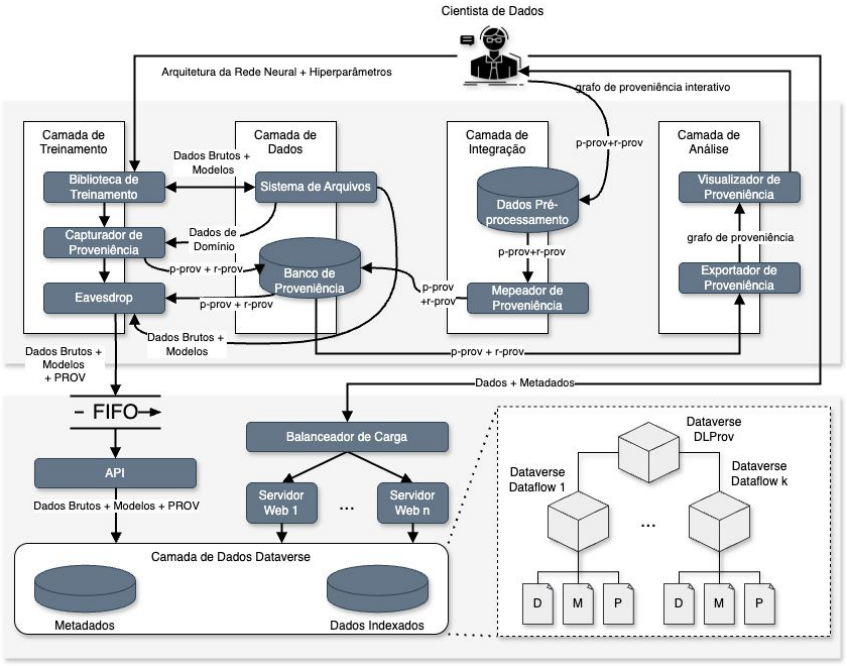

Figura 1. A arquitetura proposta para a integração da `DLProv` com o Dataverse

O *Eavesdrop*, principal contribuição deste artigo, é responsável por interceptar quando novos metadados de proveniência são capturados e armazenados no *Banco de Proveniência*, atuando como um *trigger* no banco de dados. A partir dessas informações, o *Eavesdrop* envia os dados e metadados para a *Camada de Publicação*, representada pelo Dataverse. O *Eavesdrop* envolve uma sequência estruturada de atividades: (i) consulta o *Banco de Proveniência* para gerar um documento compatível com o W3C PROV, contendo os metadados referentes ao *dataflow* de treinamento do modelo (incluindo as etapas de pré-processamento, quando especificadas); (ii) com base nos metadados de proveniência, identifica os arquivos consumidos e produzidos durante a execução do *dataflow*, acessando-os diretamente no sistema de arquivos local; (iii) os arquivos identificados, juntamente com seus metadados, são organizados e carregados em uma comunidade no Dataverse.

Para essa operação, assume-se que um usuário esteja registrado no Dataverse para a `DLProv`. A estrutura de organização dos dados no repositório é definida da seguinte forma: para cada *dataflow* distinto de treinamento, é criada uma nova comunidade no Dataverse. Dentro dessa comunidade, é criada uma subcomunidade para cada execução do *dataflow*, para representar diferentes execuções do mesmo *dataflow* com variações em parâmetros. Em cada subcomunidade, é criado um conjunto de dados, no qual os arquivos são organizados em pastas distintas, conforme a Figura 2, classificadas em: dados de entrada (D), modelos treinados (M), pesos (W), e documento de proveniência (P). Uma vez publicados, tanto o conjunto de dados quanto os arquivos individuais recebem identificadores persistentes (DOI) e URLs de acesso, atribuídos pelo próprio Dataverse. Essa funcionalidade viabiliza a rastreabilidade, encontrabilidade e o reúso dos dados por terceiros, em conformidade com os princípios FAIR.

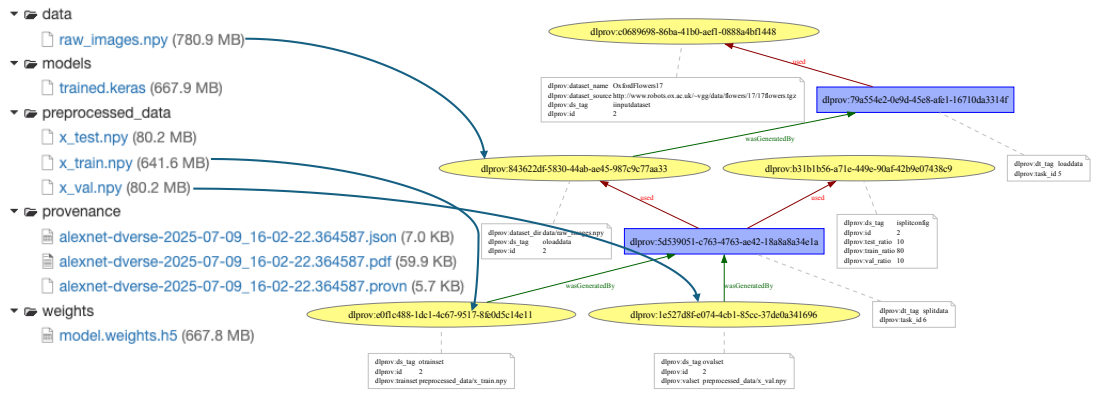

Figura 2. Exemplo de *dataset* disponibilizado no Dataverse pela `DLProv`.

5. Avaliação Experimental

Para avaliar a integração entre a `DLProv` e o Dataverse, executamos um *dataflow* para treinamento de uma CNN com a AlexNet [Krizhevsky et al. 2012], aplicada ao reconhecimento de imagens. Utilizamos dados do Oxford Flowers [Nilsback and Zisserman 2006], para imagens com 17 categorias de flores. Os experimentos foram executados no Apple M2 Pro, com 16GB de RAM e macOS 15.5. A sobrecarga computacional da `DLProv` para a captura dos dados de proveniência durante o treinamento foi avaliada em [Pina et al. 2025, Pina et al. 2024]. Assim, neste artigo, tomamos como *baseline* o tempo total de execução do *dataflow* de treinamento com a captura de proveniência ativada. A partir disso, o experimento avalia a sobrecarga imposta pelas etapas de preparação e publicação dos dados de proveniência, dados de entrada, pesos e modelos treinados na comunidade Dataverse. Além da análise de desempenho, o experimento avalia qualitativamente a integração entre a `DLProv` e o Dataverse. Buscou-se mostrar como os artefatos gerados durante a execução do *dataflow* podem ser organizados de maneira estruturada, versionados automaticamente e compartilhados de forma transparente.

A Figura 3 apresenta o tempo de execução (em segundos), o desvio padrão da execução dos *dataflows* sem e com a publicação de dados, modelos e proveniência no Dataverse. Esses valores foram obtidos a partir da média de cinco execuções de treinamento. Os resultados indicam uma sobrecarga de aproximadamente 41,4% para execuções com 10 épocas, 16,4% para 20 épocas e 6,4% para 50 épocas. Esses números revelam um impacto proporcionalmente maior no tempo total de execução em *dataflows* mais rápidos, sugerindo que o custo relativo da publicação de dados e metadados tende a diluir à medida que a duração da execução aumenta. A sobrecarga mais acentuada observada em execuções de menor tempo se deve ao tempo necessário para a ingestão dos arquivos no Dataverse, especialmente considerando que vários dos arquivos manipulados apresentam tamanhos entre 500 MB e 1 GB.

Entretanto, essa sobrecarga pode ser mitigada com a paralelização do *Eavesdrop*. Por exemplo, múltiplas *threads* podem realizar o carregamento dos arquivos concorrentemente. Além disso, tarefas de publicação, como o envio dos dados pré-processados para o Dataverse, podem ser realizadas em paralelo com o treinamento do modelo de AP, o que contribuiria para reduzir o tempo total da execução. Cabe ressaltar que muitos treinamentos, especialmente aqueles com modelos mais complexos, demandam várias horas ou mesmo dias para finalizar. Nesses casos, a sobrecarga da publicação dos artefatos científicos torna-se relativamente menor frente ao tempo total do treinamento, reforçando a viabilidade prática da abordagem proposta.

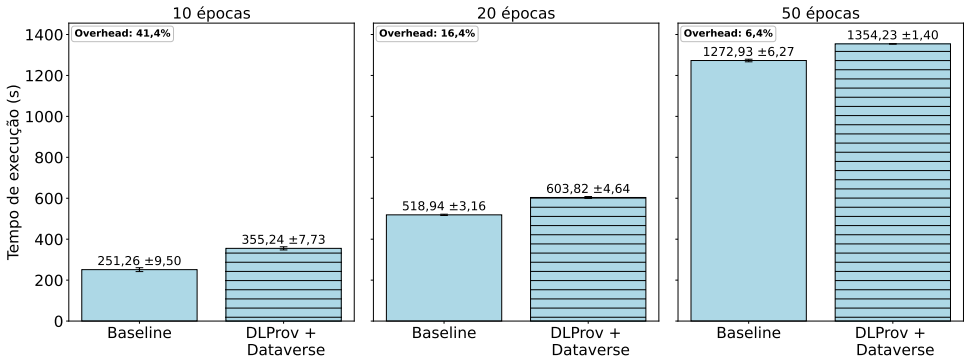

Figura 3. Tempo médio de execução entre Baseline e `DLProv` + Dataverse.

Para a avaliação qualitativa, a Figura 2 apresenta um exemplo de *dataset* denominado `alexnet-dverse-2025-07-08_16:34:45`, criado durante a execução do *dataflow* de treinamento da AlexNet. Esse *dataset* foi armazenado em uma comunidade específica para essa execução criada sob a comunidade raiz da `DLProv`, obedecendo à estrutura hierárquica descrita na Seção 4. Os arquivos com artefatos gerados na execução estão em quatro pastas: *data* com os dados brutos de entrada (`raw_images.npy`), *models* com o modelo treinado (`trained.keras`), *weights* com os pesos do modelo e *preprocessed_data* com os conjuntos de dados das fases de treinamento, validação e teste (`x_train.npy`, `x_val.npy` e `x_test.npy`). Os documentos de proveniência apontam para os artefatos e ficam no diretório denominado *provenance*, em múltiplos formatos complementares (JSON e PROVN) para atender diferentes necessidades de interpretação e reúso. Essa organização reflete o compromisso com a preservação e acessibilidade dos artefatos, promovendo não apenas a persistência dos dados utilizados e gerados, mas também a auditoria de cada execução do *dataflow*. A criação automática de um *dataset* distinto para cada execução possibilita o isolamento, versionamento e compartilhamento individualizado dos resultados, promovendo maior transparência, auditoria e organização ao longo do ciclo de vida dos experimentos.

6. Conclusão

Este artigo apresentou uma abordagem para a publicação de dados, modelos e dados de proveniência oriundos de *dataflows* de treinamento de modelos de AP, a partir da integração da `DLProv` com o Dataverse. O objetivo principal foi permitir que modelos, dados e metadados de proveniência oriundos de *dataflows* de AP pudessem ser compartilhados em conformidade com os princípios FAIR, de forma automatizada. A arquitetura desenvolvida introduziu o componente *Eavesdrop*, responsável por monitorar em tempo real a geração de dados de proveniência, estruturá-los segundo o padrão W3C PROV, identificar os artefatos produzidos e publicá-los no Dataverse. Essa integração permite a organização hierárquica dos dados, o versionamento das execuções e a atribuição de identificadores persistentes, promovendo a rastreabilidade e o reúso. A avaliação experimental, realizada com o treinamento da AlexNet, mostrou a viabilidade da proposta. A avaliação quantitativa mostrou que a sobrecarga introduzida pelo processo de publicação tende a se diluir em *dataflows* mais longos, o que reforça a adequação da abordagem em cenários reais. A avaliação qualitativa evidenciou a capacidade da proposta de estruturar, versionar e tornar acessíveis os artefatos científicos de forma compatível com práticas de governança de dados. Como trabalhos futuros, pretende-se ampliar os experimentos com volumes maiores de dados, de modo a avaliar a escalabilidade da abordagem, bem como incorporar estratégias de paralelização no carregamento e publicação dos dados, modelos e proveniência.

Referências

Blanco, G. et al. (2020). A superpixel-driven deep learning approach for the analysis of dermatological wounds. *Computer Methods and Programs in Biomedicine*, 183:105079.

Borges, G. C., dos Reis, J. C., and Medeiros, C. B. (2021). Addressing search in scientific open data repositories: A semantic metasearch platform. In *BreSci*, pages 81–88. SBC.

Crosas, M. (2011). The dataverse network®: An open-source application for sharing, discovering and preserving data. *DLib Mag.*, 17(1/2).

Dalgali, A. and Crowston, K. (2019). Sharing open deep learning models. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Demchenko, Y. et al. (2012). Addressing big data challenges for scientific data infrastructure. In *CloudCom'12*, pages 614–617. IEEE.

Flemisch, B. et al. (2024). Research data management in simulation science: Infrastructure, tools, and applications. *Datenbank-Spektrum*, 24(2):97–105.

Herschel, M., Diestelkämper, R., and Ben Lahmar, H. (2017). A survey on provenance: What for? what form? what from? *VLDB J.*, 26(6):881–906.

Kocak, B. et al. (2023). Transparency in artificial intelligence research: a systematic review of availability items related to open science in radiology and nuclear medicine. *Academic Radiology*, 30(10):2254–2266.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105.

Li, Z., Mao, F., and Wu, C. (2022). Can we share models if sharing data is not an option? *Patterns*, 3(11).

Moreau, L. and Groth, P. (2013). Provenance: an introduction to prov. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 3(4):1–129.

Nilsback, M.-E. and Zisserman, A. (2006). A visual vocabulary for flower classification. In *CVPR'06*, volume 2, pages 1447–1454. IEEE.

Pina, D. et al. (2024). Dlprov: A data-centric support for deep learning workflow analyses. In *DEEM'24*, DEEM '24, page 77–85, New York, NY, USA. ACM.

Pina, D., Kunstmann, L., et al. (2025). Dlprov: a suite of provenance services for deep learning workflow analyses. *PeerJ Comp. Sci.*, 11:e2985.

Ravi, N. et al. (2022). Fair principles for ai models with a practical application for accelerated high energy diffraction microscopy. *Scientific Data*, 9(1):657.

Schackart III, K. E., Imker, H. J., and Cook, C. E. (2024). Detailed implementation of a reproducible machine learning-enabled workflow. *Data Science Journal*.

Schlegel, M. and Sattler, K.-U. (2023). Mlflow2prov: Extracting provenance from machine learning experiments. DEEM '23, New York, NY, USA. ACM.

Waskita, A. A. et al. (2023). Open science progress: A literature assessment of open access articles. In *IC3INA'22*, page 271–275, New York, NY, USA. ACM.

Wilkinson, M. D. o. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.