Arquitetura de Curadoria e Publicação de Dados Científicos Amazônicos: o Caso do DataMap/Amazon

Danielle S. Monteiro[1], Marcio Teixeira[2], Adriano Almeida[3], Alan James Calheiros[3], Janete Machicao[1], Pedro Pizzigatti Corrêa[1]

[1]Escola Politécnica – USP – São Paulo, SP – Brasil

[2]Faculdade de Tecnologia – UNICAMP – Campinas, SP – Brasil

[3]Instituto Nacional de Pesquisas Espaciais – INPE – São José dos Campos, SP – Brasil

dsmonteiro@usp.br, mjt@unicamp.br, adriano.almeida@inpe.br,

alan.calheiros@inpe.br, machicao@usp.br, pedro.correa@usp.br

Abstract. *This article describes the DataMap/Amazon curation framework, developed to mitigate the dispersion and heterogeneity of Amazonian environmental data. The solution integrates four functional components (Dictionary, Files, Datasets, and Products) into a progressive qualification pipeline (Bronze, Silver, and Gold layers). Applied to four research initiatives (LBA, AmazonFACE, ATTO-Campina, and CHUVA), the platform transformed raw collections into FAIR scientific assets. Results include the structuring of 183 datasets, 1.58 TB of organized files, and 165 generated DOIs, strengthening reproducibility and open access to science in the region.*

Resumo. *Este artigo descreve o modelo de curadoria do DataMap/Amazon, desenvolvido para mitigar a dispersão e heterogeneidade de dados ambientais amazônicos. A solução integra quatro componentes funcionais (Dicionário, Arquivos, Datasets e Produtos) a um pipeline de qualificação progressiva (camadas Bronze, Prata e Ouro). Aplicada a quatro iniciativas de pesquisa (LBA, AmazonFACE, ATTO-Campina e CHUVA), a plataforma transformou acervos brutos em ativos científicos FAIR. Os resultados incluem a estruturação de 183 datasets, 1,58 TB de arquivos organizados e 165 DOIs gerados, fortalecendo a reprodutibilidade e o acesso aberto à ciência na região.*

1. Introdução

A produção científica orientada por dados ampliou a relevância de infraestruturas capazes de garantir não apenas armazenamento, mas também organização, proveniência, governança semântica e publicação qualificada [Hey et al. 2009]. Em domínios como monitoramento atmosférico, ecologia e modelagem ambiental, a utilidade do acervo não decorre apenas da existência de arquivos digitais, mas de um fluxo explícito de gestão do ciclo de vida dos dados que reconstrua o contexto de produção, estabilize a semântica, controle suas versões e converta observações heterogêneas em ativos científicos reutilizáveis [Higgins 2008, Michener et al. 2012].

No contexto amazônico, os desafios de dispersão, fragmentação e heterogeneidade de dados são particularmente agudos devido à alta complexidade ecossistêmica e à

infraestrutura institucional limitada. Para uma região cuja relevância científica é central [Davidson et al. 2012], grande parte dos acervos permanece em estado bruto ou semiestruturado, reduzindo o potencial de reuso. Diferentemente de infraestruturas de propósito geral, o DataMap/Amazon é concebido como infraestrutura curatorial que incorpora um fluxo de qualificação progressiva integrado ao ciclo de vida completo dos dados.

O presente trabalho vincula-se ao escopo do DS4SG pelos seguintes eixos temáticos: (i) mudanças climáticas, dado o papel crítico da Amazônia no sistema climático global [Davidson et al. 2012]; (ii) acesso das populações à ciência, tecnologia e inovação, uma vez que dados FAIR ampliam o potencial de reuso por pesquisadores de países com menor infraestrutura técnica; (iii) sustentabilidade, pois dados ambientais qualificados são insumo para políticas de conservação e uso do solo.

O objetivo deste artigo é descrever o modelo de curadoria implementado no DataMap/Amazon e analisar seus resultados na qualificação, publicação e disponibilização de dados científicos amazônicos. A contribuição resume-se em três pontos: (i) uma abordagem integrada que separa semântica, materialidade física, unidade publicável e agrupamento analítico em componentes funcionais distintos; (ii) um pipeline progressivo de maturidade curatorial inspirado na arquitetura medalhão de data lakehouse [Armbrust et al. 2020], que conduz o dado bruto ao ativo científico aderente aos princípios FAIR [Wilkinson et al. 2016] e citável; (iii) a demonstração empírica desse modelo em quatro iniciativas estratégicas da pesquisa amazônica. A contribuição central reside na integração operacional e sistematização de componentes consolidados da literatura, e não na proposição de novos algoritmos ou técnicas de processamento.

O restante do artigo está organizado assim: a Seção 2 apresenta os fundamentos conceituais e trabalhos relacionados; a Seção 3, o desenho da pesquisa; a Seção 4, o modelo proposto e o ciclo de vida dos dados; a Seção 5, a arquitetura da plataforma; a Seção 6, o estudo de caso e os resultados; a Seção 7 discute achados e limitações; e a Seção 8 conclui o trabalho.

2. Trabalhos Relacionados e Fundamentos Conceituais

A literatura sobre ciência intensiva em dados sustenta que o valor científico dos dados é produzido ao longo de um ciclo de vida que envolve planejamento, coleta, descrição, preservação, descoberta, integração e reuso [Hey et al. 2009, Michener et al. 2012]. O modelo do Digital Curation Centre (DCC) [Higgins 2008] trata a curadoria como prática contínua e sistemática, não redutível ao depósito pontual de arquivos. No plano normativo, os princípios FAIR – Findable, Accessible, Interoperable, Reusable – compõem o enquadramento mais influente para publicação e reutilização de dados científicos, pressupondo metadados ricos, identificadores persistentes, protocolos padronizados e vocabulários controlados [Wilkinson et al. 2016, UNESCO 2021].

Enquanto repositórios de propósito geral como Zenodo [CERN e OpenAIRE 2013], Figshare [Thelwall and Kousha 2016] e Dataverse [King 2007] focam em custódia e descoberta, o DataMap/Amazon integra nativamente a curadoria progressiva como parte do fluxo de trabalho, harmonizando semântica e rastreabilidade.

O modelo PROV do W3C [Groth and Moreau 2013] consolidou o entendimento de que a confiabilidade de um objeto digital depende da possibilidade de re-

construir entidades, atividades e agentes de sua produção. No DataMap/Amazon, a proveniência é registrada em três momentos: na ingestão (Bronze), com metadados mínimos de origem, data e responsável; na transição para a Prata, documentando as transformações com referência ao curador; e na Ouro, onde o versionamento com DOI vincula cada estado publicável ao histórico precedente e integra-se à DataCite (metadados Dublin Core), garantindo rastreabilidade da cadeia de custódia e interoperabilidade [Fenner et al. 2019, DataCite Metadata Working Group 2021, Brickley et al. 2019, Data Citation Synthesis Group 2014].

Do ponto de vista do bem social, iniciativas como o DataONE [Michener et al. 2012] e o Zenodo [CERN e OpenAIRE 2013] demonstraram que reduzir barreiras à descoberta e ao reuso tem impacto mensurável na produção científica colaborativa. O DataMap/Amazon distingue-se ao orientar esse objetivo não apenas à custódia e à descoberta, mas à qualificação progressiva do dado antes da publicação – especialmente relevante onde a heterogeneidade dos acervos representa uma barreira estrutural ao reuso por pesquisadores com menor infraestrutura institucional.

3. Desenho da Pesquisa e Estratégia de Avaliação

O estudo adota um desenho aplicado e descritivo, que combina a apresentação dos componentes funcionais do modelo, a descrição da arquitetura que o sustenta e a análise de sua aplicação em um estudo de caso com dados de quatro iniciativas amazônicas. Essa escolha é coerente com a natureza do objeto – um modelo de integração de práticas curatoriais – e com a tradição de avaliar infraestruturas de sistemas de informação por meio de estudos de caso reais em escala [Wieringa 2014].

3.1. Questões de Pesquisa e Indicadores

A análise foi organizada em torno de três questões de pesquisa: QP1 – o modelo converte arquivos brutos em objetos científicos estruturados, descritos e versionados?; QP2 – a proposta viabiliza publicação persistente e citável por meio de DOI e estados formais de disponibilização?; QP3 – a plataforma consolida acervos heterogêneos em escala relevante? Os indicadores correspondentes são, respectivamente: componentes funcionais aplicados, metadados enriquecidos e datasets organizados (QP1); versões publicadas, DOIs gerados e objetos em estado encontrável (QP2); e volume em TB, número de arquivos e amplitude do catálogo (QP3).

3.2. Escopo Empírico e Limitações do Recorte

A unidade empírica compreende quatro iniciativas estratégicas: LBA, AmazonFACE, ATTO-Campina e CHUVA. O recorte é intencional, selecionado por representatividade de perfis curatoriais distintos (séries temporais contínuas, dados experimentais de longo prazo, acervo histórico de campanhas encerradas e sítio observacional especializado), cobertura de diferentes escalas de acervo (de 2 a 78 datasets) e disponibilidade efetiva de dados no período de referência, assegurada por acordos institucionais. Reconhece-se explicitamente que a restrição a quatro iniciativas, a ausência de outros tipos de dados (sensoriamento remoto, imagens de satélite, dados socioeconômicos) e a seleção não probabilística limitam a generalização das conclusões: os resultados são indicativos de viabilidade operacional em contextos similares de alta heterogeneidade, não generalizáveis por inferência estatística.

O foco da avaliação não é comparar o DataMap/Amazon com soluções algorítmicas por meio de experimentos controlados. As métricas apresentadas na Seção 6 são indicadores de atividade curatorial realizada (datasets estruturados, DOIs gerados, volumes organizados), e não de qualidade do resultado da curadoria nem de impacto longitudinal no reuso efetivo dos dados. Essa distinção é discutida na Seção 7.

4. Modelo de Curadoria Proposto

O modelo abstrai a curadoria de dados amazônicos em duas dimensões complementares: a decomposição funcional do acervo (quatro componentes) e a maturidade curatorial progressiva (pipeline de três camadas). Essa dupla organização permite tratar heterogeneidade, rastreabilidade e publicabilidade como dimensões integradas do mesmo problema de gerenciamento de dados.

4.1. Estrutura Funcional de Representação dos Dados

O modelo organiza o acervo em quatro componentes funcionais com responsabilidades distintas e explicitamente separadas. O Dicionário de Dados concentra a semântica das variáveis: nomes, tipos, unidades, granularidade e domínios válidos. Os Arquivos de Dados preservam a materialidade física dos objetos digitais: medições in situ, saídas instrumentais e arquivos derivados. Os Datasets representam a unidade lógica, citável e versionável do sistema. Os Produtos de Dados agregam datasets segundo critérios temáticos ou analíticos, favorecendo descoberta e exploração em escala.

Essa organização desacopla significado, armazenamento, publicação e agregação temática – dimensões que, em muitos acervos científicos, ficam implícitas ou sobrepostas, dificultando governança e reuso. No DataMap/Amazon, cada componente cumpre função específica, fornecendo uma estrutura operacional para tratar objetos heterogêneos sem perda de proveniência.

4.2. Pipeline de Qualificação Progressiva

A qualificação ocorre em três camadas: Bronze, Prata e Ouro. A estrutura é inspirada na arquitetura medalhão (*medallion architecture*), padrão de plataformas de data lakehouse como o Delta Lake [Armbrust et al. 2020], adaptada à curadoria científica: a semântica de transformações transacionais dá lugar à lógica de progressão curatorial orientada ao ciclo de vida científico.

Na camada Bronze, os dados são preservados em estado bruto e imutável, garantindo recuperabilidade e reprocessamento. Na camada Prata, passam por limpeza, padronização de formatos, harmonização de unidades e calendários e enriquecimento com metadados mínimos – operações documentadas com referência ao curador responsável. Na camada Ouro, são submetidos à curadoria especializada: (i) validação estatística descritiva; (ii) identificação e tratamento documentado de lacunas temporais; (iii) controle de valores atípicos por métodos padronizados – Z-score, definido como

$$z = \frac{x - \mu}{\sigma}, \tag{1}$$

onde x é o valor observado, μ é a média e σ é o desvio padrão, e Intervalo Interquartil (IQR), definido como

$$IQR = Q_3 - Q_1, \tag{2}$$

onde Q_1 e Q_3 são o primeiro e o terceiro quartis, respectivamente; (iv) enriquecimento semântico com vocabulários controlados; (v) definição de licença Creative Commons; e (vi) atribuição de DOI por versão via integração com a DataCite [Fenner et al. 2019]. Cabe ressaltar que a detecção de anomalias (Z-score, Eq. 1; IQR) é balanceada com o conhecimento de domínio: um valor estatisticamente extremo pode representar um evento climático real – como uma seca severa ou uma enchente histórica – e não um erro de sensor. Por isso, os métodos estatísticos atuam como sinalizadores (*flags*), e a decisão de filtrar ou manter o dado baseia-se nos parâmetros físicos esperados para o instrumento. Esse equilíbrio reflete a divisão de trabalho do pipeline: os *scripts* executam tarefas repetitivas e de larga escala – ingestão via TUS (Bronze), padronização de formatos e harmonização de unidades (Prata) e cálculo das métricas estatísticas (Ouro) –, enquanto a curadoria humana define o dicionário semântico, enriquece metadados, seleciona licenças e valida os sinalizadores, exercendo o julgamento necessário para discernir falha instrumental de fenômeno natural.

A aderência aos princípios FAIR na camada Ouro é explícita: encontrabilidade (Findable) por metadados ricos e DOIs persistentes; acessibilidade (Accessible) por protocolos padronizados e landing pages públicas; interoperabilidade (Interoperable) por vocabulários controlados, formatos abertos e exportação de metadados Dublin Core [Brickley et al. 2019]; e reutilizabilidade (Reusable) por licença Creative Commons, proveniência detalhada e documentação suficiente para novo uso analítico [Wilkinson et al. 2016].

A frequência de atualização varia conforme o perfil de cada iniciativa – de contínua (ATTO-Campina) a por campanha (CHUVA, LBA) ou semestral/anual (AmazonFACE) –, conforme detalhado na Tabela 2.

A Figura 1 demonstra o ciclo de vida completo de um dataset no DataMap/Amazon: do upload bruto à qualificação progressiva (Bronze, Prata e Ouro, onde anomalias são tratadas) e à geração de identificadores persistentes e disponibilização externa, atendendo aos princípios FAIR.

5. Arquitetura da Plataforma

A arquitetura da plataforma operacionaliza o modelo integrando portal de descoberta, orquestração curatorial, persistência em PostgreSQL, armazenamento em MinIO e serviços de DOI, com o fluxo evoluindo do upload via protocolo TUS [tus.io 2023] até a descoberta externa via metadados FAIR (Figura 1).

A Tabela 1 ilustra a diferenciação estratégica do DataMap/Amazon: enquanto repositórios generalistas como Zenodo, Dataverse, Figshare e DataONE priorizam o depósito e a preservação de ativos já preparados, a plataforma internaliza o ciclo completo de qualificação – pipeline progressivo (Bronze-Prata-Ouro), dicionário semântico e validações automáticas (estatísticas e semânticas) –, otimizando qualidade e interoperabilidade em contextos de alta complexidade ambiental como o amazônico.

Figura 1. Ciclo de vida de um dataset no DataMap/Amazon, evidenciando o fluxo progressivo de qualificação e publicação.

Tabela 1. Comparação técnica entre o DataMap/Amazon e repositórios científicos consolidados.

Funcionalidade	DataMap /Amazon	Zenodo	Dataverse	Figshare	DataONE
Curadoria progressiva	Nativa (Bronze-Prata-Ouro)	Não	Não	Não	Não
Granularidade de proveniência	Alta (Histórico de transformações)	Mínima	Mínima	Mínima	Média
DOI por versão	Sim	Sim	Sim	Sim	Sim
Dicionário semântico	Integrado	Não	Não	Não	Não
Pipeline de validação	Automático (Estatístico/Semântico)	Não	Não	Não	Não
Especialização de domínio	Alta (Ambiental Amazônico)	Baixa (Genérico)	Média	Baixa (Genérico)	Alta (Ambiental)

6. Estudo de Caso e Resultados

O modelo foi aplicado à curadoria de dados de quatro iniciativas estratégicas: LBA (Large-scale Biosphere-Atmosphere Experiment in Amazonia), AmazonFACE (Amazon Free Air CO_2 Enrichment), ATTO-Campina (Amazon Tall Tower Observatory sítio Campina) e CHUVA (Cloud Processes of the Main Precipitation Systems in Brazil), cujos acervos encontravam-se majoritariamente em formatos brutos e heterogêneos. A Tabela 2 caracteriza o recorte empírico.

A aplicação seguiu o pipeline Bronze-Prata-Ouro: os arquivos foram ingeridos e preservados em sua forma original, harmonizados quanto a formatos, unidades, dicionários e metadados, e por fim submetidos a validação estatística, tratamento de lacunas e valores atípicos (Z-score, Eq. 1; IQR, Eq. 2), enriquecimento semântico e versionamento com DOI. As iniciativas cobrem perfis curatoriais distintos, desde séries temporais longas dependentes de consistência entre campanhas (LBA, ATTO-Campina) até alta variabilidade de formatos e diversidade instrumental (CHUVA).

Os resultados quantitativos são apresentados na Tabela 3, referentes ao estado do sistema em março de 2025. Foram estruturados 183 datasets, publicadas 371 versões, gerados 165 DOIs – dos quais 130 já estão em estado encontrável – e organizados 1,58

Tabela 2. Caracterização do estudo de caso: quatro iniciativas estratégicas da pesquisa amazônica.

Iniciativa	Foco científico	Características dominantes	Perfil de atualização
LBA	Biosfera-atmosfera	Séries ambientais de longa duração, múltiplos sítios, diversidade instrumental.	Ciclos por campanha e período de operação dos sítios.
AmazonFACE	Resposta da floresta ao CO_2	Medições experimentais, variáveis atmosféricas e ecológicas integradas.	Semestral a anual, conforme disponibilização pelas equipes.
ATTO-Campina	Sítio observacional (4 km da torre ATTO)	Observações contínuas, alta resolução temporal, integração multi-instrumento.	Periódica (subhorária a horária); ingestão em ciclos regulares.
CHUVA	Nuvens e precipitação	Dados de campanha, sensores especializados, alta heterogeneidade de formatos; acervo histórico de seis campanhas (2010-2014).	Variável conforme evento. Por campanha e instrumento.

TB distribuídos em 258.004 arquivos. Isso corresponde a uma taxa de publicação efetiva de $130/165 \approx 78,8\%$ e a uma média de $371/183 \approx 2,03$ versões por dataset.

A disparidade entre iniciativas é esperada: o AmazonFACE apresenta volume reduzido (2 datasets, 0,01 TB) por estar em fase inicial de ingestão no período de referência, enquanto as demais tinham acervos mais consolidados ou campanhas encerradas. Os 35 DOIs ainda não encontráveis (21,2%) correspondem a ativos em embargo, aguardando autorização de liberação, ou em propagação nos índices da DataCite; a plataforma opera sob política institucional de preservação digital de longo prazo, garantindo custódia perene independentemente de restrições temporárias de acesso.

Tabela 3. Resultados quantitativos por iniciativa.

Iniciativa	Datasets	Versões	DOIs	Encontráveis	TB	Arquivos
LBA	55	148	65	52	0,61	96.500
AmazonFACE	2	3	2	2	0,01	100
ATTO-Campina	78	133	55	43	0,60	95.204
CHUVA	48	87	43	33	0,36	66.200
Total	**183**	**371**	**165**	**130**	**1,58**	**258.004**

A leitura dos números responde às três questões: a estruturação de 183 datasets com metadados enriquecidos e versões internas evidencia um fluxo curatorial efetivo, e não mero depósito (QP1); a média de 2,03 versões por dataset e os 78,8% de DOIs encontráveis indicam publicação persistente e citável (QP2); e o volume de 1,58 TB em 258.004 arquivos, entre quatro iniciativas de perfis distintos, demonstra capacidade ope-

racional em escala relevante (QP3). Reitera-se que essas métricas indicam atividade curatorial realizada – não qualidade do resultado nem impacto no reuso efetivo, distinção discutida na Seção 7.

7. Discussão

Do ponto de vista FAIR, os 130 DOIs registrados na DataCite e verificados como ativos demonstram que o pipeline de qualificação produziu objetos conformes com o princípio Findable: identificadores persistentes atribuídos, metadados ricos e landing pages públicas acessíveis [Wilkinson et al. 2016, Fenner et al. 2019]. A exportação de metadados Dublin Core e o uso de vocabulários controlados atendem à interoperabilidade – embora a adoção de esquemas mais expressivos, como Schema.org ou DataCite Metadata Schema 4.x, possa ampliar a compatibilidade com sistemas de descoberta externos [Brickley et al. 2019]. Essa extensão é identificada como direção de desenvolvimento futuro.

Em comparação com Zenodo e Dataverse, o DataMap/Amazon não se distingue pelos recursos técnicos disponíveis em cada plataforma, mas pelo modelo operacional: internaliza o fluxo de qualificação como parte nativa da infraestrutura, enquanto aquelas plataformas delegam essa responsabilidade ao pesquisador. Essa diferença é particularmente relevante em contextos de alta heterogeneidade e baixa maturidade curatorial institucional – precisamente as condições predominantes nos acervos amazônicos estudados. Reconhece-se, contudo, que essa comparação é qualitativa e baseada em análise das características operacionais das plataformas, não em experimento controlado.

Do ponto de vista do bem social, o impacto potencial está em ampliar a colaboração científica sobre mudanças climáticas, o acesso de gestores públicos e da sociedade civil à evidência para políticas de conservação, e a replicabilidade do fluxo para outros domínios de dados públicos. Nesse sentido, a arquitetura medalhão vai além da eficiência técnica: pesquisadores de países em desenvolvimento frequentemente não dispõem dos recursos computacionais para limpar e estruturar terabytes de dados brutos. Ao disponibilizar os dados já na camada Ouro – com anomalias tratadas e vocabulários controlados aplicados –, o DataMap/Amazon atua como nivelador de assimetrias de capacidade técnica, transferindo o custo curatorial do pesquisador individual para a infraestrutura institucional.

Permanecem limitações importantes. Primeiro, embora a aderência aos princípios FAIR tenha sido estruturada no desenho do modelo, o estudo carece de avaliação quantitativa e automatizada dos metadados por ferramentas reconhecidas, como o F-UJI ou o FAIR Evaluator. Segundo, não se mediu longitudinalmente o reuso efetivo dos datasets nem o impacto sobre citações – distinção central entre atividade de curadoria realizada e qualidade ou impacto do resultado [Michener et al. 2012]. Terceiro, a avaliação não compara o modelo com fluxos curatoriais alternativos ou sem o uso da plataforma. Quarto, a expansão para dados de sensoriamento remoto, imagens de satélite e dados socioeconômicos ainda requer investigação. Quinto, os resultados cobrem apenas quatro iniciativas selecionadas intencionalmente, o que impede generalizações para o conjunto da pesquisa amazônica. Esses pontos definem a agenda futura, a ser conduzida nos próximos ciclos de operação da plataforma: aferição do escore FAIR por princípio (F-UJI, FAIR Evaluator), mensuração de uso e impacto (*downloads*, visualizações das

landing pages e citações via DataCite Event Data) e comparação controlada com fluxos curatoriais alternativos.

8. Conclusão

Este artigo apresentou e analisou um modelo de curadoria e publicação de dados científicos amazônicos implementado no DataMap/Amazon. Sua principal contribuição reside na integração operacional de práticas consolidadas – decomposição funcional do acervo em quatro componentes, qualificação progressiva inspirada na arquitetura medalhão de data lakehouse [Armbrust et al. 2020] e atribuição de DOI por estado curatorial – em um fluxo único de gestão do ciclo de vida dos dados, capaz de converter observações heterogêneas em ativos científicos FAIR, persistentes e citáveis.

Ao evidenciar essa integração em quatro iniciativas reais, o trabalho desloca o debate sobre repositórios científicos de uma perspectiva puramente custodial para uma perspectiva curatorial orientada à qualidade do ativo publicado. Alinhado ao eixo do bem social do DS4SG, o DataMap/Amazon amplia o acesso da sociedade a dados ambientais amazônicos qualificados – insumo essencial para pesquisa sobre mudanças climáticas, conservação da biodiversidade e políticas de uso do solo.

Uso de Inteligência Artificial

Os autores declaram que ferramentas de Inteligência Artificial foram utilizadas como auxílio para revisão linguística e gramatical, estruturação e clareza textual, e geração de imagens. A responsabilidade integral pelo conteúdo científico, pelas afirmações e pela verificação de todas as referências é exclusivamente dos autores humanos, em conformidade com o Código de Conduta da SBC para Autores (Art. 2º, III).

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

Armbrust, M. et al. (2020). Delta Lake: High-performance ACID table storage over cloud object stores. *Proceedings of the VLDB Endowment*, 13(12):3411–3424.

Brickley, D., Burgess, M., and Noy, N. F. (2019). Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *Proceedings of The World Wide Web Conference (WWW 2019)*, pages 1365–1375, New York. ACM. https://doi.org/10.1145/3308558.3313685.

CERN e OpenAIRE (2013). Zenodo: Research. shared. CERN, Geneva. https://zenodo.org.

Data Citation Synthesis Group (2014). Joint declaration of data citation principles. FORCE11, San Diego, CA. https://www.force11.org/datacitationprinciples.

DataCite Metadata Working Group (2021). DataCite metadata schema documentation for the publication and citation of research data and other research outputs. version 4.4. DataCite. https://doi.org/10.14454/3w3z-sa82.

Davidson, E. A. et al. (2012). The Amazon basin in transition. *Nature*, 481:321–328.

Fenner, M. et al. (2019). A data citation roadmap for scholarly data repositories. *Scientific Data*, 6:28.

Groth, P. and Moreau, L. (2013). PROV-overview: An overview of the PROV family of documents. W3C Working Group Note, 30 April 2013. https://www.w3.org/TR/prov-overview/.

Hey, T., Tansley, S., and Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, WA.

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation*, 3(1):134–140.

King, G. (2007). An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods & Research*, 36(2):173–199.

Michener, W. K. et al. (2012). Participatory design of DataONE – enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics*, 11:5–15.

Thelwall, M. and Kousha, K. (2016). Figshare: a universal repository for academic resource sharing? *Online Information Review*, 40(3):333–346. https://doi.org/10.1108/OIR-06-2015-0190.

tus.io (2023). Resumable uploads protocol tus: open protocol for resumable file uploads. https://tus.io/protocols/resumable-upload.

UNESCO (2021). Recommendation on Open Science. Adopted by the General Conference of UNESCO at its 41st session, 23 November 2021. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000379949.

Wieringa, R. J. (2014). *Design Science Methodology for Information Systems and Software Engineering*. Springer, Berlin.

Wilkinson, M. D. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.