Repositório de Dados para Ciência Aberta na Região Amazônica

André N. Maia[1], **Caio N. Maia**[1], **Pedro L. P. Corrêa**[1]

[1]Departamento de Engenharia de Computação e Sistemas Digitais
Universidade de São Paulo (USP) – São Paulo – SP – Brasil

Abstract. *The Amazon region is vital for climate regulation and biodiversity conservation, generating large volumes of environmental and atmospheric data. However, these data are often scattered, hard to access, and lack standardization, hindering reuse and collaborative research. Open-science initiatives require platforms that comply with FAIR principles, ensuring proper curation, persistent identifiers, and dissemination. This paper presents DataMap, a data repository specialized in managing and processing Amazonian data, supporting visualization, discovery, cataloging, and publication of datasets, with automatic DOI generation. The platform enhances accessibility and interoperability, enabling exploration through web interfaces and REST APIs.*

Resumo. *A região amazônica é essencial para a regulação climática e a conservação da biodiversidade, gerando grandes volumes de dados ambientais e atmosféricos. Contudo, esses dados frequentemente são dispersos, de difícil acesso e sem padronização, dificultando sua reutilização e pesquisa colaborativa. Iniciativas de ciência aberta requerem plataformas que sigam os princípios FAIR, promovendo curadoria adequada, identificadores persistentes e disseminação eficiente. Este trabalho apresenta o DataMap, um repositório especializado na gestão de dados amazônicos, que suporta visualização, descoberta, catalogação e publicação de datasets, com geração automática de DOIs, além de garantir acessibilidade e interoperabilidade.*

1. Introdução

A região amazônica desempenha um papel crucial na regulação climática global e na preservação da biodiversidade. Ela é responsável por uma significativa parte da absorção de carbono atmosférico, além de ser uma das maiores fontes de oxigênio do planeta e abrigar uma biodiversidade incomparável. O desmatamento e as mudanças climáticas ameaçam esses serviços ecossistêmicos, o que torna o monitoramento contínuo de seus dados ambientais e atmosféricos fundamental para entender os impactos dessa degradação e para a formulação de políticas de conservação eficazes [Foley et al. 2007]. No entanto, esses dados frequentemente se encontram dispersos, fragmentados e sem padronização, o que dificulta sua reutilização, a realização de pesquisas colaborativas e a integração com outras iniciativas científicas globais [Fleming et al. 2017]. Em um cenário onde a pesquisa científica sobre a Amazônia é crucial, a falta de acesso eficiente aos dados se tornam obstáculos para avanços significativos na compreensão dos fenômenos ambientais da região.

Nesse contexto, a ciência aberta e gestão de dados científicos [UNESCO 2021], e os princípios FAIR (*Findable, Accessible, Interoperable, Reusable*) são fundamentais.

Esses princípios garantem que os dados científicos sejam facilmente localizáveis, acessíveis a pesquisadores de diversas partes do mundo, interoperáveis com diferentes plataformas e reutilizáveis para diversas finalidades. A adoção dos princípios FAIR no gerenciamento de dados da Amazônia pode, portanto, facilitar a colaboração internacional, promover a transparência científica e maximizar o impacto das pesquisas ambientais [Wilkinson et al. 2016].

Embora o Brasil seja um dos países líderes na quantidade de repositórios de dados na América Latina, ainda está longe de alcançar a infraestrutura necessária para uma ciência aberta plena. De acordo com um levantamento do DataCite [Garduño-Magaña 2024], o Brasil possui uma quantidade significativa de repositórios, mas em comparação com líderes globais em ciência aberta, como os Estados Unidos, o país ainda apresenta uma lacuna considerável. Os Estados Unidos, por exemplo, têm mais de 1200 repositórios registrados no Re3data, o que demonstra uma infraestrutura mais robusta e estabelecida para o armazenamento, o compartilhamento e a reutilização de dados científicos. Isso reflete a necessidade de expandir a rede de repositórios no Brasil para apoiar a ciência aberta de maneira mais eficaz.

Este trabalho descreve a arquitetura da plataforma DataMap, que foi criada para ser um repositório de dados atmosféricos da Amazônia, atendendo aos requisitos de interoperabilidade e acessibilidade necessários para a colaboração científica global. Além da contribuição técnica e funcional, o artigo também apresenta uma reflexão sobre os desafios enfrentados na implementação da arquitetura proposta, como a garantia de alta escalabilidade para grandes volumes de dados, a atribuição de identificadores digitais únicos (DOI) a *datasets* permitindo o rastreamento de forma única permanente a garantir reprodutibilidade e repetibilidade a experimentos, e a necessidade de interfaces acessíveis para uma ampla gama de usuários, desde pesquisadores até gestores públicos e tomadores de decisão.

A inovação do DataMap reside não apenas em sua arquitetura técnica e de código aberto, mas também na sua abordagem para a ciência aberta, oferecendo uma solução que integra os dados da Amazônia com os mais altos padrões de acessibilidade e reutilização, fundamentais para a continuidade e expansão das pesquisas ambientais sobre a região.

2. Desafios para criação e gestão de repositórios de dados

Este capítulo apresenta os principais desafios na criação e gestão de repositórios de dados ambientais, com foco na Amazônia. O primeiro é a dispersão e fragmentação dos dados, que dificulta sua reutilização e integração. O segundo refere-se à infraestrutura insuficiente para ciência aberta no Brasil, que ainda conta com poucos repositórios especializados na região, especialmente em comparação com países como os Estados Unidos. O terceiro desafio envolve a adoção dos princípios FAIR, que requerem um ciclo de vida de dados estruturado, com ênfase em acessibilidade e interoperabilidade. O quarto diz respeito à curadoria, proveniência e segurança, aspectos essenciais para garantir a confiabilidade, rastreabilidade e proteção dos dados ao longo do tempo.

Este capítulo discute como a plataforma DataMap contribui para enfrentar esses obstáculos, promovendo a ciência aberta e o acesso seguro aos dados ambientais amazônicos.

2.1. Dispersão e fragmentação dos dados ambientais

A ciência moderna, frequentemente chamada de *eScience* ou quarto paradigma, depende do uso intensivo de dados. Esse conceito envolve tecnologias computacionais aplicadas à coleta, processamento e análise de grandes volumes, transformando a prática científica [Hey et al. 2009].

Entretanto, a *eScience* também impõe desafios, especialmente na gestão e no compartilhamento de dados. No contexto amazônico, a dispersão e fragmentação das informações, armazenadas em múltiplas fontes e formatos sem padronização, dificulta o trabalho de pesquisadores que dependem desses dados.

Segundo o *DataONE*, pesquisas baseadas em dados seguem etapas como coleta, armazenamento, organização, análise e reutilização [Michener et al. 2011]. Em ambientes onde os dados estão dispersos, como a Amazônia, executar essas etapas com eficiência se torna um desafio.

2.2. Infraestrutura insuficiente para ciência aberta no Brasil

Embora o Brasil esteja entre os líderes latino-americanos na publicação de *datasets*, com cerca de 210.000 DOIs registrados em 2024 segundo o DataCite [Garduño-Magaña 2024], sua infraestrutura de repositórios ainda é limitada. De acordo com o Re3data [RE3DATA 2025], o país possui apenas 23 repositórios científicos registrados, frente a mais de 1.200 nos Estados Unidos.

Entre esses, apenas um é dedicado exclusivamente a dados ambientais da Amazônia: o ATTO (*Amazon Tall Tower Observatory*). Essa escassez dificulta a colaboração científica e a formulação de políticas públicas baseadas em evidências. O repositório ARM (*Atmospheric Radiation Measurement*), embora relevante e com dados sobre a Amazônia, está sediado fora do Brasil e não integra sua infraestrutura nacional.

Essa disparidade evidencia a urgência de ampliar a ciência aberta no país, com plataformas nacionais que sigam os princípios FAIR e fomentem a colaboração internacional.

2.3. Desafios no Desenvolvimento de Repositórios de Dados Ambientais

Desenvolver um repositório de dados bioclimáticos conforme os princípios FAIR requer requisitos técnicos e operacionais rigorosos [Wilkinson et al. 2016]. Segundo o DataONE, o ciclo de vida dos dados deve incluir *planejamento*, *coleta*, *descrição*, *preservação*, *descoberta*, *integração* e *análise* [Michener et al. 2011].

Também é essencial descrever os dados com metadados estruturados, usando padrões como os estabelecidos em https://schema.org/Dataset, para facilitar sua descoberta e reutilização [Brickley et al. 2019]. A *interoperabilidade* depende de vocabulários controlados, enquanto a acessibilidade é garantida por protocolos abertos como HTTP e FTP.

Tais exigências demandam infraestrutura e capacitação adequadas, representando desafios relevantes à implementação de repositórios alinhados aos princípios FAIR.

2.4. Curadoria, Proveniência e Segurança dos Dados

A curadoria assegura que os dados sejam bem descritos, organizados e atualizados, facilitando sua reutilização científica. A proveniência garante rastreabilidade da origem e

integridade dos dados. Já a segurança envolve proteção contra acesso não autorizado, corrupção ou perda de dados.

Para que os dados ambientais da Amazônia sejam úteis em pesquisas colaborativas, e para assegurar reprodutibilidade e repetibilidade, é essencial que esses três aspectos sejam atendidos, garantindo confiabilidade, transparência e acesso seguro às informações.

3. Arquitetura da plataforma DataMap

A plataforma DataMap foi projetada para dar suporte ao ciclo de vida completo de dados ambientais e atmosféricos da Amazônia, desde a ingestão bruta até a publicação citável de versões de dados. A arquitetura da plataforma DataMap busca resolver os problemas anteriormente elencados: dispersão de dados, infraestrutura insuficiente para ciência aberta em território brasileiro e desafios específicos em repositórios de dados bioclimáticos.

A concepção da plataforma parte do que foi formulado em *The Fourth Paradigm* [Hey et al. 2009]: a próxima onda de descoberta científica dependerá da capacidade de transformar volumes massivos e heterogêneos de dados em conhecimento auditável. Para concretizar essa visão, a plataforma foi alinhada ao ciclo completo de gestão proposto por Michener [Michener et al. 2011] no contexto do DataONE, que inclui planejamento, coleta, garantia de qualidade, descrição, preservação, descoberta, integração e análise. A cada uma dessas oito etapas foi associado um componente ou serviço que elimina atritos recorrentes no trabalho dos pesquisadores.

Durante o planejamento e a coleta, a plataforma exige que o depositante forneça metadados estruturados mínimos e oferece mecanismos de *upload* resiliente capazes de retomar transferências interrompidas, garantindo que séries *multi gigabyte* cheguem intactas. As fases de qualidade e descrição são implementadas por um fluxo de curadoria em camadas: dados brutos são preservados tal-qual, depois passam por saneamento e, finalmente, recebem enriquecimento semântico realizado por especialistas. A preservação é obtida por replicação e versionamento integral, de forma que todo estado anterior permaneça consultável. Para a descoberta, a plataforma publica catálogos pesquisáveis por espaço, tempo e variável; a integração é suportada por APIs e esquemas abertos que permitem combinar diferentes coleções sem conversões *ad hoc*; por fim, a análise é facilitada por agrupamentos temáticos que entregam pacotes prontos para modelagem estatística ou aprendizagem de máquina.

Ao longo de todo o ciclo, a plataforma faz convergir, de forma orgânica, os quatro princípios FAIR. Os conjuntos tornam-se encontráveis porque cada versão publicada recebe automaticamente um DOI e integra um catálogo indexado por tempo, espaço e palavra-chave, exposto a buscadores de dados. Permanecem acessíveis por protocolos padronizados de *download* e por resoluções de DOI que redirecionam para a página de metadados. A adoção de formatos abertos, vocabulários controlados e esquemas de metadados compatíveis garante a interoperabilidade, permitindo combinações sem conversões idiossincráticas. Finalmente, a associação entre licenças claras, trilha de proveniência e versionamento integral, todos vinculados ao DOI, assegura que os dados sejam reutilizáveis em novas hipóteses ou análises. Assim, a DataMap converte acervos dispersos em ativos científicos robustos, alinhados ao paradigma da pesquisa intensiva em dados.

3.1. Taxonomia de Dados

A plataforma DataMap organiza os dados em quatro níveis hierárquicos, concebidos para transformar observações brutas em ativos científicos progressivamente qualificados. Os níveis são: Dicionário de Dados, Arquivos de Dados, *Datasets* e Produtos de Dados. Esses níveis estruturam o acervo de modo a viabilizar curadoria semântica, preservação física e descoberta interdisciplinar. A Figura 1 ilustra essa estrutura, enquanto a Tabela 1 detalha suas funções.

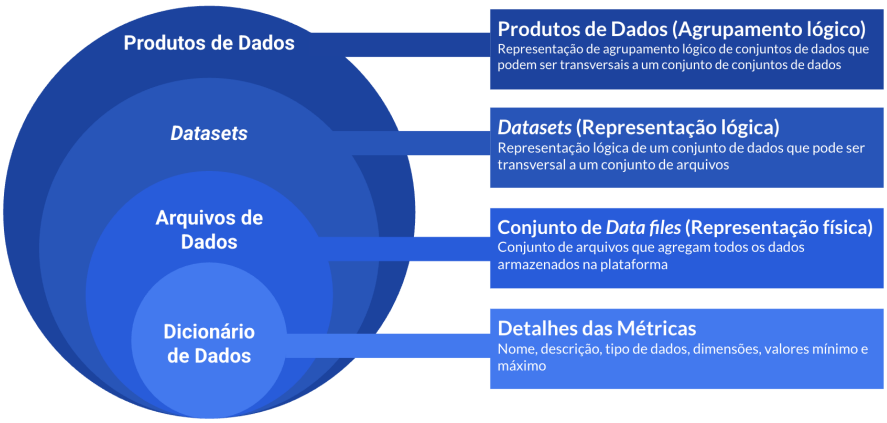

Figura 1. Estrutura hierárquica da taxonomia de dados na plataforma DataMap.

Nível	Descrição
Dicionário de Dados	Núcleo semântico da plataforma. Define cada métrica elementar (nome, tipo, unidades, dimensões, valores válidos), permitindo validações automáticas e exploração padronizada. Serve de base para a interoperabilidade entre conjuntos distintos.
Arquivos de Dados	Representação física do acervo (NetCDF, CSV, GeoTIFF, etc.), contendo medições brutas oriundas de múltiplas fontes, como sensores de campo, reanálises atmosféricas ou produtos de satélite.
Dataset	Entidade lógica e citável, composta por múltiplos arquivos relacionados. Cada *dataset* agrega metadados científicos relevantes (tempo, espaço, autores, método), possui versionamento formal e recebe identificador persistente (DOI).
Produto de Dados	Coleção temática de *datasets* agrupados por critério comum (ex.: variável, região, instrumento), oferecendo uma entrada simplificada para exploração e análise em larga escala.

Tabela 1. Níveis da Taxonomia de Dados do DataMap

Essa organização viabiliza um acesso incremental à informação, desde a preservação *bit a bit* da observação original até conjuntos qualificados para visualizações, estatísticas ou modelagem computacional. Além disso, promove rastreabilidade, reprodutibilidade e reutilização conforme os princípios FAIR, facilitando a integração entre dados heterogêneos e projetos colaborativos.

3.2. Ciclo de Vida dos Dados da Plataforma

Baseando-se na taxonomia adotada, a plataforma organiza o ciclo de vida dos dados ambientais por meio de um *pipeline* em três camadas: *Bronze*, *Silver* e *Gold*. Essa estrutura

é inspirada em boas práticas de engenharia de dados para *data lakes* e foi concebida para mitigar a heterogeneidade estrutural das fontes, a ausência de identificadores persistentes e a dificuldade de reuso interdisciplinar, conforme discutido na Seção 2.

Na camada *Bronze*, fluxos contínuos de arquivos oriundos de reanálises numéricas, produtos de sensoriamento remoto e medições *in situ* são armazenados de forma bruta e imutável, sem qualquer transformação. Esta camada funciona como um *data lake* histórico, preservando a integridade forense dos dados e garantindo reprodutibilidade ao possibilitar o retorno ao dado original em análises futuras.

A transição para a camada *Silver* inicia o processo de qualificação. Nessa etapa, os dados são saneados, padronizados em unidades e calendários, normalizados em sistemas de coordenadas e enriquecidos com metadados mínimos compatíveis com o esquema de catalogação da plataforma. Em seguida, a camada *Gold* aplica rotinas semiautomáticas de agregação multiespacial, validação estatística, verificação de lacunas e enriquecimento semântico. Especialistas contribuem com curadoria, cruzando séries complementares, documentando e removendo valores atípicos, e descrevendo variáveis segundo vocabulários controlados do domínio.

Cada versão resultante na camada *Gold* recebe um identificador persistente (DOI), assegurando rastreabilidade, persistência e citabilidade conforme os padrões internacionais de dados abertos. Esse fluxo incremental de qualificação permite análises exploratórias e inferenciais com menor sobrecarga de pré-processamento, além de fornecer conjuntos balanceados e consistentes para treinos reprodutíveis de modelos de *machine learning*. As transições entre as camadas promovem a progressiva adição de contexto, coerência e valor científico ao *dataset* original.

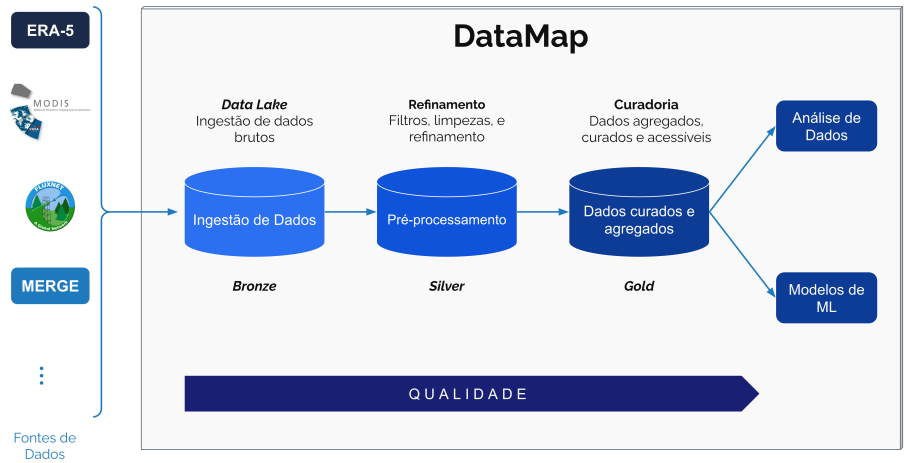

Figura 2. Representação do ciclo de vida dos dados na plataforma DataMap, estruturado nas camadas *Bronze*, *Silver* e *Gold*.

3.3. Segurança da Informação

A integridade científica de um repositório depende, em igual medida, da qualidade dos dados e da confiança no ambiente que os hospeda. Para proteger confidencialidade, integridade e disponibilidade (tríade CIA) dos recursos publicados no *DataMap*, a implementação adota um *defence-in-depth* que combina camadas criptográficas, controles de identidade e autorização fina, verificação humana de conteúdo e isolamento de rede.

Canal autenticado e confidencial. Todo o tráfego externo utiliza TLS 1.3 e termina em um *reverse proxy nginx* exposto na DMZ; os demais serviços comunicam-se apenas em rede interna, reduzindo a superfície de ataque. Qualquer requisição à API deve apresentar um par *API Key*/*API Secret* gerado pelos administradores; as chaves são armazenadas com *hashing* resistente a *rainbow tables* e rotacionadas periodicamente. Operações de *upload* incluem ainda um *JSON Web Token* (JWT, RFC 7519) assinado no algoritmo HS256 com chave privada; o *claim aud=file-upload* é verificado pelo servidor antes de aceitar qualquer fragmento de arquivo, prevenindo tentativas de *token replay* em rotas não autorizadas e *upload* de arquivos por usuários não aprovados.

Identidade federada e gestão de credenciais. A maior parte dos usuários autentica-se por **OAuth 2.0** com ORCID, eliminando o gerenciamento local de senhas e herdando políticas de autenticação forte adotadas globalmente pela comunidade científica. Quando credenciais internas são estritamente necessárias, as senhas permanecem criptografadas no banco de dados e nunca ficam acessíveis a operadores. Novos cadastros só se tornam aptos a submeter dados após *due-diligence* conduzida pela equipe curatorial, procedimento que mitiga ingestões maliciosas ou fora de escopo.

Autorização multi-nível. O controle de acesso segue o modelo RBAC: a biblioteca *Casbin* aplica políticas que vinculam papéis (administrador, curador, editor, leitor) a operações específicas (CRUD de metadados, publicação de versões, geração de DOI). Esses papéis podem ser revogados a qualquer momento em resposta a conduta inadequada ou mudança de equipe. Paralelamente, a plataforma opera em regime de *multi-tenancy* lógico; cada requisição carrega um *namespace* persistido em *cookie HttpOnly/Secure*, e filtros de *row-level security* no PostgreSQL asseguram que usuários observem apenas os recursos do seu espaço. O mecanismo isola grupos de pesquisa durante curadoria preliminar e evita vazamento entre projetos concorrentes.

Governança do ciclo de publicação. Antes que um conjunto de dados receba estado *findable* na agência DataCite, a versão passa por inspeção obrigatória de um curador, que avalia metadados, consistência dos arquivos e aderência às políticas FAIR. Esse *gate* humano reduz o risco de exposição de material sensível ou metadados incompletos. Similarmente, todo novo usuário é aprovado manualmente, garantindo que o repositório permaneça dedicado a conteúdo científico.

Isolamento de rede e serviços essenciais. Somente o *proxy* de borda se encontra visível na Internet pública; serviços internos (banco de dados, *object store*, compactador e gerador de DOI) operam em sub-redes privadas sem rota externa. Políticas de *iptables* e *network namespaces* nos contêineres impedem que processos comprometidos alcancem hosts críticos.

Em conjunto, esses mecanismos implementam boas práticas recomendadas por NIST SP 800-53 e OWASP para aplicações Web, criando um ambiente em que dados podem ser depositados, curados e publicados com garantias verificáveis de confidencialidade, integridade e rastreabilidade.

4. Implementação do DataMap

A implementação materializa a arquitetura conceitual em um conjunto enxuto de serviços contêinerizados que rodam em infraestrutura local da Universidade de São Paulo (USP),

mas já configurados para escalar, sem fricção, para nuvem pública. No coração do sistema está o serviço de domínio responsável por autenticação via ORCID, autorização por *namespace* e registro de *datasets*; ele expõe uma API REST descrita em OpenAPI, grava metadados em PostgreSQL (com colunas JSONB para flexibilidade) e aplica filtros de *row-level security* para que cada requisição enxergue apenas os recursos do seu *namespace*. Sobre essa base operam mecanismos de ingestão resiliente: o serviço TUS aceita arquivos *multi gigabyte* fragmentados em blocos, valida cada fragmento por meio de *hooks* que consultam o serviço central e grava diretamente no *object store* MinIO, evitando o gargalo de armazenamento temporário no *backend*. Assim que o pesquisador encerra uma versão inicial do *dataset*, um serviço interno chamado Zipper dispara um processo de compactação assíncrono que reúne todos os arquivos em um único arquivo ZIP, sobe o pacote de volta ao MinIO e marca a versão como pronta para *download*; o *frontend*, por sua vez, passa a exibir um botão que entrega ao usuário uma URL pré-assinada, válida por 24h, garantindo transferência direta do objeto sem sobrecarregar o servidor de aplicação.

A publicação citável do dado é viabilizada por um microserviço dedicado integrado com a API da DataCite. A cada nova versão, ele cria o identificador em estado *draft*, envia o bloco mínimo de metadados obrigatórios (título, autores, ano, tipo) e em seguida promove o DOI ao estado de *findable*, tudo em chamada síncrona para que o identificador esteja disponível no instante em que a interface libera o conjunto. Estratégias de retentativa, que utilizam *exponencial backoff* com *full jitter*, suavizam falhas transitórias do serviço externo; todos os eventos de estado são registrados tanto em logs estruturados quanto na trilha de proveniência do banco. A comunicação entre os serviços ocorre por HTTP interno.

No plano operacional, a plataforma dispõe de um eixo de observabilidade desde o primeiro *commit*: métricas de latência de *endpoint*, uso de CPU e volume trafegado em *upload* são raspadas por Prometheus; logs estratificados fluem para ElasticSearch; e painéis Grafana exibem números de prestação de serviço, como taxa de sucesso em transferências ou tempo médio entre submissão e DOI publicado. O armazenamento físico permanece em um MinIO *single-node* montado sobre volumes Docker persistentes, replicados diariamente para um segundo servidor, atendendo à política institucional de *backup* fora do *datacenter* primário, ao mesmo tempo em que mantém compatibilidade com a API S3 para futuras topologias híbridas.

O ciclo de desenvolvimento segue integração contínua: cada *push* dispara *lint*, testes, varredura de vulnerabilidade e, se aprovado, publica novas imagens de contêiner Docker em registro privado; a instância de *staging* atualiza-se automaticamente, permitindo validação funcional quase em tempo real. Desafios práticos concentraram-se em três frentes. Primeiro, a volatilidade dos *links* de internet amazônicos: a adoção de TUS e blocos de 8 MB reduziu a taxa de falhas de upload para menos de 1%. Segundo, a dependência do serviço de DOI: a retentativa síncrona, combinada a *timeouts* conservadores, manteve latência sob controle sem necessidade, por ora, de enfileiramento assíncrono. Terceiro, a clareza de contexto multi-institucional: o seletor de *namespaces* (*tenancy*) exibido logo após o login, que é persistido em *cookie HttpOnly*, eliminou ambiguidade no *frontend* e dispensou mudanças invasivas na API.

Com esse conjunto de decisões, o DataMap oferece um percurso único: o pesquisador arrasta seus arquivos para o navegador, acompanha a progressão de transferência

dos arquivos, publica o *dataset* com DOI e fornece um *link* de *download* consolidado; simultaneamente, métricas de serviço e trilhas de auditoria asseguram que cada etapa obedeça aos princípios FAIR e às oito fases de gestão de dados sugeridas por Michener (2024). A plataforma, hoje em produção piloto, demonstra que é possível implementar, com componentes abertos e relativamente leves, a infraestrutura necessária para transformar dados brutos em ativos científicos rastreáveis e reutilizáveis.

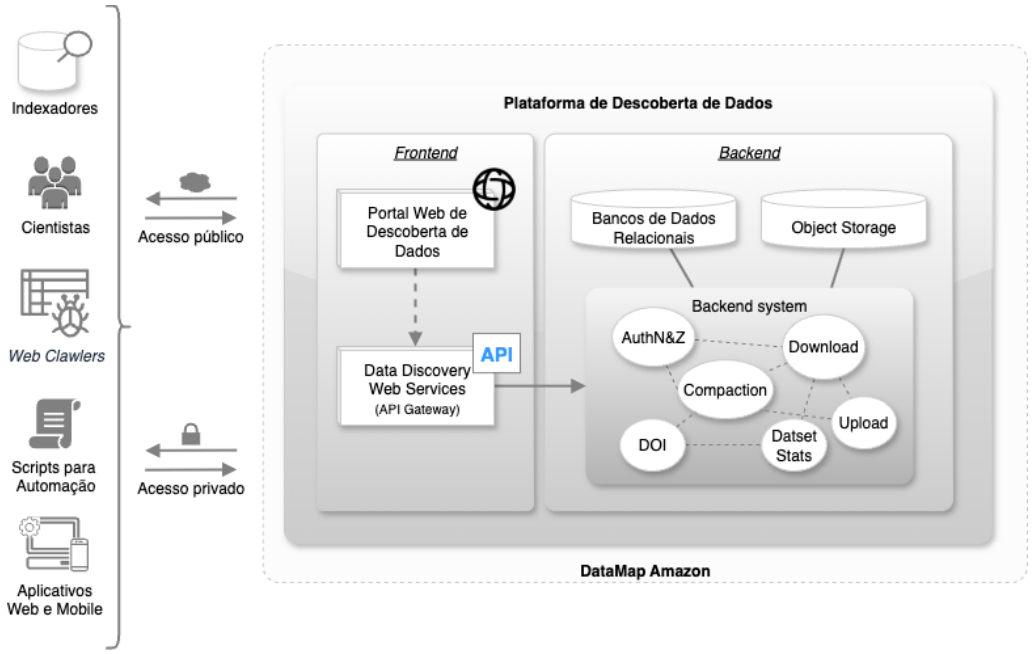

Figura 3. Diagrama da Arquitetura Implementada para a Plataforma DataMap.

A Figura 3 ilustra o ciclo de vida dos dados ambientais na plataforma DataMap, desde a coleta de dados brutos até a sua transformação em ativos científicos rastreáveis e reutilizáveis. O processo envolve etapas como a coleta de dados, a descrição com metadados, a validação e a integração dos dados, garantindo que atendam aos princípios FAIR (*Findable, Accessible, Interoperable, Reusable*). A plataforma facilita o armazenamento, a acessibilidade e a interoperabilidade dos dados, promovendo a colaboração científica e a transparência no uso de dados ambientais da Amazônia.

5. Estudo de caso

A plataforma DataMap foi implementada utilizando *datasets* de projetos de longa duração que envolvem grandes iniciativas científicas, como o LBA (*Large-Scale Biosphere-Atmosphere Experiment in Amazonia*), o AmazonFACE (*Free-Air CO_2 Enrichment*), o ATTO-Campina e o CHUVA (*Cloud and Precipitation in the Amazon*). Cada um desses projetos fornece dados essenciais para a compreensão da dinâmica ambiental da Amazônia e suas interações com o clima global.

- **LBA**: Estuda os ciclos de carbono, nutrientes, água e energia na Amazônia e sua interação com o clima global.
- **AmazonFACE**: Avalia os efeitos do aumento de CO_2 atmosférico nas florestas tropicais a longo prazo.

- **ATTO-Campina**: Coleta dados sobre interações entre atmosfera, vegetação e solo para modelagem climática.
- **CHUVA**: Investiga processos físicos em nuvens e precipitação, essenciais para entender o clima amazônico.

No entanto, muitos dos dados gerados por essas iniciativas ainda não estavam disponíveis em repositórios acessíveis, o que dificultava sua reutilização e análise colaborativa. A plataforma DataMap foi utilizada para gerenciar e curar os conjuntos de dados legados dessas iniciativas, que ainda estavam disponíveis apenas em formatos brutos ou em sistemas fragmentados. Durante a implementação, pesquisadores foram treinados para utilizar a plataforma, o que permitiu que os dados fossem organizados, enriquecidos com metadados e publicados de maneira acessível por meio de DOIs. Isso facilitou a transparência, a citação adequada e o compartilhamento dos dados dentro da comunidade científica global.

Os resultados obtidos mostram um avanço significativo na organização e disponibilização dos dados. Foram criados 183 *datasets*, que resultaram em 371 versões publicadas, gerando um total de 165 DOIs, como podemos ver na Figura 4, sendo que 130 desses DOIs estão agora *findable* e disponível para a comunidade científica.

A Figura 4 mostra o número de DOIs atribuídos aos *datasets* publicados na plataforma DataMap ao longo do tempo, desde a implementação do sistema. Ele ilustra o crescimento progressivo no número de dados acessíveis, evidenciando o impacto da plataforma na promoção de práticas de ciência aberta. O aumento contínuo no número de DOIs reflete a eficácia do DataMap em organizar e disponibilizar dados ambientais da Amazônia para a comunidade científica global.

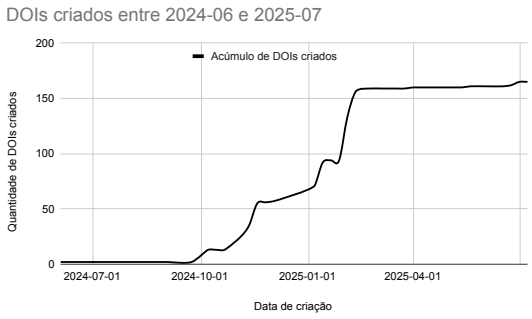

Figura 4. Número de DOIs Publicados e Dados Acessíveis ao Longo do Tempo.

Extensões mais frequentes		
Ext.	**Total**	**%**
txt	55.883	21,66%
dat	48.334	18,73%
t02	36.871	14,29%
lbn	35.352	13,70%
gz	31.131	12,07%
mvol	17.191	6,66%
png	11.418	4,43%
nc	9.380	3,64%
zip	4.759	1,84%
Outras extensões	3.651	1,17%

Tabela 2. Distribuição de arquivos por extensão na plataforma.

A Tabela 2 apresenta a distribuição dos dados armazenados na plataforma DataMap de acordo com os formatos de arquivos utilizados. A plataforma armazenou aproximadamente 1,58 TB de dados, distribuídos em 258.004 arquivos em diversos formatos. Observa-se uma predominância de extensões simples, como *txt* e *dat*, esta última frequentemente utilizada para armazenar dados no formato *NetCDF* (*.nc*) no acervo da plataforma. Formatos consolidados e explicitamente identificados, como *CSV* e *HDF5*, também estão presentes, mas com frequência significativamente menor.

Tivemos 44 usuários criados na plataforma, sendo 11 ativos no sistema, refletindo a aceitação e a utilidade da ferramenta entre os pesquisadores envolvidos nos projetos. A gestão de dados foi otimizada, com a curadoria de informações complexas e volumosas, melhorando a eficiência da pesquisa e ampliando as possibilidades de colaboração internacional.

Este estudo de caso demonstra a capacidade do DataMap de não apenas gerenciar grandes volumes de dados ambientais, mas também de superar desafios relacionados à fragmentação e à falta de infraestrutura para ciência aberta no Brasil. Com a organização e a curadoria eficazes dos dados, a plataforma contribui para a criação de um ambiente colaborativo, acessível e sustentável para a pesquisa científica na região amazônica.

6. Disponibilidade futura dos dados

A perenidade de um acervo científico requer resiliência técnica, institucional e financeira. A Tabela 3 sintetiza três vetores de risco e as soluções que a plataforma já adota (ou planeja adotar) para mitigar cada cenário.

Vetor de risco	**Estratégias de mitigação**	**Benefício esperado**
Interrupção de financiamento operacional	• Ativar modo *read-only*, reduzindo CPU e disco, mas mantendo catálogo e downloads. • Criar fundo de contingência que cubra os custos mínimos de hospedagem em *cold archive*. • Firmar convênio institucional que mantenha DOI, DNS e *object store* essenciais.	Serviço permanece disponível, assegurando resolução de DOIs e acesso aos dados publicados.
Abandono da equipe de desenvolvimento	• Código aberto em repositório público com documentação adequada. • Guia de implantação "one-click" em contêineres (*infrastructure-as-code*). • Governança aberta que permite novos *committers* ou *forks* compatíveis.	Qualquer instituição pode relançar a plataforma ou assumir manutenção, evitando obsolescência.
Perda ou corrupção de objetos	• *Backup* de dados recorrente. • Auditoria periódica DOI $\rightarrow$ URL com relatório público de links quebrados.	Múltiplas cópias verificadas asseguram integridade e disponibilidade mesmo em falhas catastróficas.

Tabela 3. Mecanismos de sustentabilidade e preservação do acervo

A combinação de redundância técnica, código aberto replicável e compromissos institucionais alinha-se às recomendações FAIR para preservação de longo prazo, garantindo que os dados permaneçam localizáveis, acessíveis e citáveis mesmo em cenários adversos.

7. Trabalhos correlatos

Este capítulo compara a plataforma DataMap com outros repositórios e sistemas voltados à gestão de dados ambientais, com foco na Amazônia. O objetivo é situar a DataMap em relação a outras iniciativas e destacar os avanços e lacunas que ela busca preencher no contexto da ciência aberta.

O ARM (*Atmospheric Radiation Measurement*) é uma referência em dados atmosféricos, com destaque para o projeto *GoAmazon*, que coleta informações sobre a interação

entre atmosfera e floresta amazônica. Apesar de robusto, o ARM possui escopo global e não é adaptado às especificidades dos dados amazônicos.

O ATTO, por sua vez, é dedicado à Amazônia, com dados coletados em uma torre de 325 metros na floresta. No entanto, sua abrangência é geograficamente restrita, enquanto o DataMap contempla uma variedade maior de dados ambientais da região, incluindo informações atmosféricas, de biodiversidade, clima e solo, com foco em acessibilidade e colaboração.

Outras plataformas relevantes incluem o MapBiomas, que monitora o uso da terra no Brasil, mas não contempla dados atmosféricos ou interações climáticas; e o Brazil Data Cube, centrado em sensoriamento remoto, mas sem estrutura para repositórios integrados de dados ambientais.

A DataMap se diferencia por sua abordagem holística, permitindo armazenamento, curadoria e publicação de diversos tipos de dados em um único repositório. Segue os princípios FAIR, com camadas adicionais de curadoria, proveniência e segurança, ausentes em plataformas como ARM ou ATTO. Também oferece geração de DOIs para cada versão de *dataset*, promovendo rastreabilidade e reutilização com maior transparência.

Comparada às demais, a DataMap apresenta arquitetura mais flexível, voltada à interoperabilidade e acessibilidade, favorecendo a colaboração entre pesquisadores, gestores e tomadores de decisão. Mantém os dados em um ciclo de vida estruturado, da ingestão à publicação citável, com foco em integridade e reusabilidade, aspectos centrais da ciência aberta.

Em síntese, embora ARM e ATTO ofereçam dados valiosos para a pesquisa ambiental na Amazônia, a DataMap propõe uma solução mais integrada e alinhada às necessidades atuais da infraestrutura de dados ambientais no Brasil, fomentando a ciência aberta e a cooperação internacional.

8. Considerações Finais

A plataforma Datamap foi projetada para superar os desafios da gestão de dados ambientais da Amazônia, promovendo a interoperabilidade e o acesso aberto a dados científicos, alinhada aos princípios FAIR. Sua implementação permitiu a organização e publicação de dados de projetos como LBA, AmazonFACE, ATTO e CHUVA, gerando 165 DOIs e armazenando mais de 1,58 TB de dados. A plataforma resolveu questões de dispersão e falta de infraestrutura, proporcionando um ecossistema coeso e acessível para a colaboração científica global.

A plataforma ainda precisa de melhorias na organização e categorização dos dados, além de ajustes na interface de usuário para facilitar o acesso. A expansão para incluir mais fontes de dados, além de fontes de dados heterogêneas, ingestão de dados assíncronos e o aprimoramento contínuo da curadoria e proveniência são essenciais para consolidar o DataMap como uma referência no gerenciamento de dados ambientais da Amazônia.

9. Agradecimentos

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

Referências

ARM (2025). Goamazon 2014/2015. Accessed: 2025-04-27.

Brickley, D., Burgess, M., and Noy, N. (2019). Google dataset search: Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference (WWW '19)*, pages 1365–1375.

Fleming, L. et al. (2017). Big data in environment and human health. *Oxford Research Encyclopedia of Environmental Science*. Accessed: 2025-04-23.

Foley, J. A. et al. (2007). Amazonia revealed: forest degradation and loss of ecosystem goods and services in the amazon basin. *Frontiers in Ecology and the Environment*, 5(1):25–32.

Garduño-Magaña, A. (2024). Infrastructure and awareness landscape analysis in latin america. Accessed: 2025-04-23.

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

Michener, W. K. et al. (2011). Dataone: Observing the earth with integrated, accessible, and reusable data. *Journal of eScience Librarianship*. Accessed: 2025-04-27.

RE3DATA (2025). Repositórios de dados nos estados unidos. Accessed: 2025-04-27.

UNESCO (2021). The role of science, technology and innovation in ensuring food security by 2030. Accessed: 2025-04-23.

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