`Triptych`: Integrando Dados, Código e Publicações para a Pesquisa Científica

Ygor Rolim[1], Liliane Kunstmann[2], Débora Pina[3], Carina F. Dorneles[4]
Marcos Lage[1], Daniel de Oliveira[1]

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

[2]Universidade Federal Rural do Rio de Janeiro (UFRRJ)

[3]Universidade Federal do Rio de Janeiro (UFRJ)

[4]Universidade Federal de Santa Catarina (UFSC)

ygorcarlos@id.uff.br,liliane.kunstmann@ufrrj.br,dbpina@cos.ufrj.br

carina.dorneles@ufsc.br,{mlage,danielcmo}@ic.uff.br

Resumo. *A Ciência Aberta promove a transparência, o compartilhamento e a reprodutibilidade na pesquisa. Atualmente, projetos científicos gerenciam dados, código e publicações em plataformas heterogêneas, como GitHub, OSF e arXiv. Embora ofereçam funcionalidades especializadas, o uso isolado dessas ferramentas fragmenta a gestão de artefatos, dificultando a rastreabilidade entre dados, software e resultados. Este artigo apresenta o* `Triptych`*, um arcabouço que integra dados, software e publicações por meio de uma camada de interoperabilidade capaz de conectar artefatos distribuídos de um mesmo projeto científico. Ao contrário de abordagens monolíticas, o* `Triptych` *integra as ferramentas já utilizadas pelos pesquisadores. Avaliações com especialistas indicam que ele melhora a integração e a rastreabilidade dos artefatos de pesquisa.*

Abstract. *Open Science promotes transparency, sharing, and reproducibility in research. Currently, scientific projects manage data, code, and publications on heterogeneous platforms such as GitHub, OSF, and arXiv. Although these tools offer specialized functionalities, their isolated use fragments artifact management, hindering traceability between data, software, and results. This article presents* `Triptych`*, a framework that integrates data, software, and publications through an interoperability layer capable of connecting distributed artifacts from the same scientific project. Unlike monolithic approaches,* `Triptych` *integrates the tools already used by researchers. Expert evaluations indicate that it improves the integration and traceability of research artifacts.*

1. Introdução

Nos últimos anos, o paradigma de Ciência Aberta (CA) [Medeiros 2021] tem promovido transparência, reprodutibilidade e compartilhamento na pesquisa científica. A UNESCO apresenta recomendações para fortalecer a CA [UNESCO 2023], destacando práticas e infraestruturas que aumentem a transparência, a confiabilidade e a integridade das pesquisas. Embora essas recomendações envolvam aspectos sociais, éticos e institucionais, a *dimensão tecnológica* é central para sua implementação. Tecnologias específicas viabilizam

a adoção de diretrizes, como os princípios FAIR (*Findable, Accessible, Interoperable and Reusable*) [Wilkinson et al. 2016], que orientam a gestão e a governança de artefatos ao longo do ciclo de pesquisa [Pereira and Pacheco 2020]. Sua implementação depende de ferramentas e infraestruturas adequadas, especialmente na gerência de dados de pesquisa [Kutha Krisnawijaya et al. 2025].

A implementação prática da CA depende da gestão adequada dos principais artefatos de um projeto científico: *(i)* dados (brutos ou processados) [Immonen et al. 2018], *(ii)* código-fonte [Flach and Kon 2021] e *(iii)* publicações [Ellerm et al. 2026]. Os dados devem seguir padrões interoperáveis, *e.g.*, JSON, XML, CSV ou HDF5 [Bonino da Silva Santos et al. 2016], e ser depositados em repositórios públicos como OSF [Foster and Deardorff 2017], Zenodo ou Dataverse [Crosas 2011], que oferecem identificadores persistentes (DOI e *Handle*) e metadados para garantir que permaneçam encontráveis, acessíveis e reutilizáveis. O código-fonte geralmente é desenvolvido em plataformas de controle de versão, como o GitHub e o GitLab, que oferecem mecanismos de versionamento e colaboração distribuída [Fraga-González et al. 2025]. Repositórios de *preprints* de acesso aberto, *e.g.*, arXiv, permitem a disseminação precoce de resultados.

Apesar da ampla adoção dessas plataformas hoje, a gestão de projetos científicos ainda permanece fragmentada. Projetos de pesquisa no contexto de experimentos computacionais tipicamente envolvem múltiplos artefatos, frequentemente armazenados em plataformas distintas. Por exemplo, os dados podem ser depositados no Zenodo ou no OSF, enquanto o código-fonte do projeto é mantido no GitHub e os *preprints* são disponibilizados no arXiv. Embora cada plataforma ofereça funcionalidades especializadas para a gestão de determinados artefatos, elas não dispõem de mecanismos integrados para representar e gerenciar esses elementos no contexto de um único projeto científico, *i.e.*, não há conexão entre os artefatos nas diversas plataformas, exceto pelo conhecimento tácito dos pesquisadores envolvidos.

Essa fragmentação gera desafios à transparência e à reprodutibilidade, *i.e.*, à implementação dos princípios FAIR. Em especial, a falta de mecanismos que conectem explicitamente dados, códigos e publicações em torno de um mesmo projeto científico dificulta a compreensão de como os resultados foram obtidos [Mattoso et al. 2010]. Como consequência, pesquisadores podem ter dificuldade para reconstruir *workflows*, rastrear e reproduzir resultados ou reutilizar artefatos. Embora algumas soluções proponham ambientes integrados para a gestão de projetos de pesquisa, *e.g.*, Zenodo oferece armazenamento de dados e publicação de *preprints*, essas abordagens frequentemente exigem a adoção de ecossistemas específicos de ferramentas, o que pode entrar em conflito com as escolhas de plataformas já estabelecidas pelos pesquisadores em seus projetos.

Diante desse cenário, é possível identificar uma série de problemas no ecossistema atual de infraestrutura para CA. Primeiramente, artefatos científicos são frequentemente gerenciados em plataformas distintas que não oferecem mecanismos nativos para representar explicitamente suas interdependências no contexto de um projeto científico [Fraga-González et al. 2025]. Em segundo lugar, a ausência de uma camada de integração entre essas plataformas dificulta a rastreabilidade e a compreensão do processo que conduz dos dados brutos aos resultados científicos publicados. Por fim, muitas soluções existentes buscam resolver esse problema por meio de ambientes monolíticos, exigindo que pesquisadores abandonem ferramentas já adotadas em favor de ecossistemas específicos, o que limita sua adoção na prática.

Para lidar com essas limitações, este artigo apresenta o arcabouço `Triptych`. O `Triptych` é projetado para integrar dados, código-fonte e *preprints* por meio de uma camada de integração capaz de conectar artefatos distribuídos em diferentes plataformas. Diferentemente de abordagens existentes que exigem a adoção de um ecossistema específico, o `Triptych` foi concebido para interoperar com ferramentas já utilizadas na prática científica, como GitHub, Zenodo e arXiv. Essa estratégia permite que pesquisadores continuem utilizando suas plataformas preferidas, enquanto o `Triptych` estabelece conexões explícitas entre os artefatos de um projeto científico por meio de um banco de dados de proveniência [Herschel et al. 2017]. Para avaliar o `Triptych`, foi realizada uma avaliação por especialistas que já conduzem projetos nos quais dados, código-fonte e publicações estão distribuídos em diferentes plataformas. Os resultados indicam que a abordagem proposta é prática e alinhada às necessidades dos pesquisadores.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta os conceitos relacionados ao projeto científico e os trabalhos relacionados. A Seção 3 descreve o arcabouço `Triptych` e sua arquitetura. A Seção 4 apresenta os resultados da avaliação realizada com especialistas. Por fim, a Seção 5 apresenta as conclusões e as direções para trabalhos futuros.

2. Referencial Teórico e Trabalhos Relacionados

Projeto Científico. Um Projeto Científico é definido como uma unidade organizacional de alto nível, responsável por coordenar objetivos, metodologias, experimentos e resultados de pesquisa [Ramos et al. 2021]. Embora o conceito de Ciência Aberta abranja qualquer tipo de projeto científico, o foco deste trabalho está em projetos que envolvem experimentos *in silico*, nos quais a produção científica resulta da combinação de três elementos principais: *(i)* dados, abrangendo desde entradas brutas até resultados processados; *(ii)* código-fonte, incluindo especificações de *workflows* via sistemas como Airflow ou *scripts* (por exemplo, em Python) que formalizam a lógica de pesquisa; e *(iii)* publicações, contendo o conteúdo textual. Para corroborar ou refutar uma hipótese, um projeto científico pode incluir um ou mais experimentos, cada um seguindo um ciclo de vida bem definido, conforme descrito por [Mattoso et al. 2010]. Como múltiplos experimentos podem coexistir em um mesmo projeto, seus ciclos de vida podem se sobrepor, exigindo gestão cuidadosa das interações realizadas pelos cientistas ao longo do tempo.

Ciclo de Vida do Experimento. O ciclo de vida de um experimento pode ser sintetizado em três etapas principais: *(i)* Composição, *(ii)* Execução e *(iii)* Análise. A composição corresponde à definição da estrutura lógica do experimento, incluindo a sequência de atividades, os dados de entrada e saída e os parâmetros de configuração. A execução é a etapa em que o código é processado em um ambiente computacional, utilizando os parâmetros definidos para gerar novos conjuntos de dados. Por fim, a análise envolve a interpretação e avaliação dos resultados, geralmente culminando na produção científica. Todas as informações de cada execução devem ser registradas no contexto do mesmo projeto, ou seja, o protocolo de experimentação deve conter referências a dados brutos e processados, grafos de dependência, código-fonte utilizado e *preprints* associados. Embora essas etapas possam ocorrer tanto de forma sequencial quanto de forma iterativa, a fragmentação dos artefatos produzidos em cada uma delas impõe desafios à reprodutibilidade. Nesse cenário, dados de proveniência [Freire et al. 2008] podem atuar como elo integrador.

Alinhamento com Princípios FAIR. A rastreabilidade é fundamental para o alinhamento

do projeto aos Princípios FAIR [Wilkinson et al. 2016], que visam maximizar a utilidade de artefatos, dados, código e publicações [Medeiros 2021]. Em um projeto científico, a conformidade com os princípios FAIR depende de: tornar os dados localizáveis (*Findable*) por meio de metadados enriquecidos que descrevam o contexto da coleta; acessíveis (*Accessible*), garantindo que o caminho para obter os dados e o código correspondente esteja documentado; interoperáveis (*Interoperable*), utilizando modelos de proveniência padronizados (*e.g.*, W3C PROV-DM) para descrever relações entre dados de diferentes atividades e *workflows*; e reutilizáveis (*Reusable*), fornecendo o contexto necessário, *i.e.*, parâmetros, versões de bibliotecas e proveniência, para que outros possam replicar ou estender o experimento com confiança.

Trabalhos Relacionados. Diversas iniciativas têm sido propostas para apoiar a gestão de artefatos científicos no contexto da CA, especialmente no que diz respeito ao gerenciamento de dados de pesquisa e à promoção de transparência e reprodutibilidade. O DataMap [Maia et al. 2025] é um repositório voltado à gestão de dados científicos, com ênfase na região amazônica. A plataforma oferece suporte à catalogação, à descoberta, à visualização e à publicação de *datasets*, incluindo a geração automática de DOIs. Sua abordagem concentra-se na organização e exposição de metadados em nível de *dataset*, bem como na representação das etapas do *workflow*. No entanto, o DataMap não modela explicitamente relações de proveniência em nível mais granular entre diferentes tipos de artefatos científicos. No contexto de domínios específicos, a proposta apresentada em [da Cruz et al. 2018] define uma infraestrutura para apoiar práticas de CA na área de ciência do solo. Essa abordagem enfatiza o compartilhamento, a organização e o acesso aberto a dados e resultados científicos, incorporando também aspectos de proveniência relacionados ao processo de curadoria dos dados. Embora contribua para a transparência e colaboração dentro de um domínio específico, sua aplicabilidade é mais restrita e não aborda de forma abrangente a integração entre múltiplos tipos de artefatos distribuídos.

Outra abordagem relevante é o RO-Crate [Soiland-Reyes et al. 2022], que propõe o empacotamento de todos os recursos envolvidos na produção de um resultado científico em uma estrutura única e autocontida. Utilizando um modelo baseado em anotações do *Schema.org* e representado em JSON-LD, o RO-Crate descreve identificadores, relações, proveniência e metadados dos artefatos, contribuindo para a reprodutibilidade e a organização de objetos de pesquisa. [Ellerm et al. 2026] utiliza a proveniência gerada no RO-Crate para ampliar o contexto de artigos científicos para além de arquivos estáticos, provendo verificabilidade do texto científico com os fragmentos de dados e códigos através do uso de LLMs que ancoram as afirmações do texto nos dados de proveniência. Entretanto, abordagens como RO-Crate pressupõem a agregação dos artefatos em um único pacote, o que pode limitar sua aplicação em cenários onde os recursos permanecem distribuídos em diferentes plataformas. Na contramão dessas abordagens, o `Triptych` adota uma perspectiva centrada na proveniência e na integração distribuída. Em vez de exigir a centralização ou o empacotamento dos artefatos, a proposta estabelece conexões explícitas entre dados, código-fonte e *preprints* mantidos em suas plataformas originais.

3. O Arcabouço `Triptych`

Esta seção apresenta os principais aspectos do `Triptych`. Inicialmente, são descritos seus requisitos e apresentados os objetivos que orientam sua concepção e uso. Em seguida, discute-se sua arquitetura, destacando os componentes que a compõem e o modelo de proveniência adotado.

3.1. Elicitação dos Requisitos

Uma série de requisitos foi elicitada (Tabela 1) a partir de diversas reuniões com pesquisadores, com o objetivo de orientar e fundamentar o desenvolvimento do `Triptych`. Esses requisitos surgiram de discussões estruturadas com pesquisadores experientes na condução e gestão de projetos científicos das áreas de Ciência da Computação, Ciência da Informação e Biologia, proporcionando uma perspectiva multidisciplinar na identificação das necessidades do `Triptych`. De forma geral, espera-se que o arcabouço ofereça mecanismos adequados para a organização, preservação e compartilhamento de diferentes tipos de artefatos de pesquisa, *i.e.*, dados, código-fonte e *preprints*, em um ambiente integrado. Além disso, o arcabouço deve favorecer o reúso de artefatos gerados em pesquisas anteriores.

Tabela 1. Resumo dos Requisitos Funcionais do `Triptych`

ID	Descrição
RF1	Suporte a armazenamento de múltiplos formatos de artefatos
RF2	Disponibilidade e reúso de artefatos ao longo do tempo
RF3	Publicação de *preprints* a partir de projetos
RF4	Integração com plataformas externas de publicação
RF5	Gestão de projetos de pesquisa
RF6	Representação de projetos previamente desenvolvidos
RF7	Suporte ao trabalho colaborativo
RF8	Relacionamento entre projetos e artefatos
RF9	Busca e indexação por palavras-chave

Do ponto de vista dos Requisitos Funcionais (RF), o `Triptych` deve permitir o armazenamento de artefatos provenientes de experimentos científicos, de análises computacionais e de testes relacionados ao desenvolvimento de modelos. Esses artefatos incluem não apenas dados brutos, mas também arquivos de diferentes naturezas, *e.g.*, documentação técnica, códigos-fonte, *scripts*, executáveis de modelos e materiais de apoio que permitam compreender e reproduzir os resultados obtidos. Assim, o `Triptych` deve oferecer suporte ao armazenamento de múltiplos formatos de arquivos, garantindo que diversos tipos de artefatos possam ser mantidos e organizados em um único ambiente (RF1).

Além do armazenamento, o `Triptych` deve garantir que os artefatos permaneçam disponíveis ao longo do tempo e sejam reutilizados em novos projetos (RF2). Esse requisito é especialmente relevante em contextos acadêmicos, em que os resultados de experimentos anteriores frequentemente servem de base para novas investigações. Outro requisito funcional diz respeito à publicação de *preprints*: o `Triptych` deve permitir que pesquisadores registrem versões preliminares de manuscritos a partir de dados e projetos já cadastrados, o que simplifica a divulgação dos artigos (RF3). Adicionalmente, o `Triptych` deve possibilitar integração com plataformas externas de publicação acadêmica, permitindo o envio automatizado de *preprints* (RF4).

O `Triptych` deve oferecer funcionalidades para a gestão de projetos de pesquisa (RF5), incluindo a criação, edição e manutenção de objetivos, metodologias, resultados e recursos de cada projeto. Um aspecto importante é a capacidade de representar projetos previamente desenvolvidos, permitindo que pesquisadores consolidem dados, código-fonte e *preprints* distribuídos em diferentes plataformas (RF6). Ainda no âmbito dos RFs, o sistema deve possibilitar trabalho colaborativo entre múltiplos pesquisadores (RF7) e permitir o estabelecimento de relações entre projetos, dados , código-fonte e *preprints*, de modo que os usuários identifiquem facilmente quais dados foram utilizados em determinado projeto ou

quais *preprints* estão associados (RF8). Por fim, o `Triptych` deve disponibilizar mecanismos de busca e indexação baseados em palavras-chave, termos ou assuntos, facilitando a localização de projetos, dados e publicações (RF9).

No que se refere aos Requisitos Não Funcionais (RNF), a interface do `Triptych` deve ser projetada de forma intuitiva, reduzindo a complexidade de navegação e o número de etapas necessárias para registrar ou acessar informações (RNF1). Outro requisito relevante é a escalabilidade do sistema: a plataforma deve ser capaz de lidar com grandes volumes de dados de pesquisa sem comprometer o desempenho, considerando que experimentos científicos frequentemente geram quantidades significativas de dados (RNF2).

3.2. Arquitetura do `Triptych`

A arquitetura do `Triptych` (Figura 1) foi projetada para atender aos requisitos apresentados na Sub-Seção 3.1, e é organizada em quatro camadas: *(i)* Serviços Externos, *(ii)* Camada de Aplicação, *(iii)* Camada de Dados e *(iv)* Camada de Apresentação. A camada de *Serviços Externos* inclui todas as plataformas invocadas pelo `Triptych` para execução de funcionalidades, como armazenamento de dados, hospedagem de código-fonte ou publicação de *preprints*. Para cada serviço integrado, o `Triptych` disponibiliza um adaptador que media a comunicação entre a camada de aplicação e o serviço externo correspondente, abstraindo detalhes de implementação e garantindo interoperabilidade. Nativamente, o `Triptych` oferece adaptadores para GitHub, Zenodo, arXiv e OSF. Adicionalmente, um adaptador para o Maritaca.AI (`https://www.maritaca.ai/`) permite a implementação de tradução automática, detalhada a seguir.

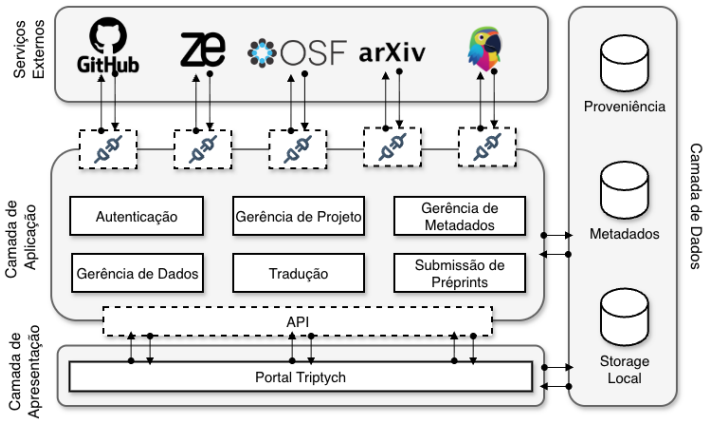

Figura 1. Arquitetura conceitual do arcabouço `Triptych`.

A *Camada de Dados* é composta por três repositórios: *(i) Storage* Local, *(ii)* Repositório de Metadados e *(iii)* Repositório de Proveniência. O *Storage* Local armazena arquivos associados aos projetos, incluindo diferentes tipos, como imagens, vídeos e documentos. O banco de dados de proveniência segue o modelo apresentado na Figura 2, no qual a entidade `Projeto` ocupa a posição central, atuando como núcleo estruturador de todo o esquema. Esse repositório armazena informações sobre os projetos registrados, incluindo dados dos usuários envolvidos e dos serviços externos associados a cada projeto. Dessa forma, o modelo assegura a rastreabilidade e o controle dos diversos recursos vinculados a um projeto no `Triptych`.

Adicionalmente, o banco de dados de proveniência armazena informações complementares, como arquivos associados ao projeto, tópicos de classificação e credenciais de

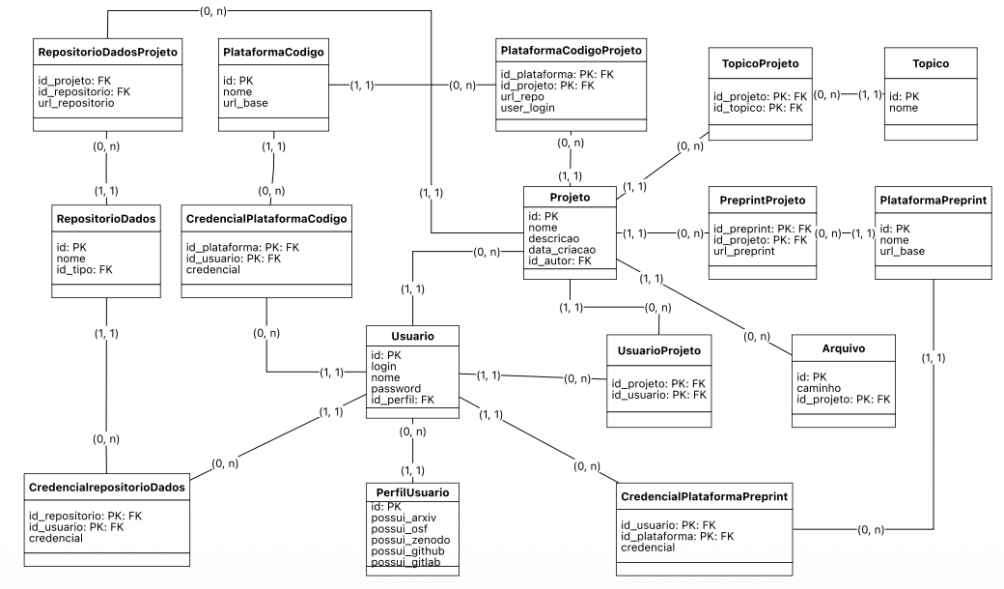

Figura 2. Esquema do Banco de Dados de Proveniência do `Triptych`.

acesso necessárias para integração com serviços externos. Esses elementos contribuem para a organização e a gestão completas do ciclo de vida de cada projeto [Mattoso et al. 2010]. Além disso, o `Triptych` mantém metadados associados aos artefatos do projeto, seguindo padrões já utilizados em repositórios de dados de pesquisa, como o *Dublin Core Metadata Element Set*, o *DataCite Metadata Schema* e o padrão *ISO 19115:2003(E)*. Esses esquemas permitem que os dados sejam identificáveis, acessíveis e reutilizáveis. No `Triptych`, os metadados são armazenados em XML [Digital Curation Centre 2024] e JSON/JSONB, garantindo interoperabilidade com diferentes sistemas e possibilitando consultas avançadas [Petkovic 2017].

A *Camada de Aplicação* constitui o núcleo do `Triptych`, sendo responsável pela implementação das regras de negócio. Essa camada recebe requisições do portal do `Triptych` e coordena a execução das operações necessárias, estabelecendo comunicação tanto com a Camada de Dados quanto com os Serviços Externos. Dois aspectos arquiteturais merecem destaque. Primeiro, a natureza *stateless* do `Triptych`. Os componentes da camada de aplicação não mantêm estado associado à sessão do usuário; cada requisição é tratada de forma independente.Nesse modelo, cada chamada representa uma transação completa: a requisição é recebida pelo portal, a transação é iniciada, as operações interagem com a camada de dados e, ao final, a resposta correspondente é retornada ao cliente [Bernstein and Newcomer 2009]. Essa abordagem favorece escalabilidade, simplicidade de manutenção e maior robustez em ambientes distribuídos. Segundo, o padrão de organização baseado no *design pattern Transaction Script*, no qual o código de cada transação é estruturado em módulos ou funções independentes, cada um com objetivo bem definido, permitindo separar claramente as operações relacionadas às diferentes funcionalidades do `Triptych`.

A camada de aplicação do `Triptych` é composta pelos seguintes componentes: *(i)* Autenticação, *(ii)* Gerência de Projetos, *(iii)* Gerência de Dados, *(iv)* Gerência de Metadados, *(v)* Submissão de *Preprints* e *(vi)* Tradução. O componente de *Autenticação* é responsável pela criação e verificação de usuários. Ele recebe requisições cujo corpo é estruturado em JSON, contendo atributos como nome de usuário, senha e endereço de e-mail. A função verifica a existência do usuário por meio de consultas à tabela de usuários armazenada no

banco de dados de proveniência da camada de dados. Caso não haja registro associado ao e-mail informado, um novo usuário é automaticamente cadastrado.

O componente de *Gerência de Projetos* é responsável pela criação e administração de projetos. No `Triptych`, um projeto pode englobar diferentes tipos de atividades e artefatos, como *preprints*, artigos e quaisquer outros conteúdos que o usuário deseje organizar e enriquecer com metadados. As requisições recebidas por esse componente utilizam o formato *form-data* e incluem dois atributos principais: o objeto JSON *project*, contendo nome, descrição, identificador do autor (usuário) e identificador do tópico associado, e o atributo *files*, referente aos arquivos que o usuário deseja anexar ao projeto. Esses arquivos podem incluir imagens, textos ou outros conteúdos, que são posteriormente renderizados juntamente com a descrição do projeto em *Markdown*. Ao receber a requisição, o `Triptych` valida o *token* de autorização e verifica se o projeto já existe no banco de dados de proveniência. Caso o projeto ainda não exista, o sistema cria, quando houver arquivos anexos, um diretório correspondente no repositório de dados local.

O componente de *Gerência de Dados* é responsável pelo carregamento de arquivos para os repositórios externos integrados à plataforma, incluindo Zenodo, OSF e o repositório de código-fonte do GitHub. As requisições direcionadas a esse componente utilizam o formato *x-www-form-urlencoded* e incluem atributos como o identificador do projeto remoto no repositório de destino, os *tokens* de autorização para autenticação nas plataformas externas e o conteúdo do arquivo a ser enviado. Com base nessas informações, o sistema constrói dinamicamente a requisição adequada para a API correspondente e envia o arquivo para o destino especificado.

Para aumentar a robustez durante a integração com serviços externos, o componente de *Gerência de Dados* implementa um mecanismo de retentativas automáticas em caso de falhas nas requisições. Esse mecanismo utiliza a estratégia conhecida como *Exponential Backoff*, que aumenta progressivamente o intervalo entre tentativas subsequentes após uma falha, evitando sobrecarga nos serviços externos e reduzindo o risco de falhas adicionais [Ve et al. 2024]. O componente de *Gerência de Metadados* permite associar metadados a arquivos previamente submetidos ao `Triptych` ou a repositórios externos integrados, complementando a descrição dos recursos armazenados e facilitando sua identificação, organização e posterior recuperação.

O componente de *Tradução* é responsável pela tradução e correção gramatical do conteúdo textual associado à submissão de *preprints* à plataforma arXiv e aos textos da página do projeto. As requisições enviadas a esse componente possuem conteúdo em JSON, contendo um objeto denominado *content*, com os subatributos título e assunto. Ao receber a requisição, o sistema envia essas informações para a plataforma Maritaca.AI, construindo um *prompt*, ou seja, um texto com instruções específicas que orientam o modelo de linguagem a realizar a tradução e a revisão gramatical. Como resultado, o texto original é convertido para o inglês e aprimorado linguisticamente antes de ser utilizado na submissão do artigo ou na publicação da página do projeto.

O componente de *Submissão de Preprints* é responsável pela submissão automatizada de *preprints* em plataformas como o arXiv. As requisições enviadas a esse componente possuem corpo em formato JSON, contendo atributos como o nome de usuário na plataforma de *preprints*, a senha da conta, o assunto do artigo, o título, a lista de autores, o resumo e o conteúdo do artigo, compactado em um arquivo .zip. O processo é realizado com o auxílio

da ferramenta de automação *Selenium WebDriver*, que simula interações com a interface *web* da plataforma. Inicialmente, o arquivo .zip é temporariamente armazenado em um diretório interno do servidor da camada de dados. Em seguida, o *Selenium WebDriver* executa automaticamente as etapas necessárias para concluir a submissão.

Cada componente do `Triptych` é acessado por meio de rotas de *endpoints* fornecidas pela camada de aplicação. Essas rotas correspondem a pontos de acesso da API que permitem ao portal do sistema interagir com os componentes da camada de aplicação. Por meio desses endereços, o portal pode solicitar operações específicas, como consultar dados armazenados, modificar informações existentes ou gerar novos conteúdos. Assim, cada requisição enviada pelo portal é direcionada a um *endpoint* específico, que aciona o componente correspondente para processar a operação e retornar a resposta apropriada.

Finalmente, a *Camada de Apresentação* contém o Portal `Triptych`, que corresponde à interface por meio da qual os usuários interagem. Nesse portal (Figura 3), os usuários podem acessar funcionalidades do arcabouço, como a criação e a gestão de projetos, a associação de arquivos e metadados, a integração com repositórios externos e a submissão de *preprints*. O portal organiza e apresenta as informações, além de coletar as requisições dos usuários e encaminhá-las à camada de aplicação por meio dos *endpoints* da API. Assim, o Portal `Triptych` atua como mediador entre o usuário e os serviços internos do sistema. O código-fonte do `Triptych` pode ser obtido em `https://github.com/UFFeScience/triptych`.

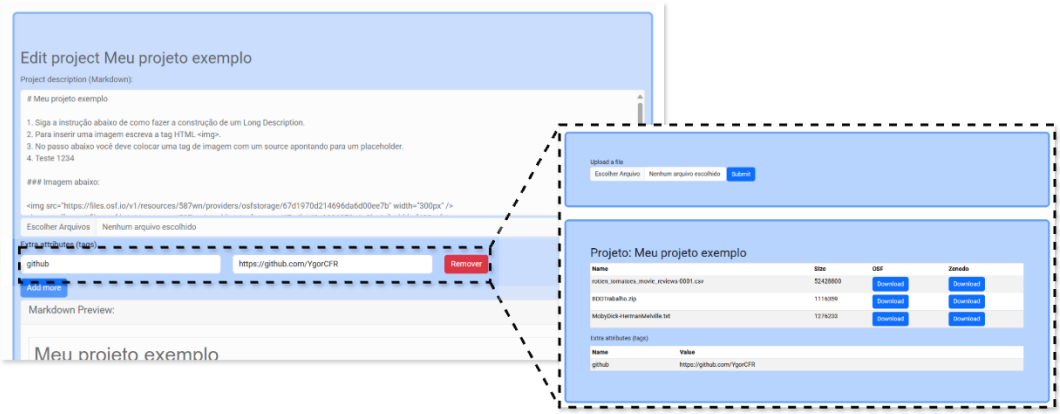

Figura 3. Interface do Portal `Triptych`

4. Avaliação do `Triptych`

Para avaliar o `Triptych`, adotaram-se duas abordagens: qualitativa e quantitativa. A avaliação qualitativa teve como objetivo investigar a percepção de gestores experientes de projetos científicos acerca do arcabouço, buscando compreender sua utilidade, aplicabilidade e potencial em projetos de pesquisa. Já a avaliação quantitativa objetivou mensurar a sobrecarga adicionada pelo `Triptych`. Para tal, o sistema foi implantado em uma instância `t2.2xlarge` da AWS, com 8 vCPUs e 32 GiB de RAM.

O Modelo TAM. Para avaliar a experiência dos usuários gestores de projeto com o `Triptych`, adotou-se o TAM (*Technology Acceptance Model*) [Davis 1989]. O TAM é utilizado para analisar fatores que influenciam a aceitação e o uso de uma tecnologia por indivíduos ou grupos, examinando variáveis relacionadas a atitudes e comportamentos frente aos sistemas de informação. Neste artigo, são considerados dois fatores [Davis 1989]: Facilidade Percebida de Uso (*Perceived Ease of Use* – PEOU), que indica o grau em que o usuário

acredita que utilizar o sistema é simples, e Utilidade Percebida (*Perceived Usefulness* – PU), que reflete o quanto o usuário acha que o `Triptych` melhora seu desempenho nas tarefas. A Tabela 2 apresenta as questões utilizadas na avaliação.

Tabela 2. Questões utilizadas na avaliação baseada no TAM

ID	Pergunta	Tipo
Q1	Qual o grau de aplicabilidade do Triptych nas suas rotinas diárias?	Utilidade
Q2	Suas atividades de publicação de *preprints* no Arxiv seriam aceleradas com seu uso?	Utilidade
Q3	Compartilhar materiais de pesquisa com várias plataformas é eficaz?	Utilidade
Q4	Em que medida o Triptych substituiria atividades manuais?	Utilidade
Q5	Qual o grau de facilidade de uso do Triptych?	Facilidade
Q6	Qual é a satisfação com o aprendizado para uso do Triptych?	Facilidade
Q7	É fácil lembrar todas as etapas necessárias para usar o sistema?	Facilidade
Q8	É fácil identificar erros no Triptych?	Facilidade

Caracterização dos Participantes. O estudo contou com quatro participantes, número intencionalmente reduzido devido ao perfil especializado necessário para a avaliação do `Triptych`. Todos são pesquisadores com qualificação acadêmica e experiência em atividades de pesquisa. Os participantes possuem título de doutorado e mais de dez anos de atuação em pesquisa, o que lhes confere familiaridade com práticas de gestão de projetos, de produção acadêmica e de uso de ferramentas de apoio à pesquisa. Entre os participantes, dois são bolsistas de produtividade do CNPq. Além disso, três dos quatro participantes já coordenaram ou coordenam projetos financiados pelo CNPq, evidenciando experiência direta na gestão de recursos, na organização de atividades de pesquisa e no acompanhamento dos resultados científicos.

Protocolo de Avaliação. Para apoiar a participação no estudo, foi elaborado um manual de instruções do `Triptych`, disponibilizado a todos os participantes. A avaliação ocorreu ao longo de dois meses, período em que os participantes puderam explorar as funcionalidades do sistema conforme sua disponibilidade, tendo todos os mesmos privilégios de uso e as mesmas condições de acesso. Em particular, os participantes foram orientados a representar no sistema o "espelho" de projetos que já gerenciavam, realizando as seguintes tarefas: *(i)* cadastrar-se no `Triptych`; *(ii)* efetuar login; *(iii)* conectar-se ao Zenodo; *(iv)* conectar-se ao OSF; *(v)* conectar-se ao GitHub; *(vi)* acessar a funcionalidade de gestão de projetos e criar um novo projeto; *(vii)* verificar os arquivos e metadados associados ao projeto; *(viii)* enviar arquivos para os repositórios de dados e de código-fonte integrados; *(ix)* criar atributos adicionais e metadados para o projeto; *(x)* editar as informações do projeto; e *(xi)* realizar a submissão de um *preprint* no arXiv. Essas atividades foram definidas para permitir que os participantes experimentassem diferentes aspectos funcionais do `Triptych`.

Resultados da Avaliação Qualitativa. A Figura 4 apresenta a distribuição das respostas na escala *Likert* para cada questão da Tabela 2. As respostas dos participantes foram organizadas em uma escala de cinco níveis, permitindo avaliar o grau de concordância ou de percepção em relação aos aspectos analisados. As opções de resposta foram: *(i)* Muito Baixo, *(ii)* Baixo, *(iii)* Médio, *(iv)* Alto e *(v)* Muito Alto, permitindo análise comparativa das percepções sobre os fatores avaliados. Observa-se uma concentração predominante nas categorias *Médio* e *Alto*, indicando percepção geral moderadamente positiva do `Triptych`. As questões relacionadas à facilidade de uso (Q5–Q8) apresentaram maior frequência no nível *Alto*, especialmente na identificação de erros e no aprendizado do sistema. Por outro lado, as questões referentes à utilidade percebida (Q1–Q4) apresentaram maior concentração no nível *Médio*,

sugerindo que, embora o `Triptych` seja considerado utilizável, ainda há oportunidades de ampliar seu impacto prático.

Uma das oportunidades de melhoria identificadas refere-se à publicação de *preprints*. Alguns participantes relataram dificuldades em compreender claramente o processo de submissão, mencionando a falta de transparência na etapa de publicação e a ausência de informações detalhadas em caso de erros. Além disso, a integração com o arXiv depende de automação da interface, tornando-a sensível a alterações no *layout* ou no código HTML da plataforma. Para complementar a análise, a Figura 5 apresenta a média ponderada das respostas, considerando a escala Likert de 5 pontos (1 = Muito Baixo, 5 = Muito Alto). A média foi calculada pela expressão $\bar{x} = \frac{\sum f_i x_i}{N}$, em que f_i representa a frequência de respostas de Q1 a Q8, x_i o valor na escala Likert e N o total de respostas. Observa-se que as maiores médias concentram-se nas questões relacionadas à facilidade de uso, especialmente na identificação de erros no sistema (3,50). Por outro lado, as questões relacionadas à aceleração de publicações apresentam médias mais moderadas (2,50).

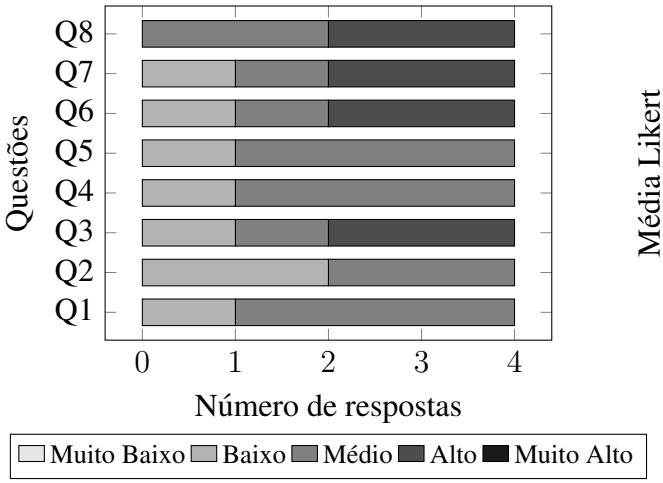

Figura 4. Respostas do TAM

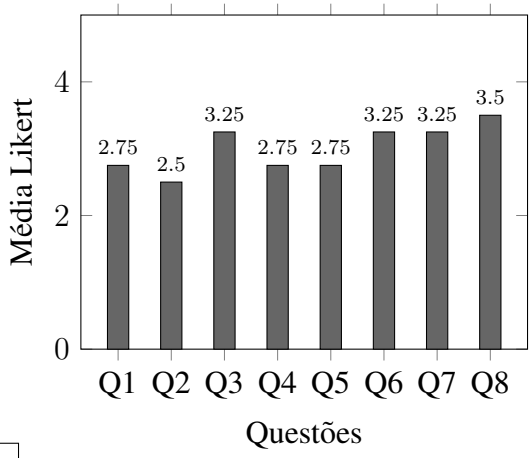

Figura 5. Média das respostas

Para avaliar a consistência interna do questionário, utilizou-se o coeficiente alfa de Cronbach [Cronbach 1951]. Considerando as oito questões do TAM ($k = 8$), o coeficiente foi calculado por meio da equação $\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_T^2}\right)$, onde k representa o número de itens do questionário, σ_i^2 a variância das respostas do item i, e σ_T^2 a variância da soma total das respostas. Obteve-se um valor aproximado de $\alpha \approx 0.96$. Valores de α superiores a 0,9 indicam consistência interna dos itens do questionário. Esse resultado sugere que as questões apresentam forte correlação entre si e são adequadas para avaliar de forma consistente a percepção dos usuários quanto à utilidade e à facilidade de uso.

Resultados da Avaliação Quantitativa. A avaliação quantitativa foi conduzida por meio da análise da sobrecarga adicional no envio de arquivos de diferentes tamanhos para os repositórios Zenodo e OSF, utilizando o `Triptych`. Foram realizados testes de desempenho para observar o comportamento do `Triptych` em cenários que simulam diferentes usos pelos usuários. O foco principal da avaliação foi a carga de arquivos CSV, escolhidos por serem amplamente utilizados no armazenamento e no compartilhamento de dados estruturados. Os resultados apresentados na Figura 6 mostram o tempo de carga em função do tamanho do arquivo, comparando execuções com e sem o uso do `Triptych`. De forma geral, observa-se que o tempo de carregamento aumenta à medida que o tamanho dos dados cresce, como

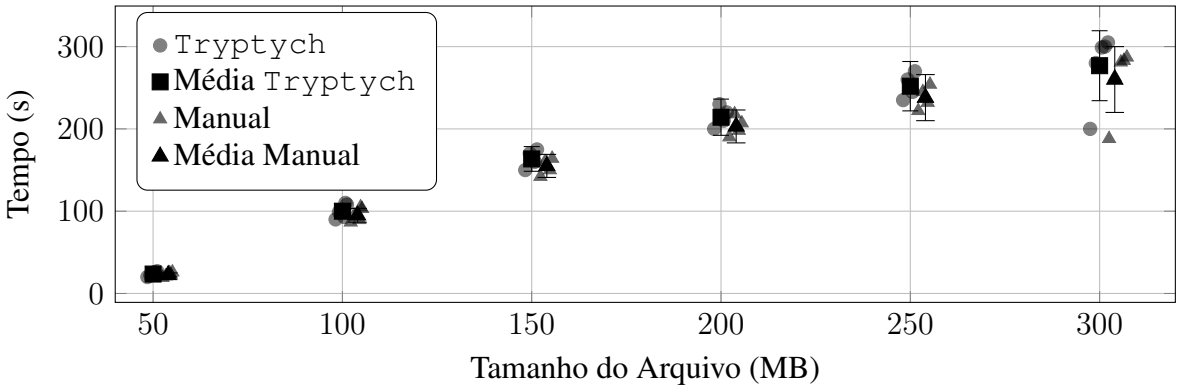

Figura 6. Avaliação da sobrecarga adicionada pelo `Triptych`.

esperado, em ambos os cenários. Para arquivos de aproximadamente 50 MB, por exemplo, o tempo médio passa de cerca de 23,6 segundos com o `Triptych` para 22,5 segundos sem o uso do `Triptych`. Essa diferença torna-se mais evidente à medida que o volume de dados aumenta, alcançando aproximadamente 276,8 segundos com o `Triptych` e 260 segundos sem ele, em arquivos de 300 MB. Apesar dessa diferença, a dispersão observada nos pontos experimentais e nas barras de erro indica variação entre as execuções, o que sugere que a sobrecarga introduzida pelo `Triptych` é relativamente pequena em relação ao custo total de carregamento.

5. Conclusões

Projetos de pesquisa frequentemente dependem de múltiplas plataformas para armazenar dados, código-fonte e publicações, o que resulta em um ecossistema fragmentado. Este artigo apresentou o `Triptych`, um arcabouço de código aberto para apoiar a organização e o compartilhamento de artefatos científicos. O `Triptych` foi concebido para reunir artefatos científicos em torno do conceito de projeto científico, conectando diferentes repositórios e serviços acadêmicos. O `Triptych` possui uma arquitetura em quatro camadas que permite integrar plataformas existentes, preservando a autonomia dos pesquisadores. Oferece funcionalidades para gerenciamento de projetos, associação de metadados, armazenamento redundante de dados, gerenciamento de código-fonte e submissão automatizada de *preprints*.

A avaliação do `Triptych` combinou análise qualitativa e quantitativa. A avaliação qualitativa indicou uma percepção positiva, especialmente quanto à facilidade de uso, enquanto os aspectos relacionados à automação de tarefas e à publicação de *preprints* receberam avaliações mais moderadas, o que aponta oportunidades de melhora. A avaliação quantitativa mostrou que a sobrecarga no envio de arquivos para repositórios externos cresce linearmente com o tamanho dos dados. Entre as limitações, destacam-se o número reduzido de participantes na avaliação qualitativa e a sensibilidade de algumas funcionalidades, *e.g.*, submissão automatizada de *preprints*, a alterações nas plataformas externas. Trabalhos futuros incluem incorporar ao `Triptych` *frameworks* de reprodução de código, como o *JupyterLite*[1].

Agradecimentos

Os autores agradecem à CAPES, ao CNPq e a FAPERJ pelo apoio parcial à realização deste trabalho. Adicionalmente, os autores declaram, de forma transparente, que ferramentas baseadas em LLMs foram utilizadas exclusivamente para revisão textual.

[1] `https://jupyterlite.readthedocs.io/en/stable/`

Referências

Bernstein, P. A. and Newcomer, E. (2009). Chapter 7 - system recovery. In Bernstein, P. A. and Newcomer, E., editors, *Principles of Transaction Processing (Second Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 185–222. Morgan Kaufmann, San Francisco, second edition edition.

Bonino da Silva Santos, L. O., Wilkinson, M., Kuzniar, A., Kaliyaperumal, R., Thompson, M., Dumontier, M., and Burger, K. (2016). *FAIR Data Points Supporting Big Data Interoperability*, pages 270–279.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

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

da Cruz, S. M. S., Ceddia, M. B., Schmitz, E. A., Rizzo, G. S., Miranda, R. C. T., Cruz, S. O., Correa, A. C., Klinger, F., Marinho, E., and Cruz, P. V. (2018). Towards an e-infrastructure for open science in soils security. In *Anais do XII Brazilian e-Science Workshop*, pages 49–56, Porto Alegre, RS, Brasil. SBC.

Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13:319–340.

Digital Curation Centre (2024). What are metadata standards? Accessed: 2025-06-17.

Ellerm, A., Adams, B., and Gahegan, M. (2026). Trustworthy scientific narrative generation through computational provenance and dynamic authoring frameworks. In *Open Conference Proceedings*, volume 8.

Flach, C. v. and Kon, F. (2021). Software livre: Pré requisito para a ciência aberta. *Computação Brasil*, 46(46):12–15.

Foster, E. D. and Deardorff, A. (2017). Open science framework (OSF). *J. Med. Libr. Assoc.*, 105(2).

Fraga-González, G., van de Wiel, H., Garassino, F., Kuo, W., de Zélicourt, D., Kurtcuoglu, V., Held, L., and Furrer, E. (2025). Affording reusable data: recommendations for researchers from a data-intensive project. *Scientific Data*, 12(1):258.

Freire, J., Koop, D., Santos, E., and Silva, C. T. (2008). Provenance for computational tasks: A survey. *Computing in Science & Engineering*, 10(3).

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.

Immonen, A., Ovaska, E., and Paaso, T. (2018). Towards certified open data in digital service ecosystems. *Software Quality Journal*, 26(4):1257–1297.

Kutha Krisnawijaya, N. N., Tekinerdogan, B., Catal, C., van der Tol, R., and Herdiyeni, Y. (2025). Implementing fair principles in data management systems: A multi-case study in precision farming. *Computers and Electronics in Agriculture*, 230:109855.

Maia, A., Maia, C., and Corrêa, P. (2025). Repositório de dados para ciência aberta na região amazônica. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 303–315, Porto Alegre, RS, Brasil. SBC.

Mattoso, M., Werner, C., Travassos, G. H., Braganholo, V., Ogasawara, E. S., de Oliveira, D., da Cruz, S. M. S., Martinho, W., and Murta, L. (2010). Towards supporting the life cycle of large scale scientific experiments. *Int. J. Bus. Process. Integr. Manag.*, 5(1):79–92.

Medeiros, C. B. (2021). Ciência aberta - colaboração sem barreiras para o avanço do conhecimento. *Computação Brasil*, 46(46):8–11.

Pereira, L. and Pacheco, R. (2020). Ciclos de vida de pesquisa com base na ciência aberta. In *Anais do XIV Brazilian e-Science Workshop*, pages 65–72, Porto Alegre, RS, Brasil. SBC.

Petkovic, D. (2017). Json integration in relational database systems. *International Journal of Computer Applications*, 168(6):14–19.

Ramos, L., Porto, F., and de Oliveira, D. (2021). Managing hypothesis of scientific experiments with phenomanager. *Journal of Information and Data Management*, 12(2).

Soiland-Reyes, S., Sefton, P., Crosas, M., Castro, L. J., Coppens, F., Fernández, J. M., Garijo, D., Grüning, B., La Rosa, M., Leo, S., et al. (2022). Packaging research artefacts with ro-crate. *Data Science*, 5(2):97–138.

UNESCO (2023). UNESCO Recommendation on Open Science. https://www.unesco.org/en/open-science/about. [Accessed 22-05-2025].

Ve, P., Sai, M., Vp, V. S., Modi, S., and Ponnusamy, S. (2024). Exponential backoff: A comprehensive approach to handling failures in distributed architectures. *International Research Journal of Modernization in Engineering Technology and Science (IRJMETS)*.

Wilkinson, M. D. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, 3(1):160018.