Biblios
ISSN 1562-4730 (online)

# Avaliação da aderência dos conjuntos de dados de pesquisa do repositório Aleia aos Princípios FAIR

## Evaluation of the adherence of the research data sets in the Aleia repository to the FAIR Principles

## Evaluación de la adhesión de los conjuntos de datos de investigación del repositorio Aleia a los Principios FAIR

**Letícia Guarany Bonetti**
**Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília, Brasil**

**Tatyane Guedes Martins da Silva**
**Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília, Brasil**

**Caterina Groposo Pavão**
**Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil**

**Samile Andréa de Souza Vanz**
**Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil**

**Rene Faustino Gabriel Junior**
**Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil**

**Ana Carolina Simionato Arakaki**
**Universidade de Brasília, Brasília, Brasil**

ORIGINAL

## Resumo

**Objetivo**. Avaliar a aderência dos conjuntos de dados de pesquisa depositados no repositório Aleia quanto aos Princípios FAIR. **Método**. A partir de uma pesquisa descritiva com abordagem quali-quantitativa, os 16 *datasets* depositados no Aleia foram avaliados quanto aos Princípios FAIR. Utilizou-se a ferramenta F-UJI *Automated FAIR Data Assessment Tool*, que automatiza a avaliação das métricas do FAIR. **Resultados**. Os maiores níveis de aderência estão ligados à localização e à interoperabilidade. Quanto à localização, 100% da amostra atingiu nível avançado em quatro das cinco métricas. Já quanto à interoperabilidade, 100% dos *datasets* aderiram em nível avançado a duas das três métricas. A acessibilidade e a reutilização foram os pontos fracos da amostra. Nenhuma das cinco métricas de reutilização foi implementada em nível avançado para 100% da amostra. **Conclusões**. No geral, os *datasets* do Aleia se encontram em níveis moderados de *FAIRness*. Para atingir níveis maiores de aderência, é preciso investir em melhorias quanto à acessibilidade e à reutilização. Em compensação, os *datasets* se destacaram em relação à localização e à interoperabilidade, atingindo níveis avançados de aderência.

**Palavras-chave:** dados de pesquisa, repositório de dados de pesquisa, repositório Aleia, princípios FAIR, ciência aberta

## Abstract

**Objective**. To assess the *FAIRness* of Aleia datasets. **Method**. Based on a descriptive study with a qualitative and quantitative approach, the 16 datasets deposited in Aleia were assessed for the FAIR Principles. The F-UJI Automated FAIR Data Assessment Tool, which automates the assessment of FAIR metrics, was used. **Results**. The highest levels of *FAIRness* are linked to findability and interoperability. Regarding findability, 100% of the sample reached an advanced level in four of five metrics. Regarding interoperability, 100% of the datasets adhered to an advanced level in two of three metrics. Accessibility and

reusability were the weak points of the sample. None of the five reusability metrics were implemented at an advanced level for 100% of the sample. **Conclusions**. Overall, the Aleia datasets are at moderate levels of *FAIRness*. To achieve higher levels of *FAIRness*, it is necessary to invest in improvements in accessibility and reusability. On the other hand, datasets have excelled in terms of findability and interoperability, reaching advanced levels of compliance.

**Keywords:** research data, research data repository, Aleia repository, FAIR principles, open science

## Resumen

**Objetivo**. Evaluar la adherencia a los Principios FAIR de los conjuntos de datos de investigación depositados en el repositorio Aleia. **Metodología**. Con base en una investigación descriptiva con enfoque cualitativo y cuantitativo, los 16 conjuntos de datos depositados en Aleia fueron evaluados de acuerdo con los Principios FAIR. Se utilizó la herramienta F-UJI Automated FAIR Data Assessment Tool, que automatiza la evaluación de métricas FAIR. **Resultados**. Los niveles más altos de adherencia están vinculados a la localización y la interoperabilidad. Respecto a la localización, 100% de la muestra alcanzó un nivel avanzado en cuatro de las cinco métricas. En cuanto a la interoperabilidad, 100% de los conjuntos de datos cumplieron con dos de las tres métricas en un nivel avanzado. La accesibilidad y la reutilización fueron los puntos débiles de la muestra. Ninguna de las cinco métricas de reutilización se implementó en un nivel avanzado para el 100% de la muestra. **Conclusiones**. En general, los conjuntos de datos de Aleia tienen niveles moderados de FAIRness. Para lograr mayores niveles de adherencia, es necesario invertir en mejoras en accesibilidad y reutilización. Sin embargo, los conjuntos de datos se destacaron en términos de localización e interoperabilidad, alcanzando niveles avanzados de adherencia**.**

**Palavras clave:** datos de investigación, repositorio de datos de investigación, repositorio Aleia, principios FAIR, ciencia abierta

# 1 Introdução

No contexto atual da e-Science, pautada pelo grande volume de dados, o avanço rápido das tecnologias, o uso intensivo de computação e o trabalho colaborativo entre pesquisadores de diversas áreas, os dados de pesquisa desempenham um papel crucial (Córdula & Araújo, 2019). De forma simplificada, esses dados são "[...] entidades usadas como evidência de fenômenos para fins de pesquisa ou erudição" (Borgman, 2015, p. 29, tradução nossa) e representam um enorme potencial para solucionar problemas e apoiar tomadas de decisões assertivas, deixando de ser meros subprodutos das atividades científicas para tornarem-se uma fonte primária de novos conhecimentos (Sales et al., 2020). Nesse cenário, sua importância está atrelada a diversos benefícios para a ciência e para a sociedade, como a validação, a transparência e a reprodutibilidade dos resultados das pesquisas, o que garante a credibilidade e a integridade científicas. Em um cenário de constante negacionismo e descredibilização da ciência, esse compartilhamento e transparência são essenciais.

O compartilhamento e a disponibilização livre dos dados de pesquisa apoiam uma ciência mais justa e transparente como preconiza o movimento da Ciência Aberta, além de aumentarem os retornos do investimento público; reforçarem a investigação científica aberta; incentivarem a diversidade de estudos e opiniões; promoverem novas áreas de trabalho e permitirem a exploração de tópicos não previstos pelos investigadores iniciais (*Organisation for Economic Co-operation and Development* [OECD], 2007). Curty (2019) corrobora ao afirmar que o compartilhamento dos dados de pesquisa está ligado a oportunidades de expansão do potencial de aplicabilidade de dados científicos pré-existentes, de modo a otimizar custos e esforços associados à coleta de dados.

Apesar da importância dos dados de pesquisa no cenário atual, estudos sugerem que até 80% deles são perdidos dentro de 20 anos (Juty et al., 2020), tornando impossível acessá-los e reutilizá-los. Muitos desses dados coletados por pesquisadores são únicos e não podem ser substituídos em casos de perda ou má gestão como, por exemplo, os dados de eventos climáticos. É preciso, portanto, adotar medidas que garantam sua preservação a longo prazo, bem como sua contextualização, uma vez que dados de pesquisa possuem um nível de abstração maior que publicações tradicionais como artigos, anais de eventos e livros. Essa abstração pode dificultar a interpretação, a validação e a reutilização adequada dos dados por outros pesquisadores, e por isso a importância da sua representação adequada por meio de metadados. Mas para isso é preciso a conscientização dos pesquisadores quanto à importância do compartilhamento e do uso de ferramentas adequadas, bem como a criação de políticas que norteiam e incentivam essa prática, junto com infraestruturas tecnológicas que apoiem a disponibilização dos dados de pesquisa (Araújo et al., 2021).

Uma solução amplamente adotada quando se fala no compartilhamento e na preservação dos dados de pesquisa são os repositórios de dados, que fornecem local seguro e confiável para o depósito, a representação, o acesso e a preservação dos dados. A criação e disseminação de repositórios de dados, segundo Curty (2019), passaram por um rápido impulso a partir das diretivas governamentais e institucionais em torno do compartilhamento de dados de pesquisa financiados com recursos públicos. Um exemplo é o caso da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), que já prevê a gestão e o compartilhamento dos dados de pesquisa que resultem de projetos financiados pela Fundação.

A criação e o fortalecimento desses ambientes tecnológicos que permitem o compartilhamento dos dados são essenciais, uma vez que a gestão de dados de pesquisa em repositórios é um dos principais recursos para o avanço da ciência (Tartarotti et al., 2019). Ao depositarem seus dados em repositórios, os pesquisadores facilitam não só o acesso e a preservação, mas também a documentação e citação adequada deles, indo de acordo com as políticas das agências de fomento e de movimentos internacionais como a Ciência Aberta e Dados Abertos.

Com isso em mente, o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) desenvolveu um repositório institucional de dados de pesquisa em acesso aberto: o Aleia, que dá suporte à comunidade científica do Instituto para a gestão adequada desses insumos valiosos, além de fortalecer o movimento da Ciência Aberta. O Aleia foi lançado oficialmente durante o Seminário Nacional de Bibliotecas Universitárias (SNBU), em 2023, motivado pelo Compromisso 3 do 4º Plano de Ação Nacional, que visa estabelecer mecanismos de governança de dados de pesquisa para o avanço da Ciência Aberta no Brasil (Aleia, 2024).

Mas, a criação de repositórios e o depósito dos dados não é suficiente quando se fala na gestão adequada, que precisa ir além do simples compartilhamento, estando também fortemente relacionada à adoção de boas práticas ao longo do ciclo de vida da pesquisa (Sales et al., 2020). Isso porque, segundo Simionato (2017, p. 1), os dados coletados pelos pesquisadores “[...] em sua grande parte, nem sempre estão estruturados de forma coerente para serem disponíveis aos usuários, como o que acontece com o fenômeno *Big Data*”, o que dificulta seu acesso e interpretação. Mas, se bem geridos, os dados de pesquisa facilitam e simplificam o processo contínuo de descoberta, avaliação e reutilização em novas pesquisas, agilizando o ciclo da ciência.

Diferentes atores beneficiam-se da adoção dessas boas práticas, como pesquisadores que querem compartilhar e obter crédito pelos dados que coletaram e aqueles que querem reutilizar os dados coletados pelos seus pares, além de uma comunidade de ciência de dados que minera, integra e analisa dados novos e existentes para avançar na inovação científica. Guandalini et al. (2019) acrescentam que as instituições de ensino e pesquisa e as agências financiadoras se atentam cada vez mais com as boas práticas ligadas à gestão dos dados de pesquisa, acompanhando o cenário internacional.

Neste sentido, uma iniciativa mundialmente conhecida e difundida são os Princípios FAIR. FAIR é um acrônimo para "*Findable*" (Localizável), "*Accessible*" (Acessível), "*Interoperable*" (Interoperável) e "*Reusable*" (Reutilizável), e esses princípios desempenham um papel central como elementos-chave para boas práticas em todos os processos de gestão de dados (Sales et al., 2020). Publicados originalmente em 2016 na revista *Scientific Data*, os Princípios FAIR foram desenvolvidos para melhorar a infraestrutura de reutilização de dados, otimizando o processamento automático por máquina (Wilkinson et al., 2016).

Nicholson et al. (2023) acrescentam que esses princípios representam diretrizes para a gestão de dados de pesquisa que, se amplamente adotados, irão transformar o cenário do compartilhamento de dados ao melhorar a capacidade de localização, acessibilidade, interoperabilidade e reutilização deles, tanto para máquinas como para humanos. Este enfoque nas máquinas é essencial porque, de acordo com a iniciativa GO FAIR (n. d., tradução nossa), “[...] os humanos dependem cada vez mais do suporte computacional para lidar com os dados, devido ao aumento do volume, da complexidade e da velocidade de criação deles”. Ou seja, é preciso investir em soluções que facilitem o processamento automático. Hodson et al. (2018) corroboram ao afirmarem que o que ficou conhecido como Princípios FAIR são um meio de melhorar a qualidade da pesquisa e aumentar os benefícios extraídos dos dados.

Entretanto, "No Brasil, mais precisamente na Ciência da Informação, ainda são incipientes os relatos na literatura científica de trabalhos que exploram a adoção dos Princípios FAIR a conjuntos de dados (ou *datasets*) de pesquisa depositados em repositórios de dados de maneira prática" (Groehs et al., 2023, p. 5). A incipiência do tema na literatura somada ao nível elevado de abstração dos Princípios FAIR na publicação original leva a dificuldades de implementação e níveis ainda baixos de *FAIRness* no país, como visto em estudo de Bonetti (2023) em repositórios de dados institucionais de São Paulo.

Nesse contexto, o objetivo deste trabalho é avaliar a aderência dos conjuntos de dados de pesquisa depositados no repositório Aleia quanto aos Princípios FAIR, contribuindo para as discussões do tema em nível nacional. Ao final das análises será possível responder à seguinte pergunta: quais Princípios FAIR são os pontos fortes e quais são os pontos fracos dos *datasets* do Aleia? A partir disso serão propostas melhorias práticas para que os dados sejam cada vez mais localizáveis, acessíveis, interoperáveis e reutilizáveis, seguindo as tendências internacionais e potencializando os benefícios atrelados ao repositório. Essas melhorias podem ser adotadas, inclusive, por outros repositórios de dados de pesquisa no Brasil.

De modo a garantir o embasamento teórico para a pesquisa, as próximas seções apresentam uma breve revisão teórica sobre dados de pesquisa, repositórios de dados e Princípios FAIR. Em seguida, são detalhados os métodos utilizados no estudo e é feita uma discussão sobre os resultados encontrados. O artigo finaliza com as conclusões, sintetizando as descobertas e sugestões de implementações.

## 2 Revisão de literatura

Dados de pesquisa podem ser definidos, de acordo com a OECD (2007, p. 13, tradução nossa), como "[...] registros factuais usados como fontes primárias na pesquisa científica, e que são geralmente aceitos na comunidade científica como necessários para validar os resultados de pesquisa". Em consonância, Sayão e Sales (2020, p. 32) definem os dados como "[...] todo e qualquer tipo de registro coletado, observado, gerado ou utilizado no âmbito da pesquisa científica, que pode ser interpretado, tratado e aceito como evidência pela comunidade científica e necessário para analisar, validar e produzir resultados de pesquisa". Ou seja, eles são registros coletados e/ou utilizados durante a pesquisa e servem como fonte primária para validar os resultados publicados.

Os dados de pesquisa possuem um alto nível de abstração e de heterogeneidade, estão presentes em todas as áreas do conhecimento, e incluem:

> [...] fatos e estatísticas recolhidas para posterior referência ou análise, documentos (texto, Word), planilhas (Excel, etc), cadernos de laboratório, cadernos de campo, diários, questionários, transcrições, fitas de áudio, fitas de vídeo, fotografias, filmes, sequências de proteínas ou genéticos, respostas de teste, slides, artefatos, amostras, coleção de objetos digitais adquiridos e gerados durante o processo de pesquisa, conteúdos de banco de dados (vídeo, áudio, texto, imagens), modelos, algoritmos, scripts, arquivos de log, software de simulação, metodologias e fluxos de trabalho, procedimentos operacionais, padrões e protocolos (Dudziak, 2016, para. 1).

O reconhecimento dessa idiossincrasia é essencial quando se fala nas escolhas tecnológicas e gerenciais em repositórios, que, muitas vezes, contemplam dados de diversos domínios (multidisciplinares) e precisam lidar com diferentes tipologias documentais. Sales et al. (2020) explicam que a gestão eficaz dos dados de pesquisa acontece através de alguns componentes essenciais, dentre eles: estruturação adequada dos dados, metadados de qualidade, potencial de interoperabilidade entre eles e atendimento a questões legais e éticas.

A adoção de metadados de qualidade e a necessidade de interoperabilidade estão ligadas às infraestruturas tecnológicas como os repositórios de dados de pesquisa, que são ambientes digitais adequados para a gestão dos dados e têm como objetivo custodiar, publicar, disseminar, preservar e compartilhar conjuntos de dados. Segundo Sayão e Sales (2016), os repositórios de dados apoiam todo o ciclo da gestão de dados de pesquisa e garantem o acesso contínuo e aberto desses objetos digitais. Sanchez et al. (2017, p. 3) trazem uma definição sucinta dos repositórios de dados como aqueles que "[...] buscam organizar, estruturar, permitir acesso, disseminar e preservar todos os dados gerados por meio de pesquisas realizadas em sua maioria por Instituições de Ensino e Pesquisa". É o caso do Ibict com o repositório Aleia, por exemplo.

Segundo Sayão e Sales (2016), os repositórios de dados trazem diversos benefícios para a ciência e para a sociedade como: ampliam a visibilidade dos dados de pesquisa depositados; facilitam a identificação dos autores dos dados, o que permite que eles sejam citados corretamente; oferecem um ambiente tecnológico seguro para a preservação a longo prazo dos dados; contribuem para a formação da memória científica das instituições; permitem que os dados estejam disponíveis *on-line* para serem acessados, baixados e visualizados e aumentam o grau de reutilização e reinterpretação dos dados.

Nesse contexto, o Aleia[1] é o repositório de dados do Ibict e configura-se como um repositório institucional, uma vez que cumpre o papel de memória institucional e é gerenciado "[...] no âmbito de uma instituição acadêmica,

[1] Disponível em: https://aleia.ibict.br/.

como universidades ou institutos de pesquisa, e são voltados para arquivar dados que são, geralmente, provenientes unicamente das atividades acadêmicas dessas instituições" (Sayão & Sales, 2016, p. 101).

Lançado oficialmente em 2023, o Aleia tem como objetivo promover, apoiar e facilitar a adoção do acesso aberto aos dados de pesquisa da comunidade científica do Instituto, constituída por seus servidores, pesquisadores visitantes, bolsistas e alunos do programa de pós-graduação (Aleia, 2024). O repositório está organizado em sete comunidades construídas a partir dos assuntos gerais tratados pelo Ibict, sendo elas: 1) Tecnologias da Informação, 2) Informação Científica, 3) Informação para Sustentabilidade, 4) Informação Tecnológica, 5) Ensino e Pesquisa, 6) Popularização da Ciência, 7) Informação para a Sociedade.

O Aleia é interoperável e possibilita que seus metadados sejam coletados por "[...] provedores de serviços (*harvesters*) e outros sistemas agregadores internacionais e nacionais como o Portal Brasileiro de Publicações e Dados Científicos em Acesso Aberto (Oasisbr), Repositórios Científicos de Acesso Aberto de Portugal (RCAAP), *Red de Repositorios de Acceso Abierto a la Ciencia* (LA Referencia), OpenAIRE" (Aleia, 2024).

Quanto à preservação, o Aleia observa as "[...] recomendações do modelo de referência de preservação digital para arquivos digitais *Open Archival Information System* (OAIS) e critérios para certificação de repositórios confiáveis" (Aleia, 2024), estando, inclusive, em processo de análise para obtenção do selo internacional *CoreTrustSeal*, que reflete as principais características indispensáveis para repositórios de dados confiáveis.

O Aleia também atua de acordo com a legislação brasileira e normativas institucionais, nacionais e internacionais como a Lei nº 13.709/2018, que dispõe sobre a Lei Geral de Proteção de Dados Pessoais (LGPD); a Resolução 466/12 do Conselho Nacional de Saúde, que trata de pesquisas e testes em seres humanos; o Decreto nº 8.777/2016, que institui a Política de Dados Abertos do Poder Executivo Federal; e a Lei nº 12.527/2011, que regula a Lei de Acesso à Informação (LAI).

Quanto à infraestrutura tecnológica, o Aleia foi desenvolvido com o *software* Dataverse, um aplicativo *web* de código aberto desenvolvido para compartilhar, preservar, citar, explorar e analisar dados de pesquisa (Rocha et al., 2021). O Dataverse foi desenvolvido em 2006 pelo *Institute for Quantitative Social Science* da Universidade de Harvard em colaboração com profissionais de todo o mundo, por meio do *Dataverse Project* (Dataverse Project, n. d.). A essência do *Dataverse Project* é reduzir a carga de trabalho de pesquisadores e editores de dados, através da automatização, com foco nos dados de pesquisa. Autores como Silva e Rodrigues (2021, p. 127) defendem que "[...] uma plataforma de repositório como o Dataverse pode facilitar muito a criação de dados científicos FAIR [...]", o que será explorado ao longo do artigo.

Os Princípios FAIR foram estabelecidos como resultado da conferência internacional *Jointly designing the data FAIRPORT*, realizada em 2014 na Holanda. A conferência reuniu especialistas de diversos países e de diferentes áreas para discutir o uso, tratamento e reutilização de dados de pesquisa no âmbito da *e-Science*, buscando criar uma infraestrutura global para dados. A partir das deliberações ocorridas na conferência, surgiu a noção de que, por meio da definição e amplo apoio a um conjunto mínimo de princípios e práticas orientadores acordados pela comunidade, todas as partes interessadas poderiam descobrir, acessar, integrar, reutilizar e citar adequadamente as vastas quantidades de informações e dados gerados pela ciência contemporânea. A reunião concluiu com um rascunho da formulação de um conjunto de princípios fundamentais que foram posteriormente elaborados em maiores detalhes e deram origem ao FAIR (Wilkinson et al., 2016). Com a sua publicação na revista *Scientific Data* em março de 2016, houve uma disseminação mais ampla.

A questão norteadora que diferencia os Princípios FAIR de outras iniciativas já existentes é a proposta de dados "acionáveis por máquinas" (*machine actionable*). Essa proposta se aplica em dois contextos complementares: primeiro, ao se referir aos metadados contextuais que cercam um objeto digital ('o que é isso?'); e segundo, ao se referir ao conteúdo do próprio objeto digital ('como eu o processo/integro?').

Wilkinson et al. (2016) contextualizam melhor essa ideia ao explicarem que humanos e máquinas frequentemente enfrentam barreiras distintas ao tentar encontrar e processar dados na *web*. Humanos têm um senso intuitivo de "semântica", sendo capazes de identificar e interpretar uma ampla variedade de pistas contextuais. Isso diminui a probabilidade de erros serem cometidos na seleção de dados apropriados ou outros objetos digitais, embora humanos enfrentem dificuldades semelhantes se houver a falta de metadados contextuais suficientes. A principal limitação dos humanos, no entanto, é que são incapazes de operar no escopo, escala e velocidade necessários no contexto da *e-Science*. É por essa razão que humanos dependem cada vez mais de agentes computacionais para realizar tarefas de descoberta e integração.

Para contornar esse problema, é fundamental que as máquinas sejam capazes de agir de forma autônoma e apropriada quando confrontadas com a ampla gama de tipos, formatos e mecanismos/protocolos de acesso que serão encontrados durante sua exploração nos ambientes digitais. Também requer que as máquinas mantenham um registro requintado de procedência, de modo que os dados que estão coletando possam ser citados de forma precisa e adequada. Auxiliar esses agentes é uma decisão crítica para todos os participantes no processo de gestão e administração de dados, e é nesse cenário que entra o FAIR.

O FAIR diz respeito a princípios de alto nível independentes de domínio que podem ser aplicados a uma ampla gama de objetos digitais (não só dados de pesquisa), tendo como foco o aprimoramento da capacidade das máquinas de localizar e usar os dados automaticamente (Wilkinson et al., 2016). É importante destacar que embora os elementos dos Princípios FAIR estejam relacionados, eles são independentes e separáveis, o que significa que podem ser implementados em qualquer combinação ou ordem (Silva et al., 2018). Levando em conta que os princípios surgiram a partir de uma conferência internacional que reuniu especialistas de diversas áreas, era esperado que o FAIR fosse flexível nesse sentido. Apesar desse benefício, a flexibilidade também leva a um grau maior de abstração, o que pode dificultar sua implementação prática e a padronização nas soluções adotadas. No Quadro 1 abaixo é possível ver, em uma tradução livre, os princípios como foram preconizados na publicação original (Wilkinson et al., 2016).

**Quadro 1**

*Princípios FAIR*

| | **Princípios FAIR** |
|---|---|
| **Localizável** | L1 - um identificador globalmente exclusivo e persistente deve ser atribuído aos (meta)dados |
| | L2 - os dados são descritos com metadados enriquecidos |
| | L3 - os metadados incluem de forma clara e explícita o identificador dos dados que descrevem |
| | L4 - os (meta)dados são registrados ou indexados em um recurso que permite pesquisas |
| **Acessível** | A1 - (meta)dados são recuperáveis por seu identificador usando um protocolo de comunicação padronizado |
| | A1.1 - o protocolo é aberto, gratuito e universalmente implementável |
| | A1.2 - o protocolo permite um procedimento de autenticação e autorização, quando necessário |
| | A2 - metadados são acessíveis, mesmo quando os dados não estão mais disponíveis |
| **Interoperável** | I1 - (meta)dados usam uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento |
| | I2 - (meta)dados usam vocabulários que seguem os Princípios FAIR |
| | I3 - (meta)dados fazem referências qualificadas a outros (meta)dados |
| **Reutilizável** | R1 - (meta)dados são descritos de forma rica com uma pluralidade de atributos precisos e relevantes |
| | R1.1 - (meta)dados são publicados com uma licença de uso de dados clara e acessível |
| | R1.2 - (meta)dados estão associados a uma proveniência detalhada |
| | R1.3 - (meta)dados atendem aos padrões da comunidade relevantes para o domínio |

***Nota.*** *Fonte: traduzido de Wilkinson et al. (2016).*


No. 88 (2025) e006 • http://biblios.pitt.edu/ • DOI 10.5195/biblios.2025.1255

Apesar de serem diretivas que buscam otimizar a localização, o acesso, a interoperabilidade e a reutilização de (meta)dados, percebe-se, pelo Quadro 1, que os Princípios FAIR não são e não exigem padrões ou ferramentas específicas. Na verdade, o FAIR busca apontar para uma direção de maior utilidade e melhores serviços de dados. Como já mencionado, isso traz certa flexibilidade para que cada comunidade científica e instituição implemente o FAIR de acordo com as necessidades específicas do seu público-alvo, mas também cria desafios quanto à implementação e à padronização das soluções adotadas.

Na prática, "[...] os Princípios FAIR estão sujeitos a diferentes interpretações, levando a diferentes aplicações e implementações e, consequentemente, reduzindo sua eficácia" (Henning et al., 2021, p. 724, tradução nossa). Somado a isso, é preciso levar em conta que "[...] o que é considerado FAIR em uma comunidade pode não ser para outra comunidade" (Henning et al., 2019, p. 181). Ou seja, diversos fatores podem influenciar nos níveis de aderência ao FAIR dos dados de pesquisa.

Sem contar que o FAIR dá foco ao processamento automático por máquinas, que devem ser capazes de "entender" e operar corretamente os dados de pesquisa. Isso significa que os princípios precisam ser implementados pensando não só no usuário humano, mas também nos agentes computacionais que irão acessar, interpretar e até mesmo reutilizar esses dados sem a intervenção direta dos indivíduos que deram o comando. Isso pode representar um desafio significativo para a conformidade com os Princípios FAIR, mas é preciso se ter em mente que sua implementação nas comunidades científicas já "[...] é um fato e tende a se tornar mais comum, na proporção em que as instituições de ensino e pesquisa demandem de seus pesquisadores o depósito dos dados de pesquisa obtidos em suas atividades de investigação em repositórios de dados específicos" (Dias et al., 2023, p. 22). Logo, investir na implementação dos princípios é essencial para otimizar o benefício extraído dos dados.

Apesar disso, como apontam Torino et al. (2023, p. 2), "no Brasil, ainda são incipientes as discussões acerca da aplicação e da maturidade dos Princípios FAIR, seja em dados, objetos digitais ou ambientes informacionais", e por isso a importância de trazê-los como objeto de pesquisa no cenário brasileiro. Com isso em mente, busca-se avaliar a aderência dos conjuntos de dados de pesquisa depositados no repositório Aleia quanto aos Princípios FAIR, identificando em que níveis de *FAIRness* se encontram os *datasets*. Este trabalho auxiliará no mapeamento da situação e no levantamento do estado da arte do tema no Brasil, junto com outros estudos como o de Bonetti (2023) e Groehs et al. (2023).

## 3 Procedimentos metodológicos

A pesquisa caracterizou-se como descritiva, que "[...] expõe as características de uma determinada população ou fenômeno, demandando técnicas padronizadas de coleta de dados" (Prodanov & Freitas, 2013, p. 127). Por meio de uma abordagem quali-quantitativa, buscou avaliar a aderência dos conjuntos de dados de pesquisa depositados no repositório Aleia quanto aos Princípios FAIR. A pesquisa se deu em 6 etapas: 1) levantamento bibliográfico para a construção do referencial teórico; 2) levantamento de todos os conjuntos de dados depositados no Aleia; 3) registro das informações de identificação (URL e código no repositório) dos conjuntos de dados em uma planilha Google; 4) avaliação dos conjuntos de dados com o uso de uma ferramenta padronizada; 5) registro das pontuações obtidas para cada conjunto de dados na planilha, que será posteriormente compartilhada em um repositório de dados; 6) análise e discussão dos resultados a partir dos dados coletados.

Para a fundamentação teórica, esta pesquisa pautou-se em buscas nas seguintes fontes: 1) Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação (BRAPCI); 2) Biblioteca Eletrônica Científica Online (SciELO); 3) periódicos da área de Ciência da Informação no Portal de Periódicos da CAPES; e 4) *Google Scholar*. Para as buscas utilizou-se os seguintes termos, em português e inglês: "dados de pesquisa/*research data*", "gestão de dados de pesquisa/*research data management*", "repositório de dados de pesquisa/*research data repository*", "Princípios FAIR/*FAIR principles*", "boas práticas/*best practices*". Realizou-se uma busca adicional usando "dados científicos", uma vez que o termo é tido como sinônimo de "dados de pesquisa".

O levantamento dos conjuntos de dados do Aleia para avaliação resultou em uma amostra de 16 *datasets*. Eles foram individualmente avaliados entre os dias 07/02/2024 e 09/05/2024, com o auxílio da ferramenta *F-UJI Automated FAIR Data Assessment Tool*[2], desenvolvida pela *Fostering Fair Data Practices in Europe - FAIRsFAIR* com financiamento da *European Union's Horizon* 2020. A ferramenta F-UJI, por padrão, apresenta seu conteúdo

[2] Disponível em: https://www.f-uji.net/.

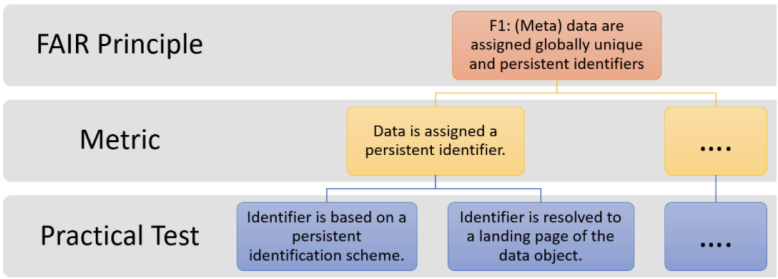
em inglês, inclusive os resultados das avaliações, mas é possível utilizar o recurso de tradução automática do navegador para traduzir para o português. Isso permite que a ferramenta seja utilizada com maior facilidade no contexto brasileiro. Neste manuscrito, optou-se por apresentar os relatórios da ferramenta F-UJI traduzidos desta forma. A lista dos 16 *datasets* analisados pode ser consultada no Apêndice A.

A F-UJI, baseada em *Representational State Transfer* (REST), faz a avaliação automática, por meio dos metadados, dos níveis de conformidade dos conjuntos de dados de pesquisa ao FAIR, entregando relatórios individuais para cada conjunto de dados avaliado. A ferramenta possui métricas baseadas nos indicadores propostos pelo *RDA FAIR Data Maturity Model Working Group*, no *WDS/RDA Assessment of Data Fitness for use checklist*, bem como nos projetos *FAIRdat*[3] e *FAIREnough*[4]. As métricas, os métodos, as escalas e o código-fonte podem ser consultados em detalhes no site da F-UJI. A ferramenta, que está em desenvolvimento e recebe constantes atualizações, vem sendo utilizada em outros estudos nacionais e internacionais (Bonetti, 2023; Groehs et al., 2023; Petrosyan et al., 2023). Seu uso foi importante para a avaliação porque, como já apontado por outros autores, os Princípios FAIR são difíceis de se colocar em prática devido ao nível de subjetividade, abstração e complexidade (Henning et al., 2019). As métricas são uma forma de medir os níveis de conformidade dos recursos digitais, como os dados de pesquisa, aos Princípios FAIR, possibilitando a implementação de melhorias e, consequentemente, aumento do nível de *FAIRness*.

Os Princípios FAIR independem de domínio e podem ser aplicados a uma ampla gama de resultados acadêmicos, definidos de forma genérica em sua publicação original (Wilkinson et al. 2016). Com isso surge espaço para interpretações ambíguas e falta de padronização na sua implementação. A F-UJI, então, especificou uma ou mais métricas para cada princípio FAIR. Como cada métrica pode ser testada de várias formas, dependendo dos contextos dos dados e das boas práticas atuais, foi proposto pela F-UJI um ou mais testes práticos para avaliar conjuntos de dados em relação a uma métrica específica. Isso resulta em um modelo hierárquico em que cada princípio FAIR recebe várias métricas, que por sua vez podem receber vários testes, conforme a Figura 1.

**Figura 1**

*Relação hierárquica entre princípios, métricas e testes práticos usados pela ferramenta F-UJI na avaliação*




***Nota.*** Fonte: captura de tela do site da ferramenta F-UJI (2024). [Descrição da imagem] Figura retangular horizontal com fundo branco. Três retângulos cinzas estão empilhados criando uma estrutura hierárquica. No primeiro retângulo está escrito “FAIR Principle” e, ao lado, em um retângulo laranja, está escrito “F1: (Meta) data are assigned globally unique and persistent identifiers”. No segundo retângulo, posicionado logo abaixo do primeiro, está escrito “Metric” e, ao lado, em um retângulo amarelo, está escrito “Data is assigned a persistent identifier”. No terceiro e último retângulo, está escrito “Practical Test” e, ao lado, em dois retângulos azuis, está escrito “Identifier is based on a persistent identification scheme” e “Identifier is resolved to a landing page of data object”. [Fim da descrição].

Todos os 16 conjuntos de dados depositados no Aleia foram avaliados pela ferramenta F-UJI, que atribui, na parte de resumo (Figura 2), uma porcentagem geral do nível de *FAIRness* de cada conjunto de dados, junto com uma pontuação para os princípios localizável (ou encontrável, na tradução livre), acessível, interoperável e reutilizável (*Score earned)*. A pontuação determina o nível de aderência, que varia numa escala de incompleto, inicial, moderado ou avançado (*Fair level*) e aparece logo ao lado da pontuação. Todas as escalas e pontuações foram

[3] Disponível em: https://www.surveymonkey.com/r/fairdat.
[4] Disponível em: https://docs.google.com/forms/d/e/1FAIpQLSf7t1Z9IOBoj5GgWqik8KnhtH3B819Ch6lD5KuAz7yn0I0Opw/viewform.

definidas e atribuídas pela ferramenta, assim como as notas máximas para cada princípio FAIR (*Score earned*), como mostra a Figura 2:

**Figura 2**

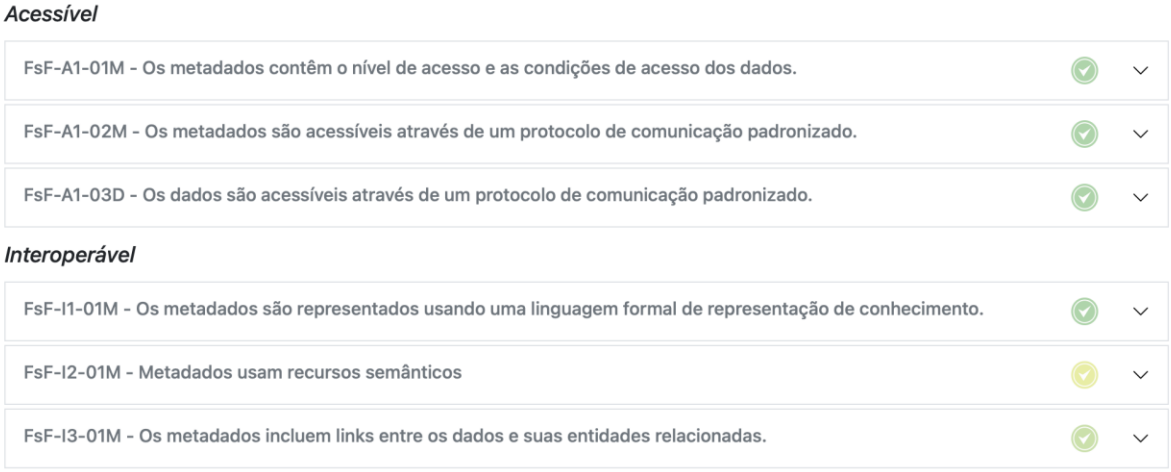
*Resumo (summary) da aderência aos Princípios FAIR fornecida pela ferramenta F-UJI para um conjunto de dados do Aleia*

***Nota.*** Fonte: captura de tela do site da ferramenta F-UJI (2024). [Descrição da imagem] Figura retangular horizontal com fundo branco. No canto esquerdo tem-se um círculo dividido em quatro partes, cada uma com uma cor. ¼ está pintado de rosa e representa o princípio reutilizável, ¼ está pintado de azul e representa o princípio localizável, ¼ está pintado de verde e representa o princípio acessível e o último ¼ está pintado de amarelo e representa o princípio interoperável. Já no canto direito, tem-se uma pequena tabela com três colunas e cinco linhas. Na primeira coluna tem-se os quatro princípios elencados na seguinte ordem: encontrável (ou localizável), acessível, interoperável e reutilizável. Na segunda coluna tem-se o título "Pontuação obtida". Na linha do princípio encontrável, vê-se a pontuação 7 de 7. Na linha do acessível, vê-se a pontuação 3 de 3. Na linha do interoperável, vê-se a pontuação 3 de 4. Na linha do reutilizável, vê-se a pontuação 7 de 10. Por fim, na última coluna, tem-se o título "Nível justo (que significa nível de FAIR)". Na linha do princípio encontrável, vê-se o nível avançado escrito em fundo verde vibrante. Na linha do acessível, vê-se o nível avançado escrito em fundo verde vibrante. Na linha do interoperável, vê-se o nível moderado escrito em fundo verde fosco. Na linha do reutilizável, vê-se o nível moderado escrito em fundo verde fosco. [Fim da descrição].

Já o relatório (Figura 3) entregue pela F-UJI exibe em detalhes quais métricas foram consideradas. Por exemplo: em localizável, a ferramenta testa a métrica FsF-F1-02D: foi atribuído um identificador persistente. Isso possibilita entender o que foi avaliado, analisar os pontos fracos e fortes dos conjuntos de dados quanto aos Princípios FAIR e identificar quais funcionalidades precisam ser aperfeiçoadas.

**Figura 3**

*Relatório (report) da aderência aos Princípios FAIR fornecido pela ferramenta F-UJI para um conjunto de dados do Aleia*




***Nota.*** Fonte: captura de tela do site da ferramenta F-UJI (2024). [Descrição da imagem] Figura retangular horizontal com fundo branco. A imagem exibe um recorte do relatório de feedback que é apresentado pela ferramenta F-UJI. Para o princípio acessível, o primeiro a ser exibido na figura, vê-se três métricas: FsF-A1-01M:

Os metadados contêm o nível de acesso aos dados, FsF-A1-02M: Os metadados são acessíveis através de um protocolo de comunicação padronizado e FsF-A1-03D: Os dados são acessíveis através de um protocolo de comunicação padronizado. Logo abaixo tem-se o princípio interoperável com três métricas listadas, na seguinte ordem: FsF-I1-01M: Os metadados são representados usando uma linguagem formal de representação de conhecimento, FsF-I2-01M: Os metadados usam recursos semânticos e FsF-I3-01M: Os metadados incluem links entre os dados e suas entidades relacionadas. Ao lado de cada métrica, há um ícone do símbolo check em tons variados de verde. [Fim da descrição].

De acordo com Hodson et al. (2018), métricas são necessárias para avaliar o *FAIRness* dos dados e certificar serviços. Elas aumentam a reutilização e a confiança, além de embasar as avaliações e decisões de financiamento. Logo, é fundamental que a ferramenta F-UJI se baseie em métricas claras para fornecer as pontuações dos *datasets*, sendo uma referência para futuras melhorias. Os autores ainda acrescentam que essas métricas precisam ser revisadas e atualizadas constantemente (Hodson et al., 2018), como é o caso da F-UJI.

No relatório exibido na Figura 3, ao clicar em uma métrica, também é possível verificar o nível de *FAIRness* (incompleto, inicial, moderado, avançado) em que o *dataset* se encontra e quais testes práticos foram realizados para chegar àquele resultado. Isso garante a transparência sobre o que está sendo avaliado e permite entender o que precisa ser implementado para aumentar os níveis de conformidade. Essa combinação de funcionalidades embasou a escolha da ferramenta F-UJI como instrumento de avaliação dos *datasets* da amostra.

É importante salientar que os resultados analisados neste trabalho se limitam aos dados entregues automaticamente pela ferramenta, que apesar do nível de detalhamento, ainda está em desenvolvimento, aplicando constantes melhorias ao código (Devaraju & Huber, 2020). Além disso, ao longo do tempo, os metadados e o próprio repositório avaliado neste estudo podem passar por aperfeiçoamentos, visto que o Aleia foi lançado recentemente, em 2023. Com isso em mente, a próxima seção apresenta os resultados do estudo.

## 4 Resultados e discussão

No momento da coleta, o Aleia continha um total de 16 conjuntos de dados depositados. Entretanto, após uma avaliação no repositório, constatou-se que seis deles (códigos dos *datasets*: IFIFO8, 8WSOGT, UNVCQO, G3BP3A, MFHBCK, 7DPOHV) não são dados de pesquisa, e sim outras tipologias como apresentações e relatórios. Somado a isso, um dos conjuntos de dados criado (código do *dataset:* KJOHIT) não possui arquivos depositados, apenas o preenchimento dos metadados. Apesar dessas inconsistências, optou-se por avaliar e apresentar os resultados para todos os 16 conjuntos de dados, uma vez que os Princípios FAIR dão destaque para os metadados, como é possível ver no Quadro 1. Inclusive, a ferramenta F-UJI faz a avaliação automática com base nos metadados disponíveis em cada *dataset*, ou seja, a pontuação final está ligada à representação/descrição do conjunto de dados. O momento em que o arquivo depositado afeta a avaliação é na métrica FsF-R1.3-02D (Os dados estão disponíveis em um formato de arquivo recomendado pela comunidade de pesquisa alvo). Logo, mesmo que o arquivo não se configure como um dado de pesquisa, é possível extrair informações importantes da avaliação feita pela F-UJI no que diz respeito aos metadados preenchidos pelo depositante, e por isso, optou-se por manter a amostra de 16 *datasets*.

Além disso, é importante destacar que a cultura do compartilhamento de dados ainda está em desenvolvimento no Brasil, o que implica em muitos pesquisadores não estarem familiarizados com a definição de dados de pesquisa. Há ainda uma falta de capacitação e sensibilização sobre o tema, o que leva os pesquisadores a depositarem arquivos que não são, de fato, dados de pesquisa. Isso é visto em outros repositórios nacionais, como o Arca Dados, que recebeu recentemente o selo *CoreTrustSeal*, que certifica a confiabilidade do repositório. Também é visto em repositórios internacionais amplamente conhecidos e utilizados, como o Zenodo. Assim, percebe-se que esse cenário se repete em outros repositórios, não sendo um problema exclusivo do Aleia. Tal fato também impactou a decisão de manter a avaliação de todos os 16 conjuntos de dados que estavam depositados no repositório, mesmo que alguns deles não configurem dados de pesquisa.

Com isso em mente, na Tabela 1 abaixo é possível visualizar a pontuação obtida por cada *dataset* quanto à localização (pontuação máxima é 7), acessibilidade (pontuação máxima é 3), interoperabilidade (pontuação máxima é 4) e reutilização (pontuação máxima é 10). Essas pontuações máximas foram definidas pela própria ferramenta F-UJI, bem como as notas que cada *dataset* recebeu.

Na Tabela 1 também é possível visualizar os percentuais de aderência geral de cada conjunto de dados do Aleia ao FAIR, que variaram entre 62% a 79%, o que representa um **nível moderado de *FAIRness***. Logo, existem

melhorias a serem adotadas, que serão detalhadas ao longo desta seção. Em um primeiro momento, serão analisados os dados presentes na Tabela 1, que indica o cenário geral da aderência ao FAIR. Por meio desta Tabela, é possível identificar quais princípios exigem melhorias e quais representam pontos fortes do Aleia. Já no Gráfico 1, que será analisado mais à frente, as métricas testadas para cada princípio serão analisadas, o que permitirá indicar quais melhorias práticas precisam ser implementadas para aperfeiçoar os níveis de conformidade ao FAIR no repositório.

**Tabela 1**

*Síntese da aderência dos datasets do Aleia aos Princípios FAIR*

| URL do dataset | Código do dataset | Pontuação em localizável (pontuação máxima 7) | Pontuação em acessível (pontuação máxima 3) | Pontuação em interoperável (pontuação máxima 4) | Pontuação em reutilizável (pontuação máxima 10) | Percentual de aderência geral |
|---|---|---|---|---|---|---|
| https://doi.org/10.48472/aleia/IFIFO8 | IFIFO8 | 7 | 2 | 3 | 7 | 79% |
| https://doi.org/10.48472/aleia/M038PL | M038PL | 7 | 2 | 3 | 7 | 79% |
| https://doi.org/10.48472/aleia/THFG9S | THFG9S | 7 | 2 | 3 | 7 | 79% |
| https://doi.org/10.48472/aleia/8WSOGT | 8WSOGT | 7 | 2 | 3 | 7 | 79% |
| https://doi.org/10.48472/aleia/XYFBKD | XYFBKD | 7 | 2 | 3 | 7 | 79% |
| https://doi.org/10.48472/aleia/HD1ND5 | HD1ND5 | 7 | 2 | 3 | 7 | 79% |
| https://doi.org/10.48472/aleia/Y1O6PD | Y1O6PD | 7 | 2 | 3 | 6 | 75% |
| https://doi.org/10.48472/aleia/UNVCQO | UNVCQO | 7 | 2 | 3 | 6 | 75% |
| https://doi.org/10.48472/aleia/G3BP3A | G3BP3A | 6,5 | 1 | 3 | 5 | 64% |
| https://doi.org/10.48472/aleia/DPW8OE | DPW80E | 6,5 | 1 | 3 | 5 | 64% |
| https://doi.org/10.48472/aleia/AOCTZO | AOCTZO | 6,5 | 1 | 3 | 5 | 64% |
| https://doi.org/10.48472/aleia/Y7FXK8 | Y7FXK8 | 6,5 | 1 | 3 | 5 | 64% |
| https://doi.org/10.48472/aleia/MFHBCK | MFHBCK | 6,5 | 1 | 3 | 5 | 64% |
| https://doi.org/10.48472/aleia/7DPOHV | 7DPOHV | 6,5 | 1 | 3 | 5 | 64% |
| https://doi.org/10.48472/aleia/21GSV8 | 21GSV8 | 6,5 | 1 | 3 | 5 | 64% |
| https://doi.org/10.48472/aleia/KJOHIT | KJOHIT | 6 | 1 | 3 | 5 | 62% |

***Nota.*** Fonte: elaborado pelos autores (2024).

Vale lembrar que o FAIR não é binário (FAIR/não FAIR), mas sim um espectro ao longo do qual vários níveis de *FAIRness* são possíveis (Higman et al., 2019). Isso significa que mesmo os conjuntos de dados depositados no mesmo repositório podem apresentar diferentes níveis de conformidade.

Conforme exposto na Tabela 1, a maior pontuação para localizável foi 7 de 7, obtida por oito *datasets* do Aleia. Sete *datasets* alcançaram a pontuação 6,5 e apenas um obteve nota 6. Isso mostra que, no geral, os conjuntos de dados avaliados estão em um nível avançado de localização, sendo um ponto forte dos *datasets* da amostra. Já quando se fala na acessibilidade, a maior pontuação foi 2 de 3, para metade dos *datasets* do Aleia. A outra metade obteve nota 1, o que equivale a um **nível inicial** de aderência (Tabela 1). A acessibilidade pode, portanto, ser aperfeiçoada, e as melhorias serão apontadas junto à análise do Gráfico 1.

No que se refere à interoperabilidade, todos os 16 conjuntos de dados do Aleia obtiveram exatamente as mesmas pontuações: 3 de 4, o que equivale a um nível moderado de conformidade, um bom resultado (Tabela 1). Autores como Henning et al. (2019) apontam que a interoperabilidade pode ser entendida como o elemento mais desafiador das práticas FAIR, uma vez que exige uma compreensão mais sofisticada de linguagens e padrões específicos da disciplina. Hodson et al. (2018) dão uma dimensão extra ao assunto ao explicarem que a interoperabilidade não é apenas semântica, mas também técnica e legal. Todas essas facetas precisam ser levadas em conta para atingir níveis altos de interoperabilidade.

Os conjuntos de dados também foram avaliados quanto à reutilização. Enquanto seis *datasets* receberam nota 7 de 10, oito receberam nota 5, o que equivale a um nível inicial de aderência (Tabela 1). Dunning et al. (2017), em estudo internacional, indicaram que a reutilização foi o princípio FAIR mais difícil de alcançar, junto com a interoperabilidade. Apesar dos *datasets* do Aleia terem se destacado quanto à interoperabilidade, é importante investir na reutilização dos dados, levando em conta que "[...] os dados precisam possuir certas características para serem considerados reutilizáveis, incluindo aspectos como relevância, completude, compreensibilidade e a confiabilidade." (Curty, 2019, p. 179).

Após identificar o cenário geral da conformidade dos *datasets* depositados no repositório Aleia aos Princípios FAIR, o Gráfico 1 aponta para as 16 métricas avaliadas pela F-UJI e o percentual de conjuntos de dados que alcançaram nível avançado, moderado, inicial ou incompleto. Como já mencionado, para cada métrica da F-UJI é indicado o nível de conformidade (incompleto, inicial, moderado, avançado) em que o *dataset* se encontra. Todos os 16 conjuntos de dados foram avaliados nas 16 métricas, e o resultado está sintetizado no Gráfico 1. Para melhor entendimento das métricas listadas no Gráfico 1, é possível consultar o Quadro 2 abaixo, que traz uma relação de cada métrica com os Princípios FAIR.

**Quadro 2**

*Relação entre as métricas avaliadas pela F-UJI com os Princípios FAIR*

| Princípios FAIR | Métricas | Requisitos |
|---|---|---|
| **Localizável** | FsF-F1-01D | Foi atribuído um identificador único aos dados |
| | FsF-F1-02D | Foi atribuído um identificador persistente aos dados |
| | FsF-F2-01M | Os metadados incluem elementos descritivos essenciais para oferecer suporte à localização dos dados |
| | FsF-F3-01M | Os metadados incluem o identificador dos dados que descrevem |
| | FsF-F4-01M | Os metadados podem ser recuperados por máquinas |
| **Acessível** | FsF-A1-01M | Os metadados contêm o nível e as condições de acesso aos dados |
| | FsF-A1-03D | Os dados são acessíveis por meio de um protocolo de comunicação padronizado |
| | FsF-A1-02M | Os metadados são acessíveis por meio de um protocolo de comunicação padronizado |
| **Interoperável** | FsF-I1-01M | Os metadados são representados usando uma linguagem formal de representação de conhecimento |
| | FsF-I1-02M | Os metadados usam recursos semânticos |
| | FsF-I3-01M | Os metadados incluem links entre os dados e suas entidades relacionadas |

| | | |
|---|---|---|
| **Reutilizável** | FsF-R1-01MD | Os metadados especificam o conteúdo dos dados |
| | FsF-R1.1-01M | Os metadados incluem as licenças de reuso dos dados |
| | FsF-R1.2-01M | Os metadados incluem informações de proveniência sobre a criação ou geração de dados |
| | FsF-R1.3-01M | Os metadados seguem um padrão recomendado pela comunidade de pesquisa alvo dos dados |
| | FsF-R1.3-02D | Os dados estão disponíveis em um formato de arquivo recomendado pela comunidade de pesquisa alvo |

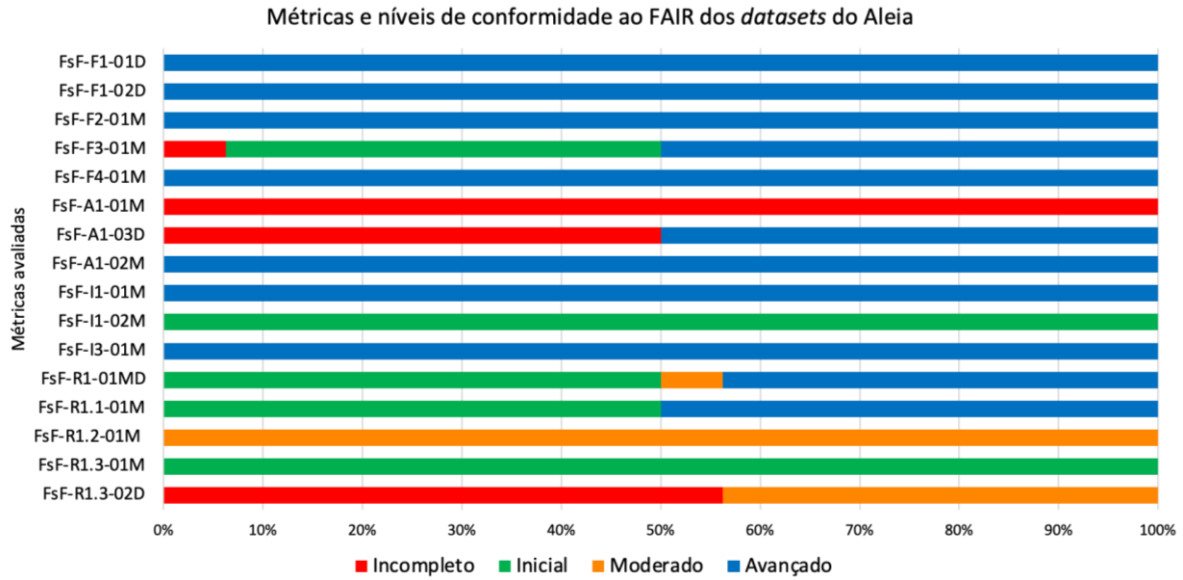
***Nota.*** Fonte: elaborado pelos autores (2024).

No Gráfico 1 abaixo é possível verificar os níveis de conformidade (incompleto, inicial, moderado e avançado) dos conjuntos de dados da amostra a cada uma das métricas citadas anteriormente.

Para se obter a pontuação final quanto à **localização**, os *datasets* da amostra foram testados em cinco métricas: 1) Um identificador globalmente único é atribuído aos dados (FsF-F1-01D); 2) Um identificador persistente é atribuído aos dados (FsF-F1-02D); 3) Os metadados incluem elementos descritivos essenciais (title, creator, data identifier, publisher, publication date, summary e keywords) para oferecer suporte à localização dos dados (FsF-F2-01M); 4) Os metadados incluem o identificador dos dados que descrevem (FsF-F3-01M) e; 5) Os metadados podem ser recuperados por máquinas (FsF-F4-01M).

## Gráfico 1

*Métricas e níveis de conformidade ao FAIR dos datasets do Aleia*




***Nota.*** Fonte: elaborado pelos autores (2024). *[Descrição da imagem]* Gráfico colorido de barras com 16 barras, uma para cada métrica que foi avaliada pela ferramenta F-UJI. As 16 métricas estão listadas no eixo vertical do gráfico na seguinte ordem: FsF-F1-01D, FsF-F1-02D, FsF-F2-01M, FsF-F3-01M, FsF-F4-01M, FsF-A1-01M, FsF-A1-03D, FsF-A1-02M, FsF-I1-01M, FsF-I1-02M, FsF-I3-01M, FsF-R1-01MD, FsF-R1.1-01M, FsF-R1.2-01M, FsF-R1.3-01M e FsF-R1.3-02D. No eixo horizontal tem-se os percentuais, na seguinte ordem: 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% e 100%. Abaixo do eixo horizontal tem-se uma legenda que indica os quatro níveis de aderência ao FAIR e suas respectivas cores no gráfico. O nível incompleto é representado pela cor vermelha, o nível inicial é representado pela cor verde, o nível moderado é representado pela cor laranja e o nível avançado é representado pela cor azul. Há sete barras que estão totalmente azuis, representando um nível avançado de conformidade, referentes às seguintes métricas: FsF-F1-01D, FsF-F1-02D, FsF-F2-01M, FsF-F4-01M, FsF-A1-02M, FsF-I1-01M e FsF-I3-01M. Há uma barra toda vermelha, representando um nível incompleto de conformidade, referente à métrica FsF-A1-01M. Há uma barra totalmente laranja, representando um nível

moderado de conformidade, referente à métrica FsF-R1.2-01M. Há duas barras totalmente verdes, representando um nível inicial de conformidade, referente às métricas FsF-I1-02M e FsF-R1.3-01M. *[Fim da descrição].*

Observa-se que 100% dos conjuntos de dados do Aleia encontram-se em nível avançado quanto à presença de 1) identificadores únicos; 2) identificadores persistentes e 3) metadados descritivos essenciais. A totalidade dos *datasets* analisados também se encontra em nível avançado quanto à recuperação desses metadados por máquinas.

De acordo com Juty et al. (2020), os identificadores persistentes são de suma importância porque a hospedagem dos objetos digitais pode mudar, editoras ou arquivos de dados podem se fundir ou serem adquiridos, ou a infraestrutura interna de uma instituição pode ser renovada, modificando a apresentação dos *Uniform Resource Identifiers* (URI) de um recurso. Isso demonstra que há um problema se os identificadores dependerem de nomes de *host* locais. Os identificadores persistentes contornam esse problema e, de acordo com Juty et al. (2020), existem quatro bons exemplos quando o assunto são objetos digitais FAIR: *Digital Object Identifier* (DOI), *Archival Resource Key* (ARK), Identifiers.org e *Persistent uniform resource locator* (PURL).

O DOI, por exemplo, é amplamente utilizado no cenário editorial para identificar documentos, conjuntos de dados e para a citação e a interoperabilidade de dados, como é o caso de todos os *datasets* do repositório Aleia. A atribuição de um DOI permite que os dados de pesquisa depositados sejam encontrados a partir da associação do seu nome único com um protocolo de recuperação (Juty et al., 2020).

Somado a isso, o uso de metadados descritivos essenciais permite a conferência e a identificação dos *datasets*, além de fornecer um contexto mínimo para sua interpretação pelo usuário. Essa métrica também foi validada para 100% da amostra. A ferramenta conseguiu recuperar todos os metadados essenciais: *creator, title, data identifier, publisher, publication date, summary* e *keywords*. Além desses, também recuperou outros elementos de metadados como: *object type, modified date, license, related resources*.

É importante destacar que metadados ricos permitem que um computador realize automaticamente tarefas rotineiras de classificação e priorização que atualmente exigem muita atenção dos pesquisadores. Além disso, o usuário humano deve ser capaz de encontrar dados com base nas informações fornecidas pelos seus metadados, mesmo sem o identificador dos dados. Não é possível presumir quem irá usá-los, ou para qual finalidade, e por isso a importância de se descrever ricamente os *datasets*. É importante destacar que, no Aleia, o depósito se dá por autoarquivamento, e os autores precisam ser incentivados a preencher corretamente e de forma extensiva os metadados. Isso se dá por meio de políticas institucionais, guias de depósito e seleção de elementos de metadados obrigatórios.

A melhoria a ser implementada diz respeito à métrica FsF-F3-01M: os metadados incluem o identificador dos dados que descrevem. Como mostra o Gráfico 1, apenas metade dos conjuntos de dados conseguiu atingir o nível avançado, enquanto os outros 50% da amostra variou entre nível inicial ou incompleto. De acordo com a F-UJI, para alcançar um nível avançado nessa métrica, é preciso que os metadados contenham um identificador persistente ou *Uniform Resource Locato*r (URL) que indique a localização do conteúdo de dados para *download*. Isso não foi validado em oito *datasets* e pode ser aperfeiçoado para garantir um nível maior de localização.

Para a **acessibilidade**, todos os conjuntos de dados foram testados em três métricas: 1) FsF-A1-01M: Os metadados contêm o nível e as condições de acesso aos dados; 2) FsF-A1-02M: Os metadados são acessíveis por meio de um protocolo de comunicação padronizado; e 3) FsF-A1-03D: Os dados são acessíveis por meio de um protocolo de comunicação padronizado. De acordo com Hodson et al. (2018), dados acessíveis podem ser obtidos tanto por humanos quanto por máquinas mediante autorização apropriada e por meio de um protocolo bem definido e universalmente implementável.

Nota-se, no Gráfico 1, que 100% dos conjuntos de dados conseguiram aderir, em nível avançado, à última métrica de acessibilidade (FsF-A1-03D), o que significa que os dados são acessíveis por meio de protocolos de comunicação padronizados: o *Hyper Text Transfer Protocol Secure* (https). Porém, apenas 50% dos *datasets* do Aleia atingiram nível avançado quanto à segunda métrica (FsF-A1-02M): metadados acessíveis por meio de um protocolo de comunicação padronizado. A outra metade atingiu um nível inicial de aderência, o que aponta para uma possível melhoria, garantindo que tanto os dados quanto os metadados sejam acessíveis por esses protocolos, que trazem as especificações técnicas reguladoras da transmissão de dados e informações entre computadores.

Por fim, quanto à métrica FsF-A1-01M (metadados contêm o nível e as condições de acesso aos dados), a ferramenta realizou alguns testes práticos para avaliar se as informações sobre restrições ou direitos de acesso

podem ser identificadas em metadados e se as informações de acesso aos dados são legíveis por computadores. A F-UJI não conseguiu recuperar e validar essas informações, e 100% dos conjuntos de dados do Aleia foi avaliado como incompleto nesta métrica (Gráfico 1).

Vale destacar que, de acordo com a GO FAIR (n. d.), a acessibilidade é, muitas vezes, mal compreendida. O 'A' em FAIR não significa necessariamente 'aberto' ou 'livre'. Implica, na verdade, na necessidade de fornecer as condições exatas sob as quais os dados podem ser acessados. E, idealmente, de forma que uma máquina possa compreender os requisitos e, em seguida, executá-los automaticamente ou alertar o usuário humano. Inclusive, os Princípios FAIR devem ser aplicados igualmente a dados que permanecerão restritos ou internos a uma determinada organização.

Apesar do Aleia ter o objetivo de promover, apoiar e facilitar a adoção do Acesso Aberto aos dados de pesquisa, promovendo a Ciência Aberta, sabe-se pela política do repositório que é possível aplicar regras específicas e controle de acesso para os conjuntos de dados de pesquisa. Eles podem ser tanto 1) públicos (acesso público tanto aos metadados como aos conjuntos de dados) ou 2) restritos (acesso público apenas aos metadados). Mas é importante que essas informações sejam publicamente declaradas para cada *dataset* e de tal forma que computadores consigam interpretar a informação automaticamente.

Quanto à **interoperabilidade**, os *datasets* do Aleia foram testados em três métricas: 1) Os metadados são representados usando uma linguagem formal de representação de conhecimento (FsF-I1-01M); 2) Os metadados fazem uso de recursos semânticos (FsF-I1-02M); e 3) Os metadados incluem links entre os dados e suas entidades relacionadas (FsF-I3-01M).

Hodson et al. (2018) explicam que para os dados serem interoperáveis, eles precisam ser descritos usando padrões e vocabulários que são reconhecidos pela comunidade de pesquisa, com um significado preciso dos conceitos e qualidades que os dados representam. É isso que permite que os dados sejam acionáveis por máquina, como descrito na publicação original do FAIR. A interoperabilidade é uma característica essencial na usabilidade dos dados e busca otimizar a comunicação entre sistemas e a integração de conjuntos de dados diferentes. Segundo Vidotti et al. (2021, p. 212), para alcançá-la "[...] os dados e os metadados precisam ser legíveis e adequados a padrões e vocabulários reconhecidos, potencializando a ligação com outros padrões e incluir referências qualificadas".

Percebe-se pelo Gráfico 1 que 100% dos conjuntos de dados do Aleia analisados atingiram nível avançado quanto à primeira e à terceira métrica de interoperabilidade (FsF-I1-01M e FsF-I3-01M). No que diz respeito à métrica FsF-I1-01M (metadados são representados usando uma linguagem formal de representação de conhecimento), a ferramenta F-UJI recuperou metadados estruturados (JSON-LD, RDFa) incorporados no código XHTML/HTML da página de destino. O *Resource Description Framework* (RDF) é um padrão do *World Wide Web Consortium* (W3C) para a descrição conceitual e representação das informações sobre os dados no ambiente *web*, fornecendo uma estrutura comum destinada a representar metadados (Silva et al., 2019). Já o *JavaScript Object Notation* (JSON) é um formato de dados usado para transmitir e armazenar dados de forma organizada e legível por máquinas. A validação desses aspectos é fundamental para a interoperabilidade dos dados de pesquisa depositados.

Em relação à métrica FsF-I1-02M (metadados fazem uso de recursos semânticos), todos os *datasets* estão em um estágio inicial de aderência. Atualmente, os metadados sobre assuntos e áreas do conhecimento são predominantemente sintáticos (baseados em termos). Há, portanto, a necessidade de transformá-los em conceituais por meio da utilização da ontologia *Simple Knowledge Organization System* (SKOS). No entanto, em língua portuguesa, ainda há escassez de tesauros SKOS com esses vocabulários controlados. Mesmo que tais tesauros sejam desenvolvidos, eles precisarão ser homologados pela ferramenta F-UJI como confiáveis. Segundo a GO FAIR (n. d.), o vocabulário controlado utilizado para descrever *datasets* deve ser documentado e resolvível por identificadores globalmente únicos e persistentes. Essa documentação deve ser facilmente acessível e disponível para qualquer pessoa que utilize os dados. Portanto, além de implementar vocabulários semânticos, será necessário homologá-los na ferramenta para que possam ser considerados na métrica FsF-I1-02M (metadados fazem uso de recursos semânticos).

Por fim, ainda em relação à interoperabilidade, os testes utilizando a ferramenta F-UJI indicam que 100% dos *datasets* apresentam *links* entre os dados e suas entidades relacionadas (FsF-I3-01M) em nível avançado. Além dos recursos relacionados serem mencionados explicitamente nos metadados, eles são indicados por *links* ou identificadores legíveis por máquina. Vale citar que esses recursos podem ser artigos, teses, dissertações, outros conjuntos de dados, dentre outros recursos. O objetivo, de acordo com a GO FAIR (n. d.), é criar tantas ligações significativas quanto possível entre recursos de (meta)dados para enriquecer o conhecimento contextual sobre

eles. Essa diretriz permite criar uma rede de descoberta de recursos relacionados, possibilitando que os usuários naveguem por diferentes elementos que compõem a pesquisa.

Considerando que autores apontaram a interoperabilidade como um dos elementos mais desafiadores do FAIR (Bonetti, 2023; Dunning et al., 2017; Henning et al., 2019), os resultados do Aleia se mostram promissores, com 100% dos *datasets* atingindo nível avançado em duas das três métricas da F-UJI.

Por fim, quanto à **reutilização**, todos os conjuntos de dados foram testados em cinco métricas: 1) FsF-R1-01MD: metadados especificam o conteúdo dos dados; 2) FsF-R1.1-01M: metadados incluem as licenças de reuso dos dados; 3) FsF-R1.2-01M: metadados incluem informações de proveniência sobre a criação ou geração de dados; 4) FsF-R1.3-01M: metadados seguem um padrão recomendado pela comunidade de pesquisa alvo dos dados; e 5) FsF-R1.3-02D: dados estão disponíveis em um formato de arquivo recomendado pela comunidade de pesquisa alvo. De acordo com Curty (2019, p. 179):

> Tais atributos são melhor verificáveis e julgados quando os dados estão documentados, acompanhados de informações suplementares e descrições que informem sobre características como, por exemplo, a proveniência dos dados e o contexto em que eles foram obtidos, os métodos para coleta, registros, transcrição, codificação e/ou processamento dos dados, suas diferentes versões, entre outros.

Pelo Gráfico 1 verifica-se que a reutilização é o ponto fraco dos dados de pesquisa depositados no Aleia, carecendo de melhorias. Quanto a métrica FsF-R1-01MD, 50% dos *datasets* foi avaliado em nível inicial, falhando em apresentar informações necessárias sobre o conteúdo dos dados nos metadados como, por exemplo, formato e tipo do arquivo. Isso evidencia a necessidade de uma avaliação nos *datasets* para aperfeiçoar a descrição do conteúdo dos metadados, ampliando a conformidade nessa métrica. É preciso lembrar que essas informações precisam estar legíveis por máquina, permitindo seu processamento automático.

O resultado foi similar para a métrica FsF-R1.1-01M (metadados incluem as licenças de reuso dos dados), em que metade dos conjuntos de dados obteve nível inicial de conformidade, enquanto a outra metade obteve nível avançado. Para os que atingiram o nível mais alto, a licença recuperada foi a *Creative Commons* (CC).

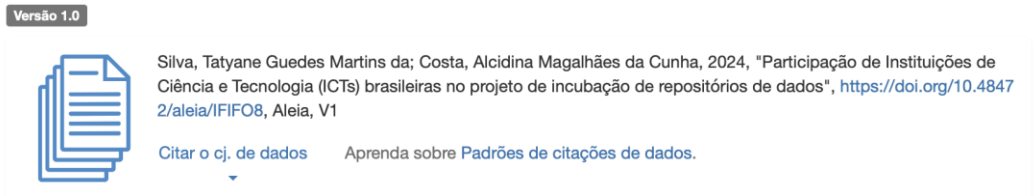
A CC é uma organização internacional sem fins lucrativos. Ela traz uma alternativa às licenças proprietárias, oferecendo aos criadores a flexibilidade para regular seu uso, modificação e distribuição (Creative Commons, n. d.). Atualmente, a CC oferece seis licenças aos criadores, que vão desde as mais restritivas (CCBY-NC-ND) às menos restritivas (CC0). A indicação clara sobre como os dados podem ser reutilizados, principalmente por meio de licenças internacionalmente conhecidas, é fundamental para o contexto FAIR. De acordo com o GO FAIR (n. d.), ambiguidades podem limitar severamente a reutilização dos dados. As condições sob as quais os dados podem ser reutilizados devem ser claras tanto para máquinas como para humanos.

Todos os oito conjuntos de dados do Aleia avaliados como avançado na métrica FsF-R1.1-01M possuíam a mesma licença atribuída: CC BY 4.0, que permite ao usuário: compartilhar (copiar e redistribuir o material em qualquer meio ou formato para qualquer finalidade, inclusive comercial) e adaptar (remixar, transformar e desenvolver o material para qualquer finalidade, até mesmo comercial). Mas, para isso, o usuário precisa atribuir os devidos créditos ao autor, fornecendo um *link* para a licença e indicando se foram feitas alterações nos dados originais. Além disso, o usuário não pode aplicar termos legais ou medidas tecnológicas que restrinjam legalmente outras pessoas de fazerem qualquer coisa que a licença permita.

Um fato interessante a ser pontuado sobre o *software* adotado para a construção do Aleia, o Dataverse, é que ele disponibiliza, de modo automático, a citação dos dados de pesquisa, conforme mostra a Figura 4. Isso facilita a reutilização dos dados de pesquisa, dando confiança ao pesquisador de que ele irá citar corretamente os dados reutilizados.

**Figura 4**

*Captura de tela de uma citação automática criada pelo Dataverse para o conjunto de dados no Aleia*




***Nota.*** Fonte: captura de tela do repositório Aleia (2024). *[Descrição da imagem]* Figura retangular horizontal com fundo branco. A imagem exibe um ícone de várias páginas uma em cima das outras no canto esquerdo. No canto direito, vê-se a seguinte citação: Silva, Tatyane Guedes Martins da; Costa, Alcidina Magalhães da Cunha, 2024, "Participação de Instituições de Ciência e Tecnologia (ICTs) brasileiras no projeto de incubação de repositórios de dados", https://doi.org/10.48472/aleia/IFIFO8, Aleia, V1". Uma linha abaixo, tem-se a opção "citar o conjunto de dados" e "Aprenda sobre Padrões de citações de dados". *[Fim da descrição].*

O *Dataverse Project* padroniza a citação de conjuntos de dados para tornar mais fácil para os pesquisadores publicarem seus dados e obterem crédito, bem como reconhecimento por seu trabalho. Quando um conjunto de dados é criado em um repositório Dataverse, a citação é gerada e apresentada automaticamente. Com uma estrutura de código aberto, o *Dataverse Project* implementa padrões aceitos pela comunidade para publicação de dados (Dataverse Project, n. d.). O padrão de citação definido pela equipe do Dataverse oferece reconhecimento adequado aos autores, bem como identificação permanente através do uso de identificadores globais e persistentes no lugar de URLs, que podem mudar com frequência. No site do *Dataverse Project*[5] é possível conferir, na íntegra, o padrão de citação adotado.

Já quanto à métrica FsF-R1.2-01M (metadados incluem informações de proveniência sobre a criação ou geração de dados), 100% dos *datasets* do Aleia forneciam um nível moderado de informações de proveniência. Alguns exemplos de metadados recuperados são: *publication_date, modified_date, creator, publisher* e *contributor*. Dados de proveniência fornecem contexto para os conjuntos de dados depositados, auxiliando outros pesquisadores a interpretarem os dados e em que cenário foram criados.

Curty (2019) explica que a reutilização está intimamente relacionada com a qualidade da documentação atrelada aos dados, fornecendo contexto. "Ao reutilizarem dados, os pesquisadores dependem diretamente de descrições e documentos suplementares que detalhem os dados e o estudo original para auxiliar na capacidade interpretativa do reutilizador sobre os pontos fortes, implicações e limitações dos dados" (Curty, 2019, p. 186). A capacidade de humanos e de máquinas de avaliar e selecionar dados com base em critérios relacionados às informações de origem é essencial para a reutilização, especialmente em escala. Para atingir o nível avançado, os metadados precisam fornecer informações de proveniência usando ontologias formais, como o caso do PROV-O. A ontologia PROV-O refere-se à especificação *Provenance Ontology* (PROV) na forma de uma ontologia *Web Ontology Language* (OWL). Ela foi desenvolvida pelo W3C para modelar e representar proveniência de dados na web e para ser usada em conjunto com outras tecnologias semânticas e padrões da web, como *Resource Description Framework* (RDF), possibilitando uma representação semântica precisa e interoperável da proveniência de dados.

Seguindo para a próxima métrica de reutilização, a FsF-R1.3-01M (metadados seguem um padrão recomendado pela comunidade de pesquisa alvo dos dados), 100% dos conjuntos de dados do Aleia atingiram um nível inicial de conformidade. É preciso implementar melhorias para que os metadados sigam um padrão recomendado pela comunidade de pesquisa alvo dos dados. O *Dublin Core* foi recuperado nos testes realizados pela F-UJI, mas ele consiste em um padrão de metadados não disciplinar, o que levou ao nível inicial de aderência relatado pela ferramenta.

De acordo com o GO FAIR (n. d.), é mais fácil reutilizar conjuntos de dados se estes forem semelhantes: mesmo tipo de dados, dados organizados de forma padronizada, documentação (metadados) seguindo um modelo comum e utilizando vocabulário comum. Se existirem padrões comunitários ou boas práticas para arquivamento e partilha de dados, estes deverão ser seguidos.

Nesse mesmo sentido, Hodson et al. (2018) afirmam que os padrões são essenciais de muitas maneiras no contexto do FAIR: desde metadados, vocabulários e ontologias para descrição de dados, até protocolos de

[5] Disponível em: https://dataverse.org/best-practices/data-citation.

transferência e troca para acesso a dados e padrões que regem a certificação de repositórios. O investimento em padrões da comunidade de pesquisa é essencial para níveis altos de *FAIRness*.

No relatório entregue pela ferramenta, alguns testes foram realizados (e não validados) e podem ser usados como base para a implementação de melhorias: 1) um padrão de metadados específico da comunidade de pesquisa é detectado usando *namespaces* ou esquemas encontrados em metadados fornecidos; 2) o padrão de metadados específico da comunidade está listado no *Registry of Research Data Repositories* (re3data) do repositório. Em ambos os testes os conjuntos de dados do Aleia não pontuaram, o que levou ao nível baixo de conformidade indicado pela F-UJI. Essas questões podem ser aperfeiçoadas para que os níveis de reutilização aumentem.

Vale citar que o re3data, lançado em 2012, é um registro global de repositórios de dados de pesquisa. Ele promove uma cultura de compartilhamento, acesso e visibilidade aos dados, sendo um serviço parceiro da *DataCite*, uma organização global sem fins lucrativos que está ativamente envolvida em várias iniciativas para melhorar a disponibilidade e citação de resultados de pesquisa (Rücknagel et al., 2021).

Por fim, de acordo com a F-UJI, apenas 40% dos conjuntos de dados estavam disponíveis em um formato de arquivo recomendado pela comunidade de pesquisa alvo (FsF-R1.3-02D). Nessa métrica era importante que os arquivos estivessem em formatos abertos e de longo prazo. Ao analisar os arquivos depositados, encontrou-se formatos proprietários como *MS Powerpoint* e *MS Excel Spreadsheet*, mas também formatos abertos como *Comma Separated Values* (CSV) e *OpenOffice Spreadsheet*. Arquivos proprietários dependem de *softwares* pagos para serem acessados, não atendendo ao que é defendido pelo movimento da Ciência Aberta. Isso pode acabar afetando também a reutilização desses dados ao longo do tempo, devido à obsolescência dos *softwares* e formatos.

Entretanto, é importante pontuar que o Guia de Usuário[6] do Aleia indica a preferência por formatos não proprietários, ou seja, esse é um desafio ligado diretamente ao pesquisador que deposita os dados. Silva e Rodrigues (2021, p. 133) retratam essa questão ao afirmarem que "[...] o acesso e sua reutilização não dependem apenas do desempenho do repositório, mas de características formais associadas com os conjuntos de dados e os processos associados à sua produção". Como nenhum formato de arquivo é barrado pelo repositório, abre margem para os usuários do serviço depositarem arquivos que não seguem as diretrizes de boas práticas. Isso leva a alguns debates importantes: será que é preciso criar políticas mandatórias para que usuários depositem arquivos em formatos abertos? A equipe gestora do repositório deveria devolver para revisão um conjunto de dados que possui arquivos proprietários? Ou, por ser apenas um formato preferencial, a equipe deve aceitar esses formatos? São pontos a serem discutidos e esclarecidos nas políticas do repositório.

De modo geral, os resultados obtidos por meio da ferramenta demonstram a necessidade de melhorias quanto à acessibilidade (oito *datasets* receberam 1 de 3) e reutilização (oito *datasets* receberam 5 de 10), em que os conjuntos de dados receberam as menores pontuações gerais de aderência. Entretanto, todos os conjuntos de dados obtiveram pontuações altas de localização e todos alcançaram níveis moderados de interoperabilidade, o que é um bom indicativo frente ao cenário nacional e internacional.

A maior pontuação geral foi 79% de *FAIRness* (para seis *datasets*), enquanto a menor foi 62% (para apenas um *dataset*), o que representa um bom resultado comparado a outros estudos similares em nível nacional, como o de Bonetti (2023). Nele, seis repositórios de dados de pesquisa institucionais de São Paulo foram avaliados com a ferramenta F-UJI, e o maior percentual de aderência ao FAIR foi 50% para 62 dos 68 conjuntos de dados depositados no repositório da Unicamp. Para os outros cinco repositórios da amostra, todos os percentuais encontrados foram inferiores a 40% de *FAIRness*. Em outro estudo similar que avaliou os *datasets* depositados no COVID-19 *Data Sharing*/BR, o nível de *FAIRness* foi de 54% (Simões et al., 2021). Ou seja, percebe-se que a implementação dos Princípios FAIR no repositório Aleia está em um estágio mais avançado.

Como visto no Gráfico 1, das 16 métricas avaliadas pela F-UJI, sete foram validadas para 100% dos conjuntos de dados em nível avançado, ou seja, todos os requisitos testados estavam presentes. Dessas sete métricas, três estão ligadas à localização e duas à interoperabilidade. Como já mencionado, esses foram os pontos fortes para o repositório Aleia. Já quanto à reutilização, melhorias podem ser implementadas para todas as cinco métricas. Nenhuma delas foi validada em nível avançado para 100% dos *datasets*. Recomenda-se que esses aspectos sejam revistos pela equipe gestora do Aleia para serem aperfeiçoados.

[6] Disponível em: https://aleia.ibict.br/dvn/guide/guiaAleia.html.

# 5 Conclusões

O volume e a velocidade com que os dados de pesquisa são produzidos ultrapassam a capacidade humana de processamento, o que leva à necessidade do uso de máquinas para contornar essas barreiras. Mas, ao contrário do humano que tem um senso intuitivo de semântica, as máquinas precisam que os dados estejam bem documentados e estruturados para que possam processá-los automaticamente. É nesse cenário que as boas práticas como os Princípios FAIR se tornam essenciais, permitindo que os agentes computacionais acessem, interpretem, combinem e realizem tarefas automáticas com os dados.

Ciente dessa importância, este estudo buscou avaliar a aderência aos Princípios FAIR dos conjuntos de dados de pesquisa depositados no repositório Aleia, repositório institucional de dados do Ibict. Sendo o Ibict um instituto de pesquisa reconhecido, é fundamental que seu repositório de dados siga as boas práticas internacionalmente adotadas, trazendo maior visibilidade para a instituição. A adoção de boas práticas é, inclusive, um passo importante para a certificação do repositório por entidades como o *CoreTrustSeal*.

Por meio dos dados coletados com a ferramenta F-UJI, foi possível identificar que os conjuntos de dados do repositório Aleia encontra-se, no geral, em um **nível moderado de *FAIRness***. Os *datasets* obtiveram percentuais de aderência geral que variaram entre 62% a 79%, o que leva em conta todos os Princípios FAIR. Isso é um bom indicativo considerando que o Aleia é um repositório recente, lançado em 2023, e que esta foi a primeira avaliação realizada no repositório em busca de melhorias quanto aos Princípios FAIR.

Nenhum dos conjuntos de dados da amostra conseguiu aderir a todos os princípios, mas existem pontos fortes no repositório como a localização e a interoperabilidade. Sete *datasets* receberam nota 6,5 de 7 enquanto oito receberam a nota máxima em localização. O outro ponto forte dos dados de pesquisa da amostra foi a interoperabilidade: todos os *datasets* receberam nota 3 de 4. Isso é um bom indicativo, principalmente considerando os estudos nacionais e internacionais revisados que apontam a interoperabilidade como um desafio das práticas FAIR.

Quanto à localização, a única métrica que demanda melhorias é a que diz respeito aos metadados incluírem o identificador dos dados que descrevem, o que pode ser aperfeiçoado pela equipe gestora. Já em interoperabilidade, outro ponto forte do Aleia, a melhoria a ser implementada diz respeito ao uso de recursos semânticos nos metadados como vocabulários controlados. A interoperabilidade é uma característica essencial no valor e na usabilidade dos dados, e por isso a importância de se atingir níveis cada vez maiores de aderência.

A acessibilidade, em contrapartida, pode ser aperfeiçoada. Para alcançar níveis maiores, é preciso investir na declaração do nível e das condições de acesso aos dados, além de verificar o protocolo de comunicação padronizado para os dados. Outro ponto fraco do Aleia é a reutilização. Para otimizá-la, os metadados do conjunto de dados precisam seguir um padrão recomendado pela comunidade de pesquisa alvo e os dados precisam estar disponíveis em um formato de arquivo recomendado pela comunidade. Mas o Aleia, em seu Guia de Usuários, já indica a preferência por formatos abertos e não proprietários, incentivando sua escolha pelos pesquisadores que depositam seus dados. Outro ponto a ser aperfeiçoado é a questão dos metadados especificarem o conteúdo dos dados (tamanho, tipo do arquivo, etc.), de tal forma que máquinas consigam processar a informação automaticamente. Logo, existem melhorias a serem implementadas.

Vale lembrar que os resultados aqui expostos se limitam ao que foi entregue automaticamente pela ferramenta auxiliar F-UJI, que se mostrou bastante simples e eficaz na avaliação dos *datasets*. Como mencionado no trabalho, métricas bem definidas são necessárias para avaliar o *FAIRness* dos dados e certificar serviços. Todas as métricas da F-UJI estão disponíveis abertamente no site e qualquer outro pesquisador pode, inclusive, reproduzir esta pesquisa utilizando a ferramenta e os mesmos conjuntos de dados, que estão em acesso aberto. É importante destacar ainda que o foco do FAIR é a otimização do processamento automático pelos computadores, e por isso a importância dada para informações e dados legíveis por máquinas. Todas as melhorias implementadas precisam levar essa funcionalidade em conta para serem, de fato, FAIR.

Identifica-se como limitação da pesquisa o volume de *datasets* analisado, que é restrito. Por outro lado, apenas com um conjunto pequeno de dados foi possível aprofundar a análise em cada um dos Princípios FAIR, o que possibilitou também explorar a ferramenta F-UJI em todas as suas perspectivas. Futuramente, pretende-se dar continuidade à pesquisa incluindo a análise dos novos *datasets* depositados no Aleia e a partir do estudo do uso de ontologias formais para descrição dos dados.

# Referências


Aleia. (2024). *Sobre o Aleia*. https://aleia.ibict.br/about.xhtml

Araújo, D. G., Dias, G. A., & Autran, M. M. M. (2021). Compartilhamento de dados no contexto da ciência brasileira: Um estudo integrativo. *Informação & Informação, 26*(3), 100-124. https://doi.org/10.5433/1981-8920.2021v26n3p100

Bonetti, L. G. (2023). *Metadados e os Princípios FAIR: Um estudo dos repositórios de dados de pesquisa em São Paulo*. [Mestrado, Universidade Federal de São Carlos]. Repositório da Universidade Federal de São Carlos. https://repositorio.ufscar.br/handle/ufscar/17448

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world***.** The MIT Press.

Córdula, F. R., & Araújo, W. J. (2019). O compartilhamento de dados científicos na era do e-Science. In G. A. Dias & B. M. J. F. Oliveira (Orgs.), *Dados científicos: Perspectivas e desafios*. (pp. 177-187). Editora UFPB. http://www.editora.ufpb.br/sistema/press5/index.php/UFPB/catalog/book/359

Creative Commons. (n.d.). *Licenses list.* https://creativecommons.org/licenses

Curty, R. (2019). Abordagens de reúso e a questão da reusabilidade dos dados científicos. *Liinc em Revista, 15*(2). 177-193. https://doi.org/10.18617/liinc.v15i2.4777

Dataverse Project. (n.d.) *About the project*. https://dataverse.org/about

Devaraju, A., & Huber, R. (2020). F-UJI: An automated FAIR Data assessment tool [Dataset]. ZENODO. https://doi.org/10.5281/zenodo.4063720

Dias, G. A., Bandeira, L. K. R., & Anjos, R. L. (2023). Objetos digitais Fair: Abordando a complexidade no processo de gestão de dados de pesquisa. *Archeion Online, 11*(2). 219-232. https://doi.org/10.22478/ufpb.2318-6186.2023v11n2.68962

Dudziak, E. (2016, Dezembro 7). Dados de Pesquisa agora devem ser armazenados e citados. *ABCD USP.* https://www.abcd.usp.br/noticias/dados-materiais-metodos-revistas-exigem-dados-pesquisa-estejam-disponiveis/

Dunning, A., Smaele, M., & Böhmer, J. (2017, February 20–23). *Are the FAIR Data Principles fair?* [Preprint]. 12th International Digital Curation Conference (IDCC 2017), Edinburgh, EDI, Scotland. https://doi.org/10.5281/zenodo.321423

GO FAIR. (n.d.). *GO FAIR initiative.* https://www.go-fair.org/go-fair-initiative

Groehs, A., Silveira, C. Z., & Lima, E. M., Dittrich, E., Matos, I. Y. R., Ribeiro, D. I., Jr. (2023). LattesData e a adoção aos Princípios FAIR: Uma análise usando a F-UJI Automated FAIR Data Assessment Tool. *Em Questão, 29*, Artigo e130018. https://doi.org/10.1590/1808-5245.29.130018

Guandalini, C. A., Furnival, A. C. M., & Arakaki, A. C. S. (2019). Boas práticas científicas na elaboração de planos de gestão de dados. *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação, 17*, Artigo e019034. https://doi.org/10.20396/rdbci.v17i0.8655895

Henning, P. C., Ribeiro, C. J. S., Sales, L. F., Moreira, J. L. R., & Santos, L. O. B. S. (2019). Desmistificando os princípios fair: Conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados fair. *Pesquisa Brasileira em Ciência da Informação e Biblioteconomia, 14*(3), 175-192. https://pbcib.com/index.php/pbcib/article/view/46969/27455

Henning, P. C., Silva, L. O. B., Pires, L. F., Sinderen, M. V., & Moreira, J. L. R. (2021). The FAIRness of data management plans: An assessment of some European DMPs. *Revista Eletrônica de Comunicação, Informação & Inovação em Saúde, 15*(3), 722-735. https://doi.org/10.29397/reciis.v15i3.2270

Higman, R., Bangert, D., & Jones, S. (2019). Three camps, one destination: The intersections of research data management, FAIR and Open. *Insights, 32*(1). https://doi.org/10.1629/uksg.468

Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., Mietchen, D., Petrauskaité, R., & Wittenburg, P. (2018, June). *Turning FAIR data into reality: Interim report from the European Commission Expert Group on FAIR data* [Version Interim draft]. https://doi.org/10.5281/zenodo.1285272

Juty, N., Wimalaratne, S. M., Soiland-Reyes, S., Kunze, J., Goble, C. A., & Clark, T. (2020). Unique, Persistent, Resolvable: Identifiers as the Foundation of FAIR. *Data Intelligence, 2*(1-2), 30-39. https://doi.org/10.1162/dint_a_00025

Nicholson, C., Kansa, S., Gupta, N., & Fernandez, R. (2023). Will it ever be FAIR? Making archaeological data findable, accessible, interoperable, and reusable. *Advances in Archaeological Practice*, *11*(1), 63-75. https://doi.org/10.1017/aap.2022.40

OECD. (2007). *OECD Principles and guidelines for access to research data*. https://www.oecd.org/en/publications/oecd-principles-and-guidelines-for-access-to-research-data-from-public-funding_9789264034020-en-fr.html

Petrosyan, L., Aleixandre-Benavent, R., Peset, F., Valderrama-Zurián, J. C., Ferrer-Sapena, A., & Sixto-Costoya, A. (2023). FAIR degree assessment in agriculture datasets using the F-UJI tool. *Ecological Informatics, 76*, Article e102126. https://doi.org/10.1016/j.ecoinf.2023.102126

Prodanov, C. C., & Freitas, E. C. (2013). *Metodologia do trabalho científico: Métodos e técnicas da pesquisa e do trabalho acadêmico*. Feevale.

Rocha, R. P., Gabriel, R. F., Jr., Vanz, S. A. S., Borges, E. N., Azambuja, L. A. B., Caregnato, S. E., Pavão, C. G., Passos, P. C. S. J., & Felicissimo, C. H. (2021). Análise dos sistemas DSpace e Dataverse para repositórios de dados de pesquisa com acesso aberto. *Revista Brasileira de Biblioteconomia e Documentação*, *17*, 1-25. https://rbbd.febab.org.br/rbbd/article/view/1572

Rücknagel, J., Vierkant, P., Ulrich, R., Kloska, G., Schnepf, E., Fichtmüller, D., Reuter, E., Semrau, A., Kindling, M., Pampel, H., Witt, M., Fritze, F., Van de Sandt, S., Klump, J., Goebelbecker, H. J., Skarupianski, M., Bertelmann, R., Schirmbacher, P., Scholze, F., … Kirchhoff, A. (2015). *Metadata schema for the description of research data repositories: Version 3.0.* GFZ. http://doi.org/10.2312/re3.008

Sales, L., Henning, P., Veiga, V., Costa, M. M., Sayão, L. F., Santos, L. O. B. S., & Pires, L. F. (2020). GO FAIR Brazil: A challenge for Brazilian data science. *Data Intelligence, 2*(1-2), 238-245. https://doi.org/10.1162/dint_a_00046

Sanchez, F. A., Vidotti, S. A. B. G., & Vechiato, F. L. (2017). A contribuição da curadoria digital em repositórios digitais. *Revista Informação na Sociedade Contemporânea, 1*, 1-17. https://doi.org/10.21680/2447-0198.2017v1n0ID12280

Sayão, L. F., & Sales, L. F. (2016). Algumas considerações sobre os repositórios digitais de dados de pesquisa. *Informação & Informação, 21*(2), 90-115. https://doi.org/10.5433/1981-8920.2016v21n2p90

Sayão, L. F., & Sales, L. F. (2020). Afinal, o que é dado de pesquisa? *BIBLOS: Revista do Instituto de Ciências Humanas e da Informação, 34*(2), 32-51. https://doi.org/10.14295/biblos.v34i2.11875

Silva, F. C. C., & Rodrigues, M. M. (2021). Implementação dos Princípios FAIR em repositórios de dados científicos: Uma análise comparativa das infraestruturas de software do DSpace e Dataverse. In L. F. Sales, V. S. O Veiga, P. Henning, & L. F. Sayão (Orgs.). *Princípios FAIR aplicados à gestão de dados de pesquisa*. (pp.123-136). Ibict. https://ridi.ibict.br/bitstream/123456789/1182/2/IBICT_Principios%20FAIR%20aplicados%20a%20gest%C3%A3o%20de%20dados%20de%20pesquisa_2021.pdf

Silva, L. C., Santarem Segundo, J. E., & Silva, M. F. (2018). Princípios fair e melhores práticas do linked data na publicação de dados de pesquisa. *Informação & Tecnologia, 5*(2), 81-103. https://doi.org/10.22478/ufpb.2358-3908.2018v5n2.44812

Silva, M. F., Martins, D. L., & Siqueira, J. (2019). Web semântica em repositórios: Ontologia para representação de bibliotecas digitais. *Ciência da Informação em Revista, 6*(1), 99-113. http://dx.doi.org/10.28998/cirev.2019v6n1f

Simionato, A. C. (2017, Outubro 23-27). *Mapeamento dos metadados para dados científicos* [Comunicação oral]. XVIII Encontro Nacional de Pesquisa em Ciência da Informação (Enancib 2017), Marília, SP, Brasil. http://enancib.marilia.unesp.br/index.php/XVIII_ENANCIB/ENANCIB/paper/view File/563/874

Simões, A. R. C., Anjos, R. L., & Dias, G. A. (2021). Análise dos conjuntos de dados disponíveis no repositório COVID-19 Data Sharing/BR à luz dos Princípios FAIR. In L. F. Sales, V. S. O. Veiga, P. Henning, & L. F. Sayão (Orgs.). *Princípios FAIR aplicados à gestão de dados de pesquisa*. (pp. 91-102). Ibict. https://ridi.ibict.br/bitstream/123456789/1182/2/IBICT_Principios%20FAIR%20aplicados%20a%20gest%C3%A3o%20de%20dados%20de%20pesquisa_2021.pdf

Tartarotti, R. C., Dal'evedove, P. R., & Fujita, M. S. L. (2019). Biblioteconomia de dados em repositórios de pesquisa: Perspectivas para a atuação bibliotecária. *Informação & Informação, 24*(3), 207-226. https://doi.org/10.5433/1981-8920.2019v24n3p207

Torino, E., Coneglian, C. S., & Vidotti, S. A. B. G. (2023, Novembro 6-10). *Princípios fair e a compreensão de dados por máquinas* [Trabalho completo]. XXIII Encontro Nacional de Pesquisa em Ciência da Informação (Enancib 2023), Aracaju, SE, Brasil. https://doi.org/10.22477/vii.widat.206

Vidotti, S. A. B. G., Torino, E., & Coneglian, C. S. (2021). #SejaJUSTOeCUIDADOSO: Princípios FAIR e CARE na gestão de dados de pesquisa. In L. F. Sales, V. S. O. Veiga, P. Henning, & L. F. Sayão (Orgs.). *Princípios FAIR aplicados à gestão de dados de pesquisa.* (pp. 209-221). Ibict. https://ridi.ibict.br/bitstream/123456789/1182/2/IBICT_Principios%20FAIR%20aplicados%20a%20gest%C3%A3o%20de%20dados%20de%20pesquisa_2021.pdf

Wilkinson, M., & Dumontier, M. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci Data, 3*, Article e160018. https://doi.org/10.1038/sdata.2016.18

# Apêndice A

**Quadro A**

*Relação de datasets avaliados*

| Identificador do dataset | Título do dataset | Autores do dataset |
|---|---|---|
| https://doi.org/10.48472/aleia/UNVCQO | Gestão de Dados FAIR: Em busca de pesquisas reprodutíveis no campo da Linguística | Luana Sales |
| https://doi.org/10.48472/aleia/G3BP3A | Relatórios Jardim Botânico de Brasília | Ibict; Cariniana |
| https://doi.org/10.48472/aleia/DPW8OE | Ideologia dos Partidos Políticos no Brasil após a redemocratização | Josir Cardoso Gomes |
| https://doi.org/10.48472/aleia/AOCTZO | Deputados Federais Evangélicos de 1933 a 2018 | Josir Cardoso Gomes |
| https://doi.org/10.48472/aleia/Y7FXK8 | Dataset Timeline Covid-19 | Tainá Regly; Priscila R. Carvalho; Ricardo M. Pimenta |
| https://doi.org/10.48472/aleia/MFHBCK | Sequenciamento de Bevacizumabe | Ibict |
| https://doi.org/10.48472/aleia/7DPOHV | Cobertura de sequências | Marcelo Valle de Sousa |
| https://doi.org/10.48472/aleia/21GSV8 | Normas do Tamanho da Categoria para 905 palavras da Lingua portuguesa | Gerson Americo Janczura; Goiara Mendonça Castilho; Nelson Rocha Oliveira |

| | | |
|---|---|---|
| https://doi.org/10.48472/aleia/KJOHIT | Normas de associação livre para 1004 palavras da língua portuguesa | Gerson Americo Janczura; Goiara Mendonça Castilho; Victor Nahuel Keller; Nelson Rocha Oliveira |
| https://doi.org/10.48472/aleia/IFIFO8 | Participação de Instituições de Ciência e Tecnologia (ICTs) brasileiras no projeto de incubação de repositórios de dados | Tatyane Guedes Martins da Silva; Alcidina Magalhães da Cunha Costa |
| https://doi.org/10.48472/aleia/M038PL | Portais agregadores: levantamento de dados sobre acessibilidade do Oasisbr, RCAAP e La Referencia | Blena Estevam dos Santos; Millena Cordeiro Matos de Lima; Denise Aparecida Freitas de Andrade; Cássio Teixeira de Morais; Priscila Machado Borges Sena; Bianca Amaro |
| https://doi.org/10.48472/aleia/THFG9S | Investigação da padronização de metadados adotados no Oasisbr | Gabriel Silveira Marques; Blena Estevam dos Santos; Millena Cordeiro Matos de Lima; Cássio Teixeira de Morais |
| https://doi.org/10.48472/aleia/8WSOGT | Uso do DSpace para a construção e manutenção de repositórios institucionais de publicações científicas | Millena Cordeiro Matos de Lima; Blena Estevam dos Santos; Cássio Teixeira de Morais |
| https://doi.org/10.48472/aleia/Y1O6PD | Conjunto de dados das retratações dos artigos a partir das bases OpenAlex e Retraction Watch | Denise Aparecida Freitas de Andrade; Fhillipe de Freitas Campos; Fábio Lorensi do Canto; Thiago Magela Rodrigues Dias |
| https://doi.org/10.48472/aleia/XYFBKD | Revisão sistemática dos acordos transformativos: utilizando dados da Scopus e Web of Science | Gabriel Silveira Marques |
| https://doi.org/10.48472/aleia/HD1ND5 | Levantamento de acessibilidade digital no Portal Oasisbr | Denise Aparecida Freitas de Andrade; Millena Cordeiro Matos de Lima; Blena Estevam dos Santos; Tatyane Guedes Martins da Silva; Letícia Guarany Bonetti; Cássio Teixeira de Morais; Priscila Machado Borges Sena; Bianca Amaro |

***Nota.*** Fonte: elaborado pelos autores (2024).

# Dados de publicação

**Letícia Guarany Bonetti**
Mestre em Ciência da Informação
**Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília, DF, Brasil**
leticiabonetti@ibict.br
https://orcid.org/0000-0002-3012-8465

Bacharel em Biblioteconomia pela Universidade de Brasília (2019) e Mestre em Ciência da Informação pela Universidade Federal de São Carlos (2023). Atualmente atua como bolsista do Programa de Capacitação Institucional do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) desenvolvendo serviços na área de Ciência Aberta e repositórios de dados.

**Tatyane Guedes Martins da Silva**
Graduada em Biblioteconomia
**Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília, DF, Brasil**
tatyanesilva@ibict.br
https://orcid.org/0000-0002-1743-0467

Bacharel em Biblioteconomia pela Universidade de Brasília (2019). Foi bolsista do Programa de Capacitação Institucional do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) desenvolvendo serviços na área de Ciência Aberta e repositórios de dados. Possui como principais temas de interesse: softwares livres para bibliotecas, ciência aberta, dados abertos, gestão de dados de pesquisa, repositórios, metadados e princípios FAIR.

**Caterina Groposo Pavão**
Doutora em Comunicação e Informação
**Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil**
caterina@cpd.ufrgs.br
https://orcid.org/0000-0003-3712-7200

Bacharel em Biblioteconomia pela Universidade Federal do Rio Grande do Sul (UFRGS), mestre e doutora em Comunicação e Informação pelo Programa de Pós-Graduação em Comunicação e Informação da mesma universidade. Atualmente exerce atividade docentes no Curso de Biblioteconomia do Departamento de Ciência da Informação da Faculdade de Biblioteconomia e Comunicação da Universidade Federal do Rio Grande do Sul e no Programa de Pós-Graduação em Ciência da Informação da Faculdade de Biblioteconomia e Comunicação da mesma Universidade. Atualmente é professora adjunta da Universidade Federal do Rio Grande do Sul.

**Samile Andréa de Souza Vanz**
Doutora em Comunicação e Informação
**Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil**
samilevanz@terra.com.br
https://orcid.org/0000-0003-0549-4567

Graduada em Biblioteconomia pela Universidade Federal do Rio Grande do Sul (1999), mestre e doutora em Comunicação e Informação pelo PPGCOM UFRGS (2004 e 2009), com estágio sanduíche na Dalian University of Technology (China, 2007-2008). Pós-doutorado pela Universidad Carlos III de Madrid (Madrid, 2016).

**Rene Faustino Gabriel Junior**
Doutor em Ciência da Informação
**Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil**
rene.gabriel@ufrgs.br
http://orcid.org/0000-0003-1021-3360

Graduação em Biblioteconomia e Documentação pela Pontifícia Universidade Católica do Paraná (2008), mestrado em Gestão da Informação pela Universidade Federal do Paraná (2011) e doutorado em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (2014). Atualmente é professor adjunto da Universidade Federal do Rio Grande do Sul.

**Ana Carolina Simionato Arakaki**
Doutora em Ciência da Informação

**Universidade de Brasília, Brasília, DF, Brasil**
**Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília, DF, Brasil**
**Universidade Federal de São Carlos, São Carlos, SP, Brasil**
acsimionato@ufscar.br
https://orcid.org/0000-0002-0140-9110

Professora Adjunta e pesquisadora da Faculdade de Ciência da Informação da Universidade de Brasília (UnB), atuando também no Programa de Pós-Graduação em Ciência da Informação (PPGCI/UFSCar) da Universidade Federal de São Carlos (UFSCar). Pós-doutorado (2022) na Universidade de São Paulo (USP) na área de Computação. Doutorado (2015) e mestrado (2012) em Ciência da Informação pela Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP). Lidera o grupo de pesquisa e o programa de extensão "Dados e Metadados".

### Originalidade
Declaro que o texto é original, e não está em avaliação em nenhuma outra publicação. Caso decida cancelar o processo de publicação, comprometo-me a informar imediatamente a equipe editorial da Revista Biblios para que a submissão seja arquivada.

### Preprint
Declaro que o manuscrito produzido não foi submetido a nenhuma plataforma de Preprints.

### Informações sobre o trabalho
Não se aplica

### Agradecimentos
Agradecemos ao Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) pelo suporte e pelo incentivo à pesquisa na área de repositórios de dados de pesquisa. Sem o suporte e a infraestrutura disponibilizados, este trabalho não teria sido possível. Esse compromisso do Ibict com a Ciência Aberta tem sido fundamental para promover a transparência, o compartilhamento e a reutilização de dados de pesquisa, impulsionando o avanço da pesquisa científica.

### Contribuição de autoria
Concepção e elaboração do manuscrito: L. G. Bonetti

Coleta de dados: L. G. Bonetti, T. G. M. Silva

Análise de dados: L. G. Bonetti, T. G. M. Silva, R. F. Gabriel Junior, S. A. S. Vanz, C. G. Pavão, A. C. S. Arakaki

Discussão dos resultados: L. G. Bonetti, T. G. M. Silva, R. F. Gabriel Junior, S. A. S. Vanz, C. G. Pavão, A. C. S. Arakaki

Revisão e aprovação: L. G. Bonetti.

### Uso de Inteligência artificial
Não se aplica.

### Financiamento
Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

### Permissão para usar imagens
Não se aplica.

### Aprovação do Comitê de Ética em Pesquisa
Não se aplica.

### Conflito de interesses
Não se aplica.

### Disponibilidade de dados e conjunto dos dados

**Os dados foram submetidos como materiais suplementares**: O conjunto de dados que dá suporte aos resultados deste estudo foi enviado para publicação na seção “Materiais Suplementares”.

**Os dados já estão disponíveis em outros locais:**

a) Título do conjunto de dados: Avaliação dos conjuntos de dados do repositório Aleia quanto aos princípios FAIR

b) DOI: https://doi.org/10.48472/aleia/LH6ZFZ

### Licença de uso

Os autores concedem à Biblios direitos exclusivos de primeira publicação, com o trabalho simultaneamente licenciado sob uma Licença *Creative Commons Atribuição (CC BY) 4.0 Internacional*. Esta licença permite que terceiros remixem, adaptem e desenvolvam o trabalho publicado, dando os devidos créditos pela autoria e publicação inicial neste periódico. Os autores estão autorizados a firmar acordos adicionais separados para distribuição não exclusiva da versão publicada do trabalho no periódico (por exemplo, publicação em um repositório institucional, em um site pessoal, publicação de uma tradução ou como um capítulo de livro), com reconhecimento de autoria e publicação inicial neste periódico.

### Publisher

Publicação da *University Library System of University of Pittsburgh*. Responsabilidade compartilhada com as universidades conveniadas. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

### Editores

Lúcia da Silveira e Fabiano Couto Corrêa da Silva.

### Histórico

Recebido em: 05-09-2024 – Aprovado em: 11-04-2025 – Publicado em: 03-07-2025.

The articles in this journal are licensed under a Creative Commons Attribution 4.0 United States License.

Pitt Open Library Publishing

This journal is published by Pitt Open Library Publishing.