



15ª Conferência Lusófona de Ciência Aberta (ConfOA) Acesso Aberto e Dados de Investigação Abertos: sistemas, políticas e práticas



Modalidade: Pecha Kucha

Níveis de *FAIRness* nos repositórios de dados Aleia e Deposita Dados

FAIRness levels in the Aleia and Deposita Dados data repositories

Letícia Guarany Bonetti

Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Brasília, Distrito Federal, Brasil

<https://orcid.org/0000-0002-3012-8465>

Tatyane Guedes Martins da Silva

Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Brasília, Distrito Federal, Brasil

<https://orcid.org/0000-0002-1743-0467>

Rene Faustino Gabriel Junior

Universidade Federal do Rio Grande do Sul (UFRGS)

Porto Alegre, Rio Grande do Sul, Brasil

<https://orcid.org/0000-0003-1021-3360>

Marcel Garcia de Souza

Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Brasília, Distrito Federal, Brasil

<https://orcid.org/0000-0003-2255-199X>

Henrique Fernandes Rodrigues

Universidade Federal do Rio de Janeiro (UFRJ)

Rio de Janeiro, Rio de Janeiro, Brasil

<https://orcid.org/0000-0002-6777-3935>

Ana Carolina Simionato Arakaki

Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Brasília, Distrito Federal, Brasil

<https://orcid.org/0000-0002-0140-9110>

RESUMO:

Os repositórios de dados de pesquisa buscam disseminar e preservar os dados. Estes, por sua vez, precisam seguir boas práticas, como os princípios FAIR, para potencializar seus benefícios. O objetivo deste trabalho é avaliar os níveis de *FAIRness* dos *datasets* depositados nos repositórios Aleia e Deposita Dados. Trata-se de uma pesquisa descritiva. Utilizou-se a ferramenta F-UJI para a avaliação automática dos conjuntos de dados. Verificou-se que os *datasets* do Aleia e do Deposita Dados encontram-se em um nível moderado de *FAIRness*. Ambos os repositórios destacaram-se quanto à localização, mas podem adotar melhorias quanto à reutilização.

Palavras-chave: dados de pesquisa; princípios FAIR; repositórios de dados de pesquisa; Ciência Aberta.

INTRODUÇÃO

No contexto das unidades de pesquisa, como o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), os dados de pesquisa são essenciais para projetos, colaboração entre grupos, validação de resultados, inovação e tomada de decisões estratégicas. O investimento em repositórios de dados se mostra, portanto, de suma importância, uma vez que eles “[...] constituem o lugar mais apropriado para que seus dados sejam preservados e possam ser recuperados, acessados e citados por outros pesquisadores, ou seja, tenham visibilidade em escala mundial” (Sayão; Sales, 2015, p. 49). Os repositórios são definidos por Sanchez, Vidotti e Vechiato (2017, p. 3) 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”.

Atualmente, o Ibict coordena dois repositórios de dados de pesquisa de acesso aberto: o Aleia e o Deposita Dados, e ambos oferecem um ambiente digital seguro para custodiar e compartilhar conjuntos de dados. Enquanto o **Aleia** é voltado para a comunidade científica do Ibict (servidores, pesquisadores visitantes, bolsistas e alunos do programa de pós-graduação), o **Deposita Dados** é voltado para pesquisadores brasileiros vinculados a instituições científicas que ainda não dispõem de seus repositórios de dados de pesquisa. Também podem depositar os brasileiros que coletaram seus dados por meio de colaboração científica em instituições estrangeiras de ensino e pesquisa.

Ambos adotam o software livre Dataverse, desenvolvido desde 2006 pelo *Institute for Quantitative Social Science* (IQSS) de Harvard em colaboração com profissionais de todo o mundo, por meio do *Dataverse Project* (Dataverse Project, 2020). Diferente do DSpace que é voltado para publicações de documentos, o Dataverse foi criado pensando nos dados de pesquisa, o que auxilia na gestão dos *datasets* e pode facilitar a implementação dos princípios FAIR. Com isso em mente, o objetivo deste trabalho é avaliar os níveis de *FAIRness* dos conjuntos de dados depositados no Aleia e no Deposita Dados.

DESENVOLVIMENTO

Os princípios FAIR são um acrônimo para “*Findable*” (Localizável), “*Accessible*” (Acessível), “*Interoperable*” (Interoperável) e “*Reusable*” (Reutilizável). De acordo com Wilkinson *et al.* (2016), tratam-se de princípios independentes de domínio que enfatizam o aprimoramento da capacidade das máquinas de localizar e usar os dados automaticamente (*machine-actionable*). Considerando que os humanos dependem cada vez mais do suporte computacional para lidar com o aumento do volume, complexidade e velocidade de criação de dados, torna-se essencial a adoção de diretrizes que possam otimizar o acesso e o uso dos dados de pesquisa, facilitando o trabalho dos agentes computacionais.

Os princípios FAIR são, nesse cenário, reconhecidos mundialmente como elementos-chave para boas práticas em todos os processos de gestão de dados (Sales *et al.*, 2020). Quando os dados são compartilhados em repositórios que investem em boas práticas, como o FAIR, há a ampliação dos benefícios extraídos dos dados depositados, o que também permite um acesso mais amplo, transparente e justo a eles, conforme preconiza o movimento da Ciência Aberta.

Procedimentos metodológicos

Por meio de uma pesquisa descritiva quali-quantitativa, buscou-se avaliar os níveis de *FAIRness* dos conjuntos de dados depositados nos repositórios de dados do Ibict, considerando a importância do investimento em boas práticas em repositórios. A avaliação se deu de forma automática por meio de uma ferramenta on-line e gratuita: a F-UJI¹ *Automated FAIR Data Assessment Tool*. Ela é resultado do projeto FAIRsFAIR², que recebeu financiamento do *European Union’s Horizon 2020*. A F-UJI é baseada em *Representational State Transfer* (REST) e a avaliação se deu por meio dos metadados dos conjuntos de dados da amostra. As métricas estabelecidas pelos desenvolvedores (Devaraju; Huber, 2020) como base do sistema de avaliação da ferramenta podem ser consultadas integralmente no site.

A F-UJI entrega: um resumo da aderência geral dos conjuntos de dados ao FAIR; e um relatório detalhado de quais métricas foram avaliadas (exemplo: em localizável, a ferramenta testa se foi atribuído um identificador persistente). Ela também indica qual foi o nível de *FAIRness* do *dataset* para cada uma das métricas. Mas, neste trabalho, o foco é analisar os dados entregues pela F-UJI na parte de seu resumo, de forma a possibilitar uma visão geral da aderência ao FAIR de cada um dos conjuntos de dados. A amostra foi composta por 12 *datasets* depositados no Aleia e no Deposita Dados, e as avaliações ocorreram no dia 7 de fevereiro de 2024. Os dados coletados foram sintetizados em planilhas³ para análises, compartilhamento e preservação em um repositório de dados.

1 Disponível em: <https://www.f-ujl.net/>.

2 Disponível em: <https://www.fairsfair.eu/>.

3 Disponível em: https://docs.google.com/spreadsheets/d/1fCFbaXjjsPOO-ghZEhFsujmwkMx3_v8RYB2scnlpUQo/edit?usp=sharing

Análise dos resultados

Nos Quadros 1 e 2, composto pelo resumo da avaliação do F-UJI, é possível verificar as pontuações para *Findable* (localizável) (a nota máxima é sete), *Accessible* (acessível) (a nota máxima é três), *Interoperable* (interoperável) (a nota máxima é quatro) e *Reusable* (reutilizável) (a nota máxima é 10). Todas essas pontuações máximas de referência foram definidas pela própria ferramenta F-UJI com base em suas métricas e escalas, assim como as pontuações atribuídas a cada *dataset*. Nos Quadros também é possível observar o percentual de aderência geral de cada *dataset* ao FAIR (0 a 100%).

QUADRO 1 - Aderência dos conjuntos de dados do repositório Aleia aos princípios FAIR

URL dos datasets	Código do dataset	Pontuação em localizável	Pontuação em acessível	Pontuação em interoperável	Pontuação em reutilizável	Percentual de aderência geral
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/DPW80E	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%

Fonte: Elaborado pelos autores (2024).

Nota-se que os conjuntos de dados depositados no Deposita Dados (**QUADRO 2**) obtiveram, no geral, pontuações mais altas de aderência ao FAIR do que os do Aleia (**QUADRO 1**). Nenhum dos cinco *datasets* do Aleia conseguiu atingir o nível máximo de *FAIRness* e a aderência variou entre 64% a 75%, o que corresponde a um nível moderado de *FAIRness*. Já no Deposita Dados a aderência variou entre 83% a 87%, o que também corresponde a um nível moderado de *FAIRness*. Vale destacar que os níveis estabelecidos pela F-UJI variam entre **incompleto, inicial, moderado e avançado**.

Os melhores resultados do Aleia estão atrelados à **localização** dos *datasets*, onde um deles (UNVCQO) conseguiu atingir nota máxima e todos os demais receberam nota 6,5 de 7, indicando um **nível avançado de localização**. Isso significa que: os *datasets* continham identificadores únicos e persistentes (no caso o *Digital Object Identifier* - DOI); os elementos de metadados essenciais para sua localização estavam presentes; os metadados incluíam o identificador dos dados que descreviam; e os metadados podiam ser recuperados por máquinas. Já no Deposita Dados (**QUADRO 2**), todos os *datasets* receberam a nota máxima, cumprindo todos os requisitos citados. Não há, portanto, melhorias a serem apontadas quanto à localização.

QUADRO 2 - Aderência dos conjuntos de dados do repositório Deposita Dados aos princípios FAIR

URL dos datasets	Código do dataset	Pontuação em localizável	Pontuação em acessível	Pontuação em interoperável	Pontuação em reutilizável	Percentual de aderência geral
https://doi.org/10.48472/deposita/S2HKHC	S2HKHC	7	2	3	8	83%
https://doi.org/10.48472/deposita/HD6DBA	HD6DBA	7	2	3	8	83%
https://doi.org/10.48472/deposita/IGXQF5	IGXQF5	7	2	4	8	87%
https://doi.org/10.48472/deposita/CYIXIF	CYIXIF	7	2	3	8	83%
https://doi.org/10.48472/deposita/CR8RMT	CR8RMT	7	2	3	8	83%
https://doi.org/10.48472/deposita/2KRQZT	2KRQZT	7	2	3	8	83%
https://doi.org/10.48472/deposita/ESD70V	ESD70V	7	2	3	8	83%

Fonte: Elaborado pelos autores (2024).

Quanto à **acessibilidade**, quatro dos cinco *datasets* do Aleia obtiveram pontuação 1 de 3, falhando em cumprir duas métricas: os metadados contêm o nível e as condições de acesso aos dados; e os **dados** são acessíveis por meio de um protocolo de comunicação padronizado. Em compensação, todos pontuaram quanto à terceira e última métrica avaliada: os **metadados** são acessíveis por meio de um protocolo de comunicação padronizado. Já no Deposita Dados, todos os sete *datasets* obtiveram a mesma pontuação: 2 de 3. Assim como no caso do Aleia, todos possuíam **metadados** acessíveis por meio de um protocolo de comunicação padronizado. Somado a isso, os **dados** também eram acessíveis por meio de um protocolo de comunicação padronizado. Para ambos os repositórios é preciso investir na declaração das condições de acesso aos dados, de tal forma que tanto humanos quanto máquinas consigam interpretar a informação.

No que diz respeito à **interoperabilidade**, todos os cinco *datasets* do Aleia obtiveram a mesma pontuação: 3 de 4. Isso significa que os metadados são representados usando uma linguagem formal de representação do conhecimento e incluem *links* entre os dados e suas entidades relacionadas, como artigos, teses, dissertações, *etc.* Em compensação, podem ser implementadas melhorias quanto ao uso de recursos semânticos como ontologias para garantir a interpretação inequívoca dos termos utilizados na representação. O mesmo vale para os *datasets* do Deposita Dados, com exceção do conjunto de dados IGXQF5, que cumpriu todos os requisitos de interoperabilidade (**QUADRO 2**).

Por fim, tem-se a **reutilização**, que é o objetivo final dos princípios FAIR. No Aleia, apenas um *dataset* obteve a nota 6, enquanto os demais obtiveram nota 5 de 10. Alguns requisitos foram cumpridos como: metadados incluem as licenças de reutilização dos dados (a licença é a *Creative Commons*); e metadados incluem informações de proveniência sobre a origem dos dados (exemplos: data de publicação e de modificação). Outros demandam melhorias como: metadados especificarem o conteúdo dos dados (exemplos: tipo e tamanho de arquivo, *etc.*); metadados seguirem um padrão recomendado pela comunidade de

pesquisa alvo; e dados estarem disponíveis em um formato de arquivo recomendado pela comunidade (o que considera formatos abertos e de longo prazo como XML, CSV, JPEG, *etc*). Já no Deposita Dados os resultados foram melhores: todos os sete *datasets* obtiveram a pontuação 8 de 10, o que equivale a um nível moderado de aderência.

Vale destacar que o cenário internacional aponta para uma dificuldade de conformidade quanto à reutilização (Dunning; Smaele; Böhmer, 2017), que foi um ponto fraco do Aleia. Em compensação, a interoperabilidade também é indicada como de difícil aderência pelo estudo, mas os *datasets* do Aleia e do Deposita Dados encontram-se num nível moderado quanto a esse princípio, um bom indicativo.

CONSIDERAÇÕES FINAIS

Sendo assim, é possível observar que os conjuntos de dados dos repositórios Aleia e Deposita Dados, ambos geridos pelo Ibict, encontram-se, no geral, em um nível moderado de *FAIRness*. Entretanto, há algumas diferenças a serem apontadas: os percentuais de conformidade ao FAIR dos *datasets* do Deposita Dados variaram entre 83% e 87%, enquanto os do Aleia variaram entre 64 a 75%. Motivados principalmente pela qualidade dos metadados informados pelos pesquisadores, tanto nos *datasets* como nos arquivos (*files*).

Ambos os repositórios se destacaram quanto à localização: dos 12 *datasets* da amostra, oito atingiram a nota máxima, e os demais receberam nota 6,5 de 7, o que equivale a um nível avançado. A aderência mais baixa se deu na reutilização para o Aleia, seguindo tendências internacionais apontadas por estudos similares (Dunning; Smaele; Böhmer, 2017). Apesar de todos os *datasets* do Deposita Dados receberem nota 8 de 10, os conjuntos de dados do Aleia pontuaram entre 5 e 6, indicando a necessidade de algumas melhorias como: metadados seguindo padrões recomendado pela comunidade de pesquisa e dados disponíveis em formatos abertos e de longo prazo. Quanto à acessibilidade e à interoperabilidade, os conjuntos de dados tiveram um desempenho similar, alcançando níveis moderados de aderência. Pensando nisso, é importante investir no uso de recursos semânticos como ontologias e na declaração formal das condições de acesso aos dados de pesquisa depositados, aumentando assim os níveis de conformidade.

Este estudo foi o ponto inicial para conhecimento e adequação da aderência aos princípios FAIR nos repositórios de dados do Ibict. Novos estudos serão conduzidos para adequação e melhorias dos indicadores. Em trabalhos futuros pretende-se analisar o relatório completo fornecido pela F-UJI para verificar, em detalhes, o desempenho dos *datasets* em cada princípio FAIR, realizar melhorias e capacitar os pesquisadores em seus depósitos.

REFERÊNCIAS

DATAVERSE PROJECT. **About**. 2020. Disponível em: <https://dataverse.org/about>. Acesso em: 27 mar. 2024.

DEVARAJU, A.; HUBER, R. F-UJI: an automated FAIR Data assessment tool. **Zenodo**, Genève, 2020. DOI: <https://doi.org/10.5281/zenodo.4063720>. Disponível em: <https://zenodo.org/records/4063720>. Acesso em: 20 jun. 2024.

DUNNING, A.; SMAELE, M.; BÖHMER, J. **Are The Fair Data Principles Fair?** 2017. Disponível em: <https://zenodo.org/record/321423>. Acesso em: 27 jul. 2024.

SALES, L. F. *et al.* GO FAIR Brazil: A Challenge for Brazilian Data Science. **Data Intelligence**, v. 2, n. 1–2, p. 238–245, 2020. DOI: https://doi.org/10.1162/dint_a_00046. Disponível em: <https://direct.mit.edu/dint/article/2/1-2/238-245/10004>. Acesso em: 27 jul. 2024.

SANCHEZ, F. A.; VIDOTTI, S. A. B. G.; VECHIATO, F. L. A contribuição da curadoria digital em repositórios digitais. **Revista Informação na Sociedade Contemporânea**, [S. l.], v. 1, p. 1–17, 2017. DOI: <https://doi.org/10.21680/2447-0198.2017v1n0ID12280>. Disponível em: <https://periodicos.ufrn.br/informacao/article/view/12280>. Acesso em: 6 jun. 2024.

SAYÃO, L. F.; SALES, L. F. Algumas considerações sobre os repositórios digitais de dados de pesquisa. **Informação & Informação**, [s. l.], v. 21, n. 2, p. 90-115, 2016. DOI: <https://doi.org/10.5433/1981-8920.2016v21n2p90>. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/27939>. Acesso em: 15 abr. 2024.

WILKINSON, M. D *et al.* The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, London, v. 3, p. 1-9, 2016. DOI: <https://doi.org/10.1038/sdata.2016.18>. Disponível em: <https://www.nature.com/articles/sdata201618>. Acesso em: 22 jul. 2024.