Avaliando a conformidade dos conjuntos de dados com os princípios FAIR no Dataverse

Assessing Dataset Compliance with FAIR Principles in Dataverse

Evaluación del cumplimiento del conjunto de datos con los principios FAIR en Dataverse

Rene Faustino Gabriel Junior
Professor adjunto da Universidade Federal do Rio Grande do Sul e do Programa de Pós-Graduação em Ciência da Informação (PPGCIN), da Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brasil.
http://lattes.cnpq.br/5900345665779424
https://orcid.org/0000-0003-1021-3360

Letícia Guarany Bonetti
Pesquisadora bolsista no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, Distrito Federal, Brasil.
http://lattes.cnpq.br/1895977717955732
https://orcid.org/0000-0002-3012-8465

Tatyane Guedes Martins da Silva
Pesquisadora bolsista no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília, Distrito Federal, Brasil.
http://lattes.cnpq.br/7310861285054095
https://orcid.org/0000-0002-1743-0467

Samile Andrea de Souza Vanz
Professora associada do Departamento de Ciências da Informação, da Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brasil.
http://lattes.cnpq.br/5243732207004083
https://orcid.org/0000-0003-0549-4567

Caterina Groposo Pavão
Professora do Departamento de Ciências da Informação, da Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brasil.
http://lattes.cnpq.br/4834791532698069
https://orcid.org/0000-0003-3712-7200

Resumo

Introdução: Neste trabalho busca-se verificar a compatibilidade das ferramentas de avaliação dos princípios FAIR com o *software* Dataverse. **Metodologia:** A pesquisa é caracterizada como descritiva e exploratória. Foi realizado o levantamento de ferramentas de avaliação dos princípios FAIR em um *dataset* com DOI em bases de dados. Para a coleta criou-se um *dataset* no repositório Deposita Dados do Ibict de forma a poder manipular os metadados e, por fim, realizou-se uma análise geral das ferramentas e sua aplicação no Dataverse. **Resultados:** O Dataverse não

Anais do **VII Workshop de Informação, Dados e Tecnologia**, Porto Velho, 25-27 jun. 2024.
https://doi.org/10.22477/vii.widat.191

conseguiu atingir nota máxima, mas não devido a problemas da descrição dos metadados do contexto da pesquisa, mas por utilizar representação sintática (texto), não tendo as propriedades provenientes de ontologias e vocabulários interoperáveis, sugere-se o uso do *Ontology Lookup Service*, *BioPortal* ou *Linked Open Vocabularies*. **Conclusão:** Verifica-se neste trabalho os desafios específicos do Dataverse, como a necessidade de vocabulários controlados e a representação de metadados em formatos legíveis por máquinas. O estudo ressalta a importância de adaptar os metadados às ontologias e aos vocabulários interoperáveis, enfatizando a necessidade de ferramentas que facilitem essa conformidade, contribuindo para a maturidade FAIR dos conjuntos de dados.

Palavras-chave: dados de pesquisa; princípios FAIR; Dataverse.

Abstract

Introduction: *This work seeks to verify the compatibility of the FAIR principles assessment tools with the Dataverse software.* ***Methodology:*** *It is a descriptive and exploratory research. A survey of assessment tools for the FAIR principles was carried out in a dataset with DOI in databases. For gathering, a dataset was created in the Ibict Deposita Data repository in order to manipulate the metadata and, finally, a general analysis of the tools and their application in Dataverse was carried out.* ***Results:*** *Dataverse was unable to achieve maximum scores, but not due to problems describing the metadata of the research context, but because it uses syntactic representation (text), not having properties originating from interoperable ontologies and vocabularies, as suggested by Ontology Lookup Service, BioPortal or Linked Open Vocabularies.* ***Conclusion:*** *This work highlights the specific challenges of Dataverse, such as the need for controlled vocabularies and the representation of metadata in machine-readable formats. The study call attention to the importance of adapting metadata to interoperable ontologies and vocabularies, emphasizing the need for tools that facilitate this compliance, contributing to the FAIR maturity of datasets.*

Keywords: *research data; Fair principles; Dataverse.*

Resumen

Introducción*: Este trabajo busca verificar la compatibilidad de las herramientas de evaluación de los principios FAIR con el software Dataverse.* ***Metodología****: La investigación se caracteriza por ser descriptiva y exploratoria. Se realizó un relevamiento de herramientas de evaluación de los principios FAIR en un conjunto de datos con DOI en bases de datos. Para la recolección se creó un dataset en el repositorio Deposita dados del Ibict con el fin de manipular los metadatos y, finalmente, se realizó un análisis general de las herramientas y su aplicación en Dataverse.* ***Resultados****: Dataverse no pudo alcanzar las calificaciones máximas, pero no por problemas para describir los metadatos del contexto de investigación, sino porque utiliza representación sintáctica (texto), al no tener propiedades provenientes de ontologías y vocabularios interoperables, como sugerencia para el uso de Ontology Lookup Service, BioPortal o Linked Open Vocabularies.* ***Conclusión****: Este trabajo destaca los desafíos específicos de Dataverse, como la necesidad de vocabularios controlados y la representación de metadatos en formatos legibles por máquina. El estudio destaca la importancia de adaptar los metadatos a ontologías y vocabularios interoperables, enfatizando la necesidad de herramientas que faciliten este cumplimiento, contribuyendo a la madurez FAIR de los conjuntos de datos.*

Palabras clave: *datos de investigación; principios FAIR; Dataverse.*

1 INTRODUÇÃO

Desde a introdução dos princípios FAIR em 2016, a relevância e o potencial dessas diretrizes para a ciência impulsionaram sua rápida adoção no meio acadêmico. FAIR são diretrizes destinadas a melhorar a gestão e o uso de dados na era digital, orientando a produção e o compartilhamento de dados de forma que possam ser facilmente acessados, entendidos e utilizados por humanos e máquinas (GOFAIR, 2024). Organizações de pesquisa e pesquisadores, reconhecendo a importância desses princípios, têm buscado aprimorar a gestão de seus dados, alinhando-se às práticas recomendadas de compartilhamento de dados. Com os princípios FAIR servindo como guia, esses esforços visam promover e ampliar a reutilização de dados no âmbito científico (Henning; Ribeiro; da Silva Santos, 2019).

Para avaliar se os repositórios e seus *datasets* atendem os princípios FAIR, foram desenvolvidas algumas ferramentas *open source*. Esses aplicativos funcionam a partir do preenchimento de questionários de autoavaliação ou através da avaliação automática. Muitas pesquisas estão utilizando as ferramentas automáticas para avaliar conjuntos de dados, demonstrando que no Brasil os *datasets* não tem atendido completamente os princípios FAIR (Bonetti; Arakaki, 2022; Groehs *et al.*, 2023).

A aplicação prática dos princípios FAIR necessita avaliar o nível de maturidade e adesão dos dados a esses princípios, o que motivou o desenvolvimento de diversos indicadores de maturidade e *frameworks* de avaliação. Este estudo investiga como é possível satisfazer todos os critérios avaliados pelas ferramentas no Dataverse, considerado por Rocha *et al.* (2021) o melhor *software* para atender as demandas de depósito de dados de pesquisa em análise comparativa com CKAN, DSpace e Invenio.

O objetivo da pesquisa é verificar a compatibilidade das ferramentas de avaliação dos princípios FAIR com o *software* Dataverse. Para atingir esse objetivo,

ela se desdobra nos seguintes objetivos específicos: a) identificar, na literatura, ferramentas *open-source* de avaliação dos princípios FAIR ; b) gerar indicadores avaliativos para um *dataset* nas ferramentas identificadas; c) analisar a compatibilidade das ferramentas de avaliação com o Dataverse. A pesquisa se justifica pelo grande volume de estudos que vêm utilizando essas ferramentas para avaliar a maturidade dos repositórios em relação aos princípios FAIR, apresentando resultados aquém do esperado. Desta forma, percebe-se a importância de compreender como é realizada a avaliação por estas ferramentas e sua compatibilidade com o *software* Dataverse.

2 PROCEDIMENTOS METODOLÓGICOS

Para identificar ferramentas *open-source* de avaliação dos princípios FAIR recorreu-se a uma pesquisa realizada em janeiro de 2024 nas bases de dados Brapci[1], Google Acadêmico, Scopus e Web of Science. No mesmo período, foi criado um conjunto de dados (*Dataset*) que foi submetido como *dataset* no repositório Deposita Dados do Ibict, permitindo a manipulação e análise dos metadados. Para este *dataset* foi atribuído um DOI (10.48472/deposita/ASMJKS). Na primeira fase dos testes foi criado o *dataset* de forma a atender as propriedades mínimas do *DataCite Identifier*, *Creator*, *Title*, *Publisher*, *PublicationYear*, *ResourceType* (DataCite Metadata Working Group, 2024),

Na revisão de literatura identificaram-se seis ferramentas de avaliação, no entanto, dada a limitação de espaço, este trabalho apresenta o relato de uso de duas delas. O uso das ferramentas para avaliação *do Dataset* com DOI foi acontecendo em diversos momentos ao longo dos meses de janeiro e fevereiro de 2024, a partir de várias etapas de identificação, correção de problemas e reavaliação até atingir a máxima pontuação. Para atender o último objetivo de

[1] https://brapci.inf.br

analisar a compatibilidade das ferramentas de avaliação com o Dataverse, fez-se uma análise geral das ferramentas e sua aplicação no Dataverse.

3 RESULTADOS

Na tentativa de cumprir o primeiro objetivo, que consiste em identificar ferramentas de avaliação e maturidade em relação a adesão aos princípios FAIR, constatou-se a divisão dessas ferramentas em duas categorias principais: a primeira abrange ferramentas de autoavaliação, que utilizam questionários de preenchimento próprios para avaliar os princípios FAIR, enquanto a segunda inclui ferramentas que examinam os metadados e produzem uma pontuação indicativa da maturidade.

A revisão de literatura e a organização The Hyve (2023) revelaram seis ferramentas de código aberto desenvolvidas para avaliar a qualidade dos metadados de dados de pesquisa e verificar sua conformidade com os princípios FAIR: ARDC *FAIR Data Self-Assessment Tool*, *FAIR-Checker*, F-UJI e *FAIR Evaluation Services*, FAIR *Aware*, e FAIR *Data Maturity Model*. Todas as seis ferramentas foram analisadas e descritas, no entanto, tendo em vista a limitação de espaço, as ferramentas de autoavaliação, preenchidas pelo próprio usuário, não são apresentadas. Este trabalho apresenta a avaliação do *dataset* pelas ferramentas automatizadas FAIR-Checker e F-UJI.

3.1 Fair-*Checker*

O FAIR-*Checker*[2], uma ferramenta de avaliação automatizada criada pelo Grupo de Trabalho de Interoperabilidade do Instituto Francês de Bioinformática, permite a análise FAIR de recursos digitais, incluindo conjuntos de dados e *softwares* disponíveis *on-line*. Acessível ao público geral através da internet, seu

[2] Disponível em: https://fair-checker.france-bioinformatique.fr/. Acesso em: 20 mar. 2024.

uso é livre de quaisquer procedimentos de registro ou necessidade de permissões especiais.

Ao aplicar a avaliação do aplicativo, a ferramenta avaliou a maturidade do *dataset* do estudo de forma satisfatória, porém apontou algumas inconformidades nos princípios F2A, F2B, I1, I2 e em R1.3 (GOFAIR, 2024), quais sejam:

F2 os dados devem ser descritos com metadados ricos;

I1 os dados devem ser representados em formatos que sigam padrões formais, abertos e acessíveis;

I2 os dados devem usar vocabulários que também sigam padrões formais;

R1.3 os dados devem ser associados a informações detalhadas sobre sua proveniência.

Ao analisar esses resultados de F2A e I1 observou-se que as inconformidades ocorreram nos esquemas de metadados, o qual o Dataverse utiliza o *Dublin Core Terms*, porém com o *namespace* "*dcterms*", enquanto a ferramenta sugere o mesmo padrão, porém chama de "*dct*". A ferramenta também pede o *Data Catalog Vocabulary* (DCAT), o qual o Dataverse não informa em seu esquema de metadados[3].

Quanto às incongruências observadas nos critérios F2B e I2, elas estão ligadas à recomendação de "Expressar todos os seus metadados utilizando propriedades de ontologias e vocabulários interoperáveis". Isso implica que os metadados de qualquer conjunto de dados devem ser articulados por meio de termos (classes, propriedades) oriundos de ontologias e vocabulários padronizados e interoperáveis. Essencialmente, os termos empregados devem ser normatizados, reconhecidos e adotados por comunidades específicas, facilitando o entendimento e a utilização dos dados por pessoas e sistemas automatizados. No contexto do Dataverse, embora seja possível adicionar termos (palavras-chave) e vinculá-los

[3] Disponível em: https://depositadados.ibict.br/api/datasets/export?exporter=dcterms&persistentId=doi%3A10.48472/deposita/ASMJKS Acesso em: 10 mar. 2024.

às URLs de seus respectivos conceitos, a plataforma exige que essas conexões sejam estabelecidas mediante o uso de ferramentas especializadas para a localização das ontologias adequadas, como o LOV (*Linked Open Vocabularies*), *Ontology Lookup Service* (OLS) e BioPortal. No entanto, essa exigência pode não ser plenamente atendida em algumas áreas do conhecimento que não estão integralmente representadas no Dataverse.

A ferramenta só analisa os metadados do *dataset*, não avaliando os metadados dos arquivos e dos dados.

3.2 F-UJI

Desenvolvida pela FAIRsFAIR, a F-UJI é uma ferramenta automatizada e de acesso livre que aplica as métricas FAIRsFAIR para avaliar objetos de dados (Groehs *et al.,* 2023). Ela requer apenas um PID/URL válido da página inicial do conjunto de dados como entrada. Por padrão, a F-UJI utiliza o DataCite para obter metadados em formato JSON a partir do DOI fornecido, embora essa configuração possa ser alterada nas opções do usuário para, em vez disso, utilizar o URL da página de destino. Adicionalmente, há a possibilidade de inserir um serviço de metadados, com a necessidade de definir seu tipo, que pode ser OAI-PMH, OGC CSW ou SPARQL, por exemplo.

Das ferramentas analisadas, a F-UJI demonstrou-se a mais completa, principalmente por fazer a avaliação diretamente ao fornecer um DOI, e por ela analisar não só os indicadores gerais, mas também alguns específicos como proposto no instrumento FAIR *Data Maturity Model*.

A análise da aderência do *dataset* aos princípios FAIR foi verificada a partir de relatórios extraídos da F-UJI. As métricas foram desenvolvidas pelo projeto FAIRsFAIR (Devaraju; Huber, 2020), que possibilitam uma avaliação baseada em 16 das 17 principais métricas, que correspondem a uma parte ou à totalidade de um princípio FAIR (Devaraju *et al*., 2022), traduzidas para português por Groehs *et al. (*2023).

A figura abaixo apresenta o retorno da avaliação desta primeira fase.

Figura 1 - Resultado da primeira rodada de avaliação do *dataset* de teste

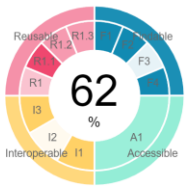
FAIR level:	moderate
Resource PID/URL:	https://doi.org/10.48472/deposita/ASMJKS
DataCite support:	enabled
Metric Version:	metrics_v0.5
Metric Specification:	https://doi.org/10.5281/zenodo.6461229
Software version:	3.1.0
Download assessment results:	{JSON}
Save and share assessment results:	

Fonte: Autores (2024).

Nota-se que, apesar do Dataverse estar em conformidade com os princípios FAIR, alcançou um nível de maturidade FAIR considerado moderado. Demonstrou adequação moderada nos princípios de Encontrabilidade (*Findable*) e Interoperabilidade (*Interoperable*), enquanto os princípios de Acessibilidade (*Accessible*) e Reusabilidade (*Reusable*) foram atendidos de maneira inicial (Figura 2).

Figura 2 - Distribuição dos critérios FAIR do dataset avaliado

Summary:



	Score earned:	Fair level:
Findable:	6 of 7	moderate
Accessible:	1 of 3	initial
Interoperable:	3 of 4	moderate
Reusable:	5 of 10	initial

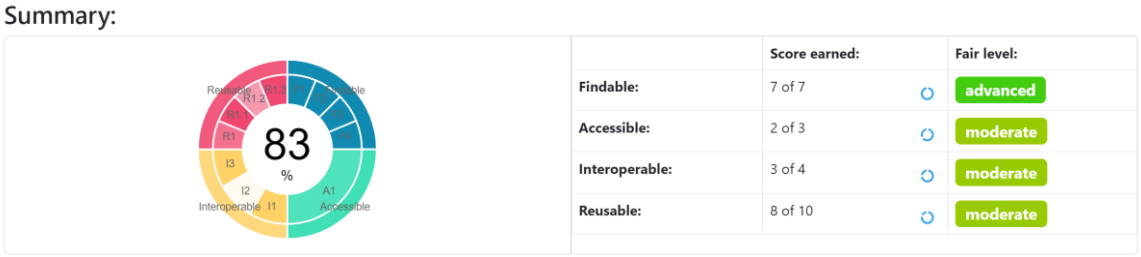
Fonte: Autores (2024).

De forma mais detalhada, observou-se a não conformidade com os princípios:

F3-01M, F3-01M-1, F3-01M-2, I2-01M-1, I2-01M-2b, R1-01MD-2a, R1-01MD-2b, R1-01MD-3, R1-01MD-4, R1.3-02D-1a, R1.3-02D-1b, R1.3-02D-1c. Ou seja, nenhum dos princípios obteve total conformidade.

Ao incorporar outros campos no *dataset* como autores, data da coleta, uri nas palavras-chave e inserção de arquivos com dados no formato .csv, o nível de maturidade aumentou para 83%, o maior percentual atingido nas avaliações. Como mostra a Figura 3.

Figura 3 - Distribuição dos critérios FAIR do dataset avaliado



Fonte: Autores (2024).

Observou-se que os critérios não são contemplados pelo Dataverse possibilitar a inserção de palavras-chave "livros", o sistema exige que os *namespace* devam estar registrados na própria ferramenta, por meio de um vocabulário de recursos semânticos. Então para possibilitar essa avaliação é necessário criar vocabulários controlados semânticos e registrá-los na ferramenta, e ainda impedir que seja registrado palavras-chaves não controladas pelo sistema e que atendam os critérios semânticos, descritos pelas ontologias *Ontology Lookup Service*, BioPortal ou *Linked Open Vocabularies* para encontrar as classes mais adequadas que se deseja usar.

4 Considerações Finais

Este estudo explora ferramentas de avaliação FAIR, identificando duas categorias principais: autoavaliação e análise de metadados. Artigos e ferramentas como ARDC FAIR *Data Self-Assessment Tool*, FAIR-Checker, F-UJI e FAIR *Evaluation Services* são destacados por sua capacidade de avaliar a conformidade com os princípios FAIR. Desafios específicos do Dataverse, como a necessidade de

vocabulários controlados e a representação de metadados em formatos legíveis por máquinas, são abordados. O estudo ressalta a importância de adaptar os metadados às ontologias e aos vocabulários interoperáveis, enfatizando a necessidade de ferramentas que facilitem essa conformidade, contribuindo para a maturidade FAIR dos conjuntos de dados.

Com o estudo pode-se observar que o Dataverse não conseguiu pontuar nota máxima, mas não devido a problemas da descrição dos metadados do contexto da pesquisa, mas por utilizar representação sintática (texto), não tendo as propriedades provenientes de ontologias e vocabulários interoperáveis, como sugestão de uso do *Ontology Lookup Service*, BioPortal ou *Linked Open Vocabularies*. Em continuidade à demanda de aumentar a pontuação do Dataverse, em novas pesquisas busca-se compreender como é possível essa integração a fim de obter a aderência máxima aos princípios FAIR. Para futuros trabalhos, pretende-se ampliar e divulgar a análise das seis ferramentas identificadas na literatura.

REFERÊNCIAS

DATACITE METADATA WORKING GROUP. *DataCite Metadata Schema for the Publication and Citation of Research Data and Other Research Outputs*. Version 4.5. DataCite e.V. DOI: https://doi.org/10.14454/g8e5-6293 Acesso em: 20 fev. 2024.

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

DEVARAJU, A.; HUBER, R. *FAIRsFAIR Data Object Assessment Metrics – FAIRsFAIR*. [*s.l.*], 2020. Disponível em: FAIRsFAIR Data Object Assessment Metrics_v0.4_PublicFeedback - Documentos Google Acesso em: 20 fev. 2024.

GOFAIR. *FAIR principles*. Disponível em: https://www.go-fair.org/fair-principles/ Acesso em: 20 fev. 2024.

HENNING, P. C.; RIBEIRO, C. J. S.; DA SILVA SANTOS, L. O. B.; DOS SANTOS, P. X. GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. *Em Questão*, Porto Alegre, v.

25, n. 2, p. 389–412, 2019. DOI: https://doi.org/10.19132/1808-5245252.389-412.

GROEHS, A.; *et al.* Lattesdata e a adoção aos princípios fair: uma análise usando a F-UJI automated fair data assessment tool. *Em Questão*, Porto Alegre, v. 29, n., 2023. DOI: https://doi.org/10.1590/1808-5245.29.130018

ROCHA, R. P.; *et al*. Análise dos sistemas DSpace e Dataverse para repositórios de dados de pesquisa com acesso aberto. *Revista Brasileira de Biblioteconomia e Documentação*, Brasília, v. 17, p. 1-25, 2021. Disponível em: https://rbbd.febab.org.br/rbbd/article/view/1572 Acesso em: 20 fev. 2024.

THE HYVE. The road to FAIRness: an evaluation of FAIR data assessment tools. Disponível em: https://www.thehyve.nl/articles/evaluation-fair-data-assessment-tools Acesso em: 20 fev. 2024.