WIDaT 2022

REPOSITÓRIOS DE DADOS DE PESQUISA: análise à luz dos princípios FAIR

Letícia Guarany Bonetti – UFSCar
Ana Carolina Simionato Arakaki - UFSCar

FAPESP

ufscar

Introdução

- Avanços das tecnologias levou a um aumento exponencial do volume de dados;
- Os dados, quando compartilhados, trazem benefícios: transparência, reprodutibilidade, economia de recursos e agilização do ciclo científico;
- Não basta compartilhar, é preciso fazer uma gestão dos dados → repositórios;
- A gestão adequada deve ir além do seu armazenamento e acesso, estando ligada às boas práticas internacionalmente adotadas → princípios FAIR (SALES et al., 2020).

Princípios FAIR

- Acrônimo para: encontrável, acessível, interoperável e reutilizável;
- Foram estabelecidos como resultado da conferência internacional *Jointly designing the data FAIRPORT*, de 2014;
- A conferência reuniu especialistas de diversos países e áreas para discutir uma infraestrutura global para publicação, descoberta, compartilhamento e reutilização de dados;
- O repositório é essencial no ecossistema de dados FAIR (HODSON et al., 2018);
- Enfatizar o aprimoramento da capacidade das máquinas de encontrar e processar os dados de forma automática (WILKINSON et al., 2016).

Objetivo

Avaliar o nível de conformidade dos dados de pesquisa depositados nos repositórios institucionais do Estado de São Paulo quanto aos princípios FAIR, em foco a UFSCar, a Unicamp e a USP.

Procedimentos metodológicos

- Pesquisa exploratória, descritiva e quantitativa;
- Amostra: a partir do metabuscador da FAPESP, que reúne repositórios de dados de pesquisa de SP;
- 9 repositórios mapeados → por se tratar de um trabalho derivado da dissertação da autora, ainda em andamento, foram considerados apenas os 3 repositórios analisados até o momento;
- Buscou sistematizar achados e fazer comparações.

Procedimentos metodológicos

- Para a avaliação da aderência aos princípios FAIR foi utilizada uma ferramenta auxiliar: ***F-UJI Automated FAIR Data Assessment Tool***;
  - serviço web para avaliar automaticamente o nível de aderência dos dados;
- A **pontuação geral** varia de 0 a 100% → apresenta uma visão ampla da aderência ao FAIR;
- A **avaliação individual** varia entre inicial, moderado ou avançado → indica o nível da aderência quanto a cada um dos quatro princípios separadamente**;**
- Todos os 199 conjuntos de dados depositados nos 3 repositórios da amostra foram individualmente avaliados.

Dados entregues pela ferramenta F-UJI:

Summary:

Pontuação geral

Pontuação individual

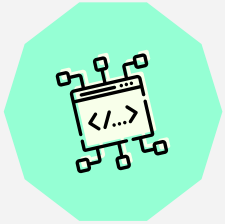
	Score earned:	Fair level:
Findable:	4 of 7	moderate
Accessible:	1.5 of 3	initial
Interoperable:	1 of 4	initial
Reusable:	2 of 10	initial

Report:

Findable

Detalhamento dos aspectos cobertos em cada princípio

- FsF-F1-01D - Data is assigned a globally unique identifier.
- FsF-F1-02D - Data is assigned a persistent identifier.
- FsF-F2-01M - Metadata includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability.
- FsF-F3-01M - Metadata includes the identifier of the data it describes.
- FsF-F4-01M - Metadata is offered in such a way that it can be retrieved programmatically.



01

UFSCar

15 conjuntos de dados

Aderência geral ao FAIR

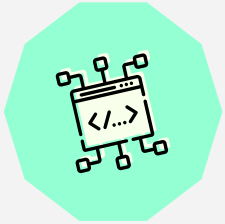
- maior pontuação: 35%;
- menor pontuação : 27%;
- média de pontuações: 30,8%;
- só 3 dos 15 conjuntos de dados conseguiram alcançar a pontuação mais alta (35%);
- falta de aderência geral aos princípios FAIR, com uma certa padronização entre as pontuações alcançadas pelos 15 conjuntos de dados.

Aderência individual

- Dois princípios com maior dificuldade para aderência:
  - interoperável (maior nota: 1/4);
  - reutilizável (maior nota: 4/10);
- Encontrável (maior nota: 4/7);
- Acessível (todos receberam a nota 1,5/3);
- Dunning, Smaele e Böhmer (2017) corroboram esse resultado ao afirmar que, em avaliação feita em 38 repositórios da Holanda, "interoperável" e "reutilizável" foram os princípios mais difíceis de aderir. 38% dos repositórios não possuíam metadados ricos e apenas 41% atribuíam uma licença clara.

Alguns pontos para melhoria:

- Nível de aderência da UFSCar: inicial;
- não foram encontrados identificadores persistentes para todos os conjuntos de dados (encontrável);
    - Juty et al. (2020) afirmam que existem 4 tipos comuns de identificadores FAIR: Digital Object Identifier (DOI), Archival Resource Key (ARK), Identifiers.org e Persistent Uniform Resource Locator (PURL)
- metadados não incluíam um link resolvível para os dados com base em protocolos de comunicação padronizados (acessível);
- necessidade do uso de metadados estruturados como JSON e RDF (interoperável);
- melhorar o nível de especificação do conteúdo dos dados nos metadados → contextualização dos dados (reutilizável);
- declarar nos metadados a licença sob as quais os dados podem ser reutilizados (reutilizável) → precisa ser legível por máquina.



02

Unicamp

67 conjuntos de dados

UNICAMP

Aderência geral ao FAIR

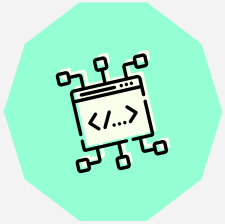
- maior pontuação: 50%;
- menor pontuação : 45%;
- média de pontuações: 49,5%;
- há também uma certa padronização entre as pontuações alcançadas pelos conjuntos de dados;
- percebe-se uma melhoria de resultado quando comparado à UFSCar.

Aderência individual

- Encontrável (todos receberam a nota 6/7);
  - todos possuíam um DOI atribuído;
- Acessível (todos receberam a nota 1/3);
  - Unicamp obteve notas menores que as da UFSCar;
- Interoperável (maior nota: 3/4);
  - diferente da UFSCar, a Unicamp obteve nível avançado para o uso de metadados estruturados;
- Reutilizável (todos receberam a nota 2/10);
  - diferente da UFSCar, oferece informações sobre o **versionamento** dos dados → boa prática incentivada pela W3C.

Alguns pontos para melhoria:

- Nível de aderência da Unicamp: moderado;
- metadados, assim como no caso da UFSCar, não incluíam o identificador dos dados que descreviam (encontrável);
- metadados não incluíam um link resolvível para os dados com base em protocolos de comunicação padronizados (acessível);
- obteve nível inicial quanto aos metadados utilizarem recursos semânticos como *namespaces* (interoperável);
- melhorar o nível de especificação do conteúdo dos dados nos metadados (reutilizável);
- ferramenta F-UJI chegou a localizar os metadados de licença, mas alegou que a representação está incorreta (reutilizável) → precisa ser legível por máquina.



03

USP

117 conjuntos de dados

Aderência geral ao FAIR

- maior pontuação: 25%;
- menor pontuação: 18%;
- média de pontuações: 22,8%;
- em comparação aos dados de pesquisa do repositório da UFSCar e da Unicamp, os conjuntos de dados da USP possuem um nível mais baixo de aderência ao FAIR.

Aderência individual

- Encontrável (maior nota: 3/7);
  - menor nota entre os três repositórios;
- Acessível (todos receberam a nota 1/3);
  - resultado semelhante ao da Unicamp;
- Interoperável (todos receberam a nota 0/4);
  - menor nota entre os três repositórios;
  - a ferramenta até chegou a encontrar URIs de *namespace* nos metadados, mas em nível inicial, o que impediu que fosse pontuado;
- Reutilizável (maior nota 2/10);
  - resultado semelhante ao da Unicamp.

Alguns pontos para melhoria:

- Nível de aderência da USP: inicial;
- não foram encontrados identificadores persistentes para todos os conjuntos de dados (encontrável);
- metadados não incluíam um link resolvível para os dados com base em protocolos de comunicação padronizados (acessível);
- não foram identificados metadados estruturados e nem metadados que incluíssem links entre os dados de pesquisa e suas entidades relacionadas (interoperável);
- assim como no caso da UFSCar/Unicamp, foi identificado um nível inicial de especificação do conteúdo dos dados (reutilizável);
- não foi identificada uma licença declarada em um elemento de metadados apropriado (reutilizável).

Considerações finais

- O escopo deste estudo limita-se aos resultados encontrados a partir da ferramenta F-UJI, trazendo um feedback para as instituições avaliadas;
- Para pontuar, era importante que as informações fornecidas nos repositórios fossem legíveis por máquinas;
- Os níveis de aderência dos dados avaliados ainda se encontram baixos → UFSCar e USP: inicial;
- Maior pontuação geral dos conjuntos de dados foi 50% (Unicamp), já a menor foi 18% (USP);
- Os repositórios têm pontos fracos em comum;
- A reutilização e interoperabilidade são princípios de difícil aderência no contexto dos repositórios da amostra → está em conformidade com o cenário internacional.

Obrigada!

Tem alguma pergunta?

leticia.bonetti@estudante.ufscar.br
acsimionato@ufscar.br