23° enancib
ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO
06 a 10 de novembro de 2023
ARACAJU-SE

GT-8 – Informação e tecnologia

ISSN 2177-3688

DADOS DE PESQUISA EM REPOSITÓRIOS DE SÃO PAULO: AVALIAÇÃO DOS PRINCÍPIOS FAIR

RESEARCH DATA IN SÃO PAULO REPOSITORIES: EVALUATION OF FAIR PRINCIPLES

Letícia Guarany Bonetti - Universidade Federal de São Carlos (UFSCar)
Ana Carolina Simionato Arakaki - Universidade Federal de São Carlos (UFSCar) / Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)

Modalidade: Trabalho Completo

Resumo: Os dados de pesquisa têm um papel de destaque no paradigma da *e-Science*, mas não basta simplesmente depositá-los na *web* para que seus benefícios possam ser extraídos. É preciso que, durante a sua gestão, sejam adotadas boas práticas, e um exemplo de grande relevância no contexto internacional são os princípios FAIR. Com isso em mente, o objetivo deste trabalho foi avaliar o quão alinhados aos princípios FAIR estavam os conjuntos de dados depositados nos repositórios de dados de pesquisa do Estado de São Paulo. A pesquisa se caracteriza como exploratória e descritiva, com abordagem quali-quantitativa. A amostra foi definida a partir do metabuscador de repositórios de dados de pesquisa da Fundação de Amparo à Pesquisa do Estado de São Paulo, e as avaliações dos dados foram feitas com o auxílio da ferramenta automática F-UJI. Como resultado, identificou-se que a aderência dos conjuntos de dados dos repositórios, no geral, foi baixa. A maior pontuação de *FAIRness* foi de 50% de aderência (Unicamp), seguida de 37% de aderência (UFABC) e 35% de aderência (UFSCar). A menor aderência foi encontrada no repositório da Unifesp, onde todos os conjuntos de dados obtiveram 14% de *FAIRness*. A USP e a UNESP obtiveram pontuações parecidas, variando entre 22 a 29% de aderência. Foi possível identificar que a interoperabilidade e a reutilização foram as facetas mais difíceis de aderir. Percebe-se, portanto, que ainda é necessário mais investimento no contexto regional em prol de dados de pesquisa mais FAIR.

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

Abstract: Research data plays a prominent role in the e-Science paradigm, but it is not enough to simply deposit it on the web for its benefits to be extracted. It is necessary that, throughout its management, good practices are adopted, and an example of great relevance in the international context are the FAIR principles. Therefore, the objective of this work was to evaluate how aligned to FAIR principles were the datasets deposited in the research data repositories of the State of São Paulo. The research is characterized as exploratory and descriptive, with a qualitative-quantitative approach. The sample was defined using the research data repository metasearch engine of the Fundação de Amparo à Pesquisa do Estado de São Paulo, and data evaluations were carried out with the help of the automatic tool F-UJI. As a result, it was identified that the adherence of the repositories' datasets, in general, was low. The highest FAIRness score was 50% adherence (Unicamp), followed by 37% adherence (UFABC) and 35% adherence (UFSCar). The lowest adherence was found in the Unifesp repository, where all datasets obtained 14% FAIRness. USP and UNESP obtained similar scores, varying between 22 and 29% adherence. It was possible to identify that, in fact, interoperability and reuse were the most difficult facets to adhere to. It is clear, therefore, that more investment is still needed in the regional context in favor of more FAIR research data.

Keywords: research data; research data repository; FAIR principles.

1 INTRODUÇÃO

Os dados de pesquisa ou científicos, como são denominamos na literatura, têm um papel de destaque na ciência contemporânea, com potencial para acelerar os avanços científicos, trazendo benefícios como economia de recursos, transparência e aumento da visibilidade institucional. O volume intenso de dados (e seu valor competitivo), as tecnologias existentes, o uso intensivo de computação, a colaboração entre cientistas e a preocupação com o acesso leva a um novo paradigma científico, fortemente baseado em dados e conhecido como *e-Science* (Ferreira, 2018). Esse movimento levou a vários debates na comunidade científica sobre padrões e boas práticas, como será abordado neste trabalho.

Guandalini, Furnival e Arakaki (2019) explicam que as instituições de ensino e pesquisa e as agências financiadoras se atentam cada vez mais com as boas práticas ligadas aos dados de pesquisa, que precisam ser devidamente geridos. Os princípios FAIR, por exemplo, publicados originalmente por Wilkinson *et al.* (2016), já são reconhecidos mundialmente como elementos-chave para boas práticas em todos os processos de gestão de dados (Sales *et al*., 2020). Eles permitem aprimorar a capacidade das máquinas de processar os dados automaticamente, aumentando a probabilidade de serem localizáveis, acessíveis, interoperáveis e reutilizáveis na *web*. Com isso em mente, é objetivo deste trabalho avaliar a aderência dos conjuntos de dados depositados nos repositórios de dados de pesquisa do Estado de São Paulo aos princípios FAIR.

2 DADOS DE PESQUISA, REPOSITÓRIOS E PRINCÍPIOS FAIR

Souza e Almeida (2021) argumentam que, de forma isolada, o termo "dado" apresenta um significado restrito e pouco informativo, porém serve como matéria-prima para uma série de observações, medidas ou fatos. Já Borgman, Scharnhorst e Golshan (2019, p. 2, tradução nossa) afirmam que "[...] dados assumem muitas formas e podem se originar de observações, experimentos, minerações, espécimes físicos ou outros métodos". Determinar o que são dados é uma tarefa complexa. Não existe um consenso absoluto sobre a definição do termo, que pode variar de acordo com cada área do conhecimento ou do contexto em que é usado.

Isso também vale para “dados de pesquisa” ou “dados científicos” que, de forma resumida, podem ser definidos como aqueles coletados por pesquisadores durante suas atividades científicas. A *European Commission* (2017), numa descrição mais detalhada, define os dados de pesquisa como informações, em particular fatos ou números, coletados para serem examinados e considerados como base para raciocínio, discussão ou cálculo. Exemplos de dados incluem estatísticas, resultados de experimentos, medições, observações resultantes de trabalho de campo, resultados de pesquisas, gravações de entrevistas e imagens. Todos eles demandam uma gestão adequada para sua preservação e acesso.

No cenário delineado, os repositórios de dados de pesquisa têm um papel importante, auxiliando na representação adequada dos dados por meio de metadados e identificadores. Isso porque não basta só disponibilizá-los na *web*, é preciso contextualizá-los e garantir sua preservação, para que eles sejam encontrados e reutilizados. Rodrigues, Dias e Lourenço (2022, p. 297) definem os repositórios de dados como aqueles que "[...] executam papéis centrais nas infraestruturas do conhecimento como entidades que facilitam o fluxo de dados entre as partes, geralmente ao longo do tempo".

Sayão e Sales (2016) explicam que os repositórios de dados podem ser divididos, de acordo com a literatura, em quatro tipos: institucionais, disciplinares, multidisciplinares e orientados por projetos. Os institucionais, contemplados nesta pesquisa, são aqueles gerenciados e mantidos no âmbito de uma instituição “[...] 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). Mas é importante citar que, como as universidades abrangem um amplo escopo de cursos, de diferentes domínios, os seus repositórios também tendem a ser multidisciplinares, lidando com uma grande variedade de dados de pesquisa. Isso pode dificultar o trabalho das equipes com a consistência da representação dos dados, afetando, inclusive, seus níveis de aderência ao FAIR, que pressupõe o uso de padrões da comunidade e soluções padronizadas.

Os princípios FAIR são um acrônimo para "*Findable", "Accessible", "Interoperable" e "Reusable*", que em português significa "Localizável", "Acessível", "Interoperável" e "Reutilizável”. Eles foram estabelecidos como resultado da conferência internacional *'Jointly designing the data FAIRPORT*' realizada em janeiro de 2014, reunindo especialistas de diversos países e de diferentes áreas de pesquisa para discutir o uso, tratamento e

reutilização de dados de pesquisa no âmbito da *e-Science*. Mas sua disseminação mais ampla começou em março de 2016, com a sua publicação no *Nature's Journal Scientific Data*.

É importante lembrar que o potencial de reutilização dos dados de pesquisa "[...] está fortemente relacionado à adoção de melhores práticas na gestão, na estruturação dos dados para interoperabilidade, no assinalamento de metadados de qualidade, no licenciamento apropriado e na acessibilidade" (Henning *et al*., 2019, p. 176). E, de acordo com Sales *et al*. (2020), o avanço da ciência, em todos os campos do conhecimento, está fortemente ligado à reutilização dos dados de pesquisa.

Logo, avaliar os níveis de aderência desses dados aos princípios FAIR pode trazer *insights* importantes para melhorar sua gestão, tanto por parte das equipes gestoras dos repositórios na escolha de soluções como softwares, identificadores persistentes, padrões de metadados, *etc.*; como por parte dos pesquisadores que fazem o depósito dos dados de pesquisa (em determinado formato) e atribuem os metadados. Somado a isso, as diretrizes propostas pelos princípios FAIR já se constituem como expectativas de agências e editoras internacionalmente (Wilkinson *et al*., 2016). Por isso se torna tão importante buscar sua implementação no cenário nacional, e este estudo busca trazer alguns *feedbacks*.

3 PROCEDIMENTOS METODOLÓGICOS

Para a avaliação dos conjuntos de dados de pesquisa quanto aos princípios FAIR foi realizada uma pesquisa exploratória e descritiva, com abordagem quali-quantitativa. A amostra foi definida a partir do metabuscador de repositórios de dados de pesquisa da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). O metabuscador, em 21 de março de 2022, continha nove repositórios mapeados, mas para a amostra foram considerados apenas os repositórios institucionais diretamente ligados às instituições de ensino superior, sendo elas: Universidade Federal de São Carlos (UFSCar), Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Universidade Estadual de Campinas (Unicamp), Universidade de São Paulo (USP), Universidade Federal do ABC (UFABC) e Universidade Federal de São Paulo (Unifesp).

Excluiu-se, portanto, o Repositório de Dados de Pesquisa da Embrapa (Redape) e o FAPESP COVID-19 Data Sharing/BR, que não são mantidos no âmbito de uma instituição de ensino superior. O repositório de dados de pesquisa do Instituto Tecnológico de Aeronáutica

(ITA) não foi considerado na amostra porque, no momento da pesquisa, continha um total de zero *datasets*, impossibilitando sua avaliação.

A análise dos dados de pesquisa dos repositórios da amostra levou em conta os princípios FAIR, que estão relacionados, mas são independentes entre si. Logo, os conjuntos de dados depositados poderiam se encontrar em diferentes estágios de "*FAIRness*" (WILKINSON *et al.*, 2016). Para a verificação da aderência dos dados de pesquisa aos princípios foi utilizada uma ferramenta auxiliar, a *F-UJI Automated FAIR Data Assessment Tool*[1], desenvolvida pelo projeto *FAIRsFAIR*, que busca alcançar soluções práticas para a adoção dos princípios FAIR ao longo do ciclo de vida dos dados na Europa.

A ferramenta F-UJI é um serviço *web* baseado em *Representational State Transer* (REST) para a avaliação automatizada do "*FAIRness*" dos conjuntos de dados, o que foi essencial para este trabalho, uma vez que foi necessário avaliar 266 *datasets* individualmente. É importante frisar que por se tratar de uma avaliação automática, e não manual, os resultados encontrados estão diretamente ligados aos metadados atribuídos aos dados de pesquisa no repositório. Todas as pontuações/níveis que serão explorados neste trabalho foram entregues pela ferramenta F-UJI, com métricas baseadas nos indicadores propostos pelo *RDA FAIR Data Maturity Model Working Group* e *WDS/RDA Assessment of Data Fitness for use checklist*. Logo, o escopo deste trabalho limitou-se aos resultados encontrados a partir da ferramenta. As métricas, métodos e código podem ser consultados no *site* da F-UJI.

Todos os 266 conjuntos de dados que estavam depositados nos seis repositórios foram individualmente avaliados no período entre janeiro e março de 2022. Para a avalição era necessário apenas inserir (manualmente) o identificador de cada conjunto de dados na F-UJI e ela realizava os testes de acordo com as métricas, devolvendo os resultados conforme as Figuras 1 e 2. Após a avaliação, os dados coletados foram compilados em planilhas[2] *Google* para análises, organizados por repositórios e ordenados de acordo com seus identificadores e códigos. O armazenamento em planilha foi importante porque, por se tratar de uma ferramenta *web* em constante aprimoramento pelos desenvolvedores, os resultados entregues podem mudar ao longo do tempo.

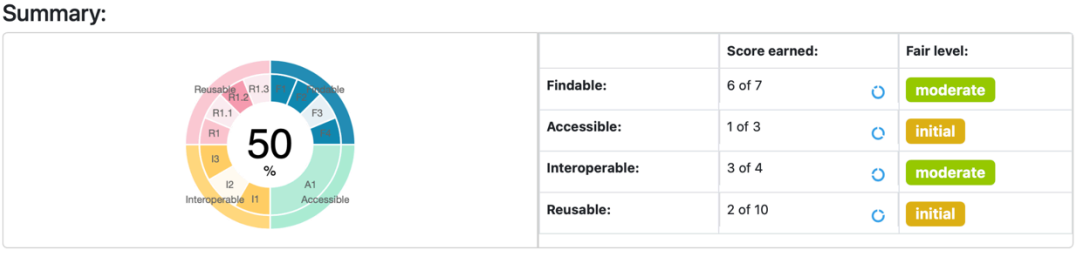
[1] Disponível em: https://www.f-uji.net
[2] Disponível em: https://docs.google.com/spreadsheets/d/1RXKYeWHF-UNSORQhnIDK5Bbl5E3blULTpVhFgHLfKx8/edit?usp=sharing.

A planilha também possibilitou apontar alguns pontos fortes e fracos com relação à aderência ao FAIR. Estes pontos estão ligados tanto às decisões técnicas do repositório (software, identificador persistente, padrão de metadados, protocolo de comunicação padronizado, *etc*.); como às decisões dos pesquisadores que depositam os dados (formato do arquivo, descrição rica com metadados, criação de *links* entre os dados de pesquisa e outras publicações como artigos/livros, *etc*.). Logo, os *feedbacks* levam a um trabalho conjunto daqueles que depositam e daqueles que gerenciam o repositório.

Todos os conjuntos de dados foram igualmente avaliados pela ferramenta F-UJI, que apresentou, na parte de resumo/*summary* (Figura 1), uma porcentagem geral do nível de *FAIRness* de cada *dataset* individualmente, junto com uma nota para localizável, acessível, interoperável e reutilizável. Essas notas (por exemplo: 6 de 7 em localizável) determinam o nível de aderência de cada *dataset* a cada faceta do FAIR, que varia numa escala de "incompleto, inicial, moderado e avançado" e aparece logo ao lado da nota. É possível verificar, em detalhes, a pontuação individual de cada *dataset* em cada faceta do FAIR na planilha disponível na nota de rodapé 2, bem como seu percentual de *FAIRness*.

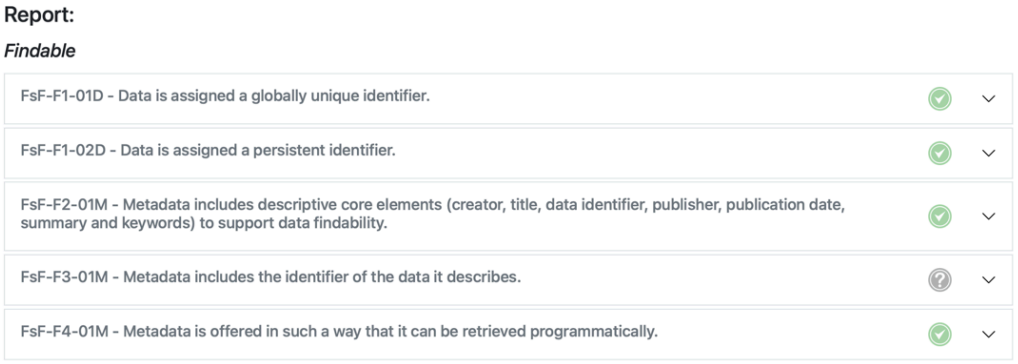
Todas as escalas e notas foram definidas pela ferramenta F-UJI. A nota máxima em localizável é sete, em acessível é três, em interoperável é quatro e em reutilizável é 10. A ferramenta também oferece um relatório/*report* (Figura 2) que mostra em detalhes quais aspectos foram considerados para avaliação em cada faceta (exemplo: em localizável, a ferramenta testa o aspecto FsF-F1-02D - foi atribuído um identificador persistente). O relatório permite entender em quais pontos os conjuntos de dados não foram validados (e perderam nota) e precisam de melhorias para se tornarem mais FAIR.

Figura 1 – Resumo da aderência aos princípios FAIR fornecido pela ferramenta F-UJI



Fonte: captura de tela retirada do site da ferramenta F-UJI (2022).

Figura 2 – Relatório da aderência aos princípios FAIR fornecido pela ferramenta F-UJI



Fonte: captura de tela retirada do site da ferramenta F-UJI (2022).

É importante frisar que algumas recomendações que serão feitas podem não ser aplicáveis ao contexto do repositório e cabe à equipe gestora determinar quais pontos são relevantes para futuras melhorias, pensando nas necessidades da sua coleção e comunidade. Vale lembrar que, por se tratar de auto arquivamento (guiado por manuais institucionais), a equipe gestora pode analisar em quais pontos é preciso melhorias na hora do depósito, passando as recomendações para os pesquisadores nos seus manuais (ou por contato direto caso sejam necessárias mudanças em um depósito já realizado). Como os princípios FAIR pressupõem certas padronizações, as equipes dos repositórios têm um papel essencial elaborando os manuais de depósito e incentivando as boas práticas dentro do seu ambiente digital em busca de maior consistência nas representações.

4 AVALIAÇÃO DOS DADOS DEPOSITADOS QUANTO AOS PRINCÍPIOS FAIR

Como já mencionado, os princípios são recentes e podem gerar equívocos e interpretações ambíguas quando colocados em prática, levando a um nível baixo de aderência ao FAIR (Henning *et al*., 2019). A tendência é que ainda haja vários pontos a serem aperfeiçoados e trabalhados, principalmente com relação à interoperabilidade e à reutilização, conforme indica cenário internacional (Dunning; Smaele; Böhmero, 2017). Ainda mais porque os repositórios da amostra são recentes, ligados a uma demanda da FAPESP de meados de 2017. Isso pode contribuir para estágios iniciais de FAIR.

Além disso, é difícil que dados alcancem 100% de aderência, uma vez que algumas das diretrizes estão abertas à interpretação e ao debate. Mas é importante que as partes interessadas invistam nesses princípios, buscando níveis cada vez maiores de *FAIRness,* já que esse investimento otimiza a capacidade das máquinas de processar os dados automaticamente. No contexto da *e-Science*, essa qualidade é essencial.

4.1 Repositório institucional UFSCar

O repositório institucional da UFSCar[3] optou por um esquema híbrido, ou seja, os *datasets* são depositados no mesmo ambiente que as demais publicações como teses, dissertações, *etc*. O *DSpace* é o software adotado e, no momento da coleta, o repositório continha um total de 15 conjuntos de dados depositados. Nenhum deles conseguiu atingir a nota máxima de *FAIRness*, o que se aplica a todas as quatro facetas do FAIR. A maior nota obtida em localizável foi 4 de 7; em acessível 1,5 de 3; em interoperável 1 de 4 e em reutilizável foi 4 de 10. Ou seja, percebe-se um ponto fraco com relação à interoperabilidade e reutilização, seguindo a tendência internacional (Dunning; Smaele; Böhmero, 2017).

Eles foram também avaliados em uma escala de 0 a 100% quanto à aderência geral ao FAIR, sendo a maior pontuação atingida igual a 35% de *FAIRness* e a média entre os 15 conjuntos de dados igual a 30,6% de aderência. Apenas três *datasets* alcançaram a pontuação mais alta (35%), se destacando na reutilização. Em contrapartida, a menor pontuação alcançada dentre os 15 *datasets* foi de 27% de aderência, verificada para três conjuntos de dados. Em comum, nenhum deles possuía metadados que incluíssem *links* entre os conjuntos de dados e suas entidades relacionadas, prejudicando a interoperabilidade. Num âmbito geral, percebe-se que o percentual de *FAIRness* dos dados de pesquisa da UFSCar é baixo, indo de acordo com o que é defendido por Henning *et al.* (2019).

Algumas melhorias podem ser sugeridas: em "**localizável**", a ferramenta F-UJI não conseguiu recuperar as informações relacionadas ao conteúdo dos dados nos metadados, como nome do arquivo, tamanho e tipo. Isso aponta para uma barreira de processamento por máquina, uma vez que as informações estão presentes no registro dos *datasets*. Além disso, tem-se a questão da atribuição de identificadores persistentes*:* há o uso do *Handle System* para todos os dados de pesquisa depositados, mas a ferramenta não conseguiu validá-los como persistentes, o que é essencial para que os dados sejam encontrados e devidamente citados. Já quando se fala na **acessibilidade**, é importante declarar, nos metadados, o nível e as condições de acesso aos dados. É importante frisar que além da informação ter que ser pública, é preciso que seja legível por máquina, utilizando padrões da

[3] Disponível em: https://repositorio.ufscar.br.

comunidade. Isso vale para todos os repositórios da amostra. A ferramenta também não conseguiu acessar os dados através de um protocolo de comunicação padronizado.

Em **interoperabilidade** pode-se citar alguns pontos a serem trabalhados como: os metadados não são representados usando uma linguagem de representação de conhecimento formal. A ferramenta buscou por metadados estruturados como *JavaScript Object Notation* (JSON-LD) e *Resource Description Framework* (RDFa) no código, mas não encontrou. O repositório também pode investir para aumentar o uso de recursos semânticos nos metadados, como *namespaces*, que garantem que determinado conjunto de objetos tenha nomes exclusivos para que possa ser facilmente identificado.

Por fim, quanto à **reutilização**, é preciso que as informações de licença estejam incluídas nos metadados, de tal forma que sejam legíveis por máquina. Isso permite que os usuários dos dados saibam como podem reutilizá-los, respeitando os direitos autorais. Outro ponto que pode ser trabalhado é o formato do arquivo depositado, que deve seguir padrões da comunidade, preferencialmente abertos para evitar barreiras de software. Alguns exemplos são: *comma-separated values* (CSV), *Portable Network Graphic* (PNG) e *Extensible Markup Language* (XML). Mas isso pode demandar a criação de políticas mandatórias para os usuários, já que é realizado auto arquivamento. Esses pontos citados foram declarados como “incompletos” ou “iniciais” pela F-UJI, cabendo melhorias.

Como pontos fortes pode-se citar: uma boa descrição dos *datasets* (uso dos principais elementos de metadados); metadados são fornecidos de forma que os principais mecanismos de pesquisa conseguem inserí-los em seus catálogos; metadados são acessíveis por meio de protocolos de comunicação padronizados; e metadados incluem as informações de proveniência dos dados (criação e origem). Ou seja, os conjuntos de dados já seguem alguns princípios FAIR. Os esforços são para aumentar esses níveis de *FAIRness*.

4.2 Repositório institucional UNESP

O repositório institucional da UNESP[4] também optou por um esquema híbrido e o software adotado é o *DSpace*, mesmo caso da UFSCar. No momento da coleta o repositório continha um total de 44 conjuntos de dados depositados e nenhum deles conseguiu atingir a nota máxima, o que se aplica a todas as quatro facetas do FAIR. A maior nota obtida em localizável foi 3,5 de 7; em acessível 1,5 de 3; em interoperável 1 de 4 e em reutilizável foi 1

[4] Disponível em: https://repositorio.unesp.br.

de 10. Os *datasets* obtiveram baixas pontuações no geral, mas principalmente quanto à interoperabilidade (em que a maioria recebeu nota 0) e reutilização (em que todos receberam nota 1), seguindo a mesma tendência dos conjuntos de dados do repositório da UFSCar.

Os 44 *datasets* foram também avaliados em uma escala de 0 a 100% de *FAIRness*, e a maior pontuação foi igual a 29%. A média para os 44 *datasets* foi igual a 25,3% de aderência. Apenas três *datasets* alcançaram a pontuação mais alta (29%), e todos os demais receberam a mesma pontuação: 25% de *FAIRness.* Num âmbito geral, percebe-se que o percentual de aderência dos dados de pesquisa é baixo. Os *datasets* que obtiveram os percentuais mais altos foram os que pontuaram em interoperabilidade, recebendo 1 de 4 em vez de 0 de 4.

Algumas melhorias podem ser sugeridas: em "**localizável**", a ferramenta F-UJI não conseguiu recuperar as informações relacionadas ao conteúdo dos dados nos metadados, como nome do arquivo, tamanho e tipo. A ferramenta também indicou um nível inicial de descrição dos dados, ou seja, o repositório pode procurar aumentar o número de elementos de metadados utilizados. A F-UJI elenca os seguintes elementos básicos: *creator, title, data identifier, publisher, publication date, summary e keywords*. Já quando se fala na **acessibilidade**, é importante declarar, nos metadados, o nível e as condições de acesso aos dados. A ferramenta também não conseguiu acessar os dados através de um protocolo de comunicação padronizado.

A **interoperabilidade**, como já visto, exige mais aperfeiçoamentos. Um ponto a ser trabalhado é que os metadados não são representados usando uma linguagem de representação de conhecimento formal. A ferramenta buscou por metadados estruturados como JSON-LD e RDFa no código, mas não encontrou. O repositório também pode investir para aumentar o uso de recursos semânticos nos metadados, como *namespaces*. Somado a isso, é importante que a instituição busque criar *links* entre os *datasets* e suas entidades relacionadas, como artigos, livros, e outros materiais que se originaram a partir dos dados.

Por fim, quanto à **reutilização**, é preciso que as informações de licença estejam incluídas nos metadados, de tal forma que sejam legíveis por máquina. Além disso, os metadados precisam conter informações mínimas sobre o conteúdo dos dados disponíveis, como tipo, formato e tamanho. Elas precisam ser declaradas em metadados apropriados, legíveis por máquina. Outro ponto que pode ser trabalhado é o formato do arquivo, seguindo padrões da comunidade, preferencialmente abertos.

Como pontos fortes a ferramenta F-UJI detectou o uso de identificadores persistentes (*Handle System*); os metadados são fornecidos de forma que os principais mecanismos de pesquisa conseguem inseri-los em seus catálogos; os metadados são acessíveis por meio de protocolos de comunicação padronizados; e os metadados incluem as informações de proveniência dos dados (informações sobre criação e origem).

4.3 Repositório de dados de pesquisa da Unicamp (REDU)

O repositório institucional da Unicamp[5] é destinado exclusivamente a dados de pesquisa e adota o software *Dataverse*. No momento da coleta continha um total de 68 conjuntos de dados depositados e nenhum dos *datasets* conseguiu atingir nota máxima, o que se aplica a todas as facetas do FAIR. A maior nota obtida em localizável foi 6 de 7; em acessível 1 de 3; em interoperável 3 de 4 e em reutilizável foi 2 de 10. Nota-se bons resultados quanto ao "localizável" e "interoperável", mas a reutilização se mostra um ponto fraco novamente.

Os 68 *datasets* foram também avaliados em uma escala de 0 a 100% quanto à aderência, sendo a maior pontuação igual a 50% e a média igual a 49,5%. No total, 62 conjuntos de dados apresentaram uma aderência de 50% ao FAIR, enquanto os outros seis obtiveram um nível de 45% de *FAIRness*. A diferença de pontuação (45% em vez de 50%) se deu devido à interoperabilidade, onde os seis *datasets* receberam nota dois em vez de três.

Algumas melhorias podem ser sugeridas: em "**localizável**" foi verificado que os metadados não incluem o identificador dos dados que descrevem. Para que fosse validado pela ferramenta, era preciso que os metadados possuíssem informações relacionadas ao conteúdo dos dados (nome do arquivo, tamanho, tipo). Novamente, isso aponta para uma barreira no processamento por máquina, uma vez que as informações estão presentes no registro dos *datasets*. Já quando se fala na **acessibilidade**, é importante declarar, nos metadados, o nível e as condições de acesso aos dados. Se existe, por exemplo, alguma restrição. A ferramenta também não conseguiu acessar os dados através de um protocolo de comunicação padronizado.

A **interoperabilidade**, ao contrário do que foi visto até o momento, se mostrou um ponto forte dos *datasets* depositados. A única sugestão é o investimento para aumentar o uso de recursos semânticos nos metadados, como *namespaces*. Mas a Unicamp apresentou

[5] Disponível em: https://redu.unicamp.br.

um ótimo nível de *links* entre os dados e seus recursos relacionados: a F-UJI conseguiu recuperar relações como "*HasPart*" e "*isPartOf*". Por fim, quanto à **reutilização**, é preciso que as informações de licença estejam inseridas nos metadados, de tal forma que sejam legíveis por máquina. Outro ponto que pode ser trabalhado é o formato do arquivo, seguindo padrões da comunidade, preferencialmente abertos.

Como pontos fortes a ferramenta F-UJI detectou o uso de identificadores persistentes; uma descrição rica dos dados de pesquisa; os metadados são fornecidos de forma que os principais mecanismos de pesquisa conseguem inserí-los em seus catálogos; os metadados são acessíveis por meio de protocolos de comunicação padronizados; há o uso de metadados/dados estruturados (RDF e JSON-LD); há *links* entre dados e seus recursos relacionados e os metadados incluem as informações de proveniência dos dados (informações sobre criação e origem). Esses são ótimos indicativos com relação às boas práticas. As melhorias sugeridas são para que níveis cada vez maiores de *FAIRness* sejam alcançados, otimizando a localização, acesso, interoperabilidade e reutilização dos dados de pesquisa.

4.4 Repositório de Dados Científicos da Universidade de São Paulo

O repositório institucional da USP[6] é destinado exclusivamente aos dados de pesquisa e adota o software *DSpace*. No momento da coleta continha um total de 118 conjuntos de dados depositados e nenhum deles conseguiu atingir nota máxima, o que se aplica a todas as facetas do FAIR. A maior nota obtida em localizável foi 3 de 7; em acessível 1 de 3; em interoperável 0 de 4 e em reutilizável foi 2 de 10. Os *datasets* obtiveram baixas pontuações, principalmente quanto à interoperabilidade e reutilização.

Os 118 *datasets* foram também avaliados em uma escala de 0 a 100% quanto à aderência geral aos princípios FAIR, sendo a maior pontuação igual a 25%. Apenas 39 dos 118 *datasets* alcançaram essa pontuação mais alta. Em contrapartida, a menor pontuação alcançada foi de 18% de aderência, verificada para três conjuntos de dados. Em comum, todos eles receberam nota 1 de 10 em reutilizável, enquanto todos os outros conjuntos de dados receberam nota 2. A média de pontuações para os 118 *datasets* foi 22,8% de aderência.

[6] Disponível em: https://repositorio.uspdigital.usp.br.

Algumas melhorias podem ser sugeridas: em "**localizável**", a F-UJI não foi capaz de recuperar identificadores persistentes, apesar do uso do *Handle System* no repositório. Também não conseguiu recuperar as informações relacionadas ao conteúdo dos dados nos metadados, como nome do arquivo, tamanho e tipo. A descrição dos dados de pesquisa também pode ser mais rica porque, de acordo com a F-UJI, os elementos essenciais para a citação dos dados só estavam presentes nos *datasets* que pontuaram com 25% de *FAIRness* (maior nota obtida). Já quando se fala na **acessibilidade**, é importante declarar, nos metadados, o nível e as condições de acesso aos dados. A ferramenta também não conseguiu acessar os dados através de um protocolo de comunicação padronizado. É o mesmo caso dos demais repositórios avaliados.

A **interoperabilidade**, como visto, demanda mais melhorias. A ferramenta F-UJI não conseguiu recuperar *links* entre o *dataset* e suas entidades relacionadas para nenhum dos conjuntos de dados. Além disso, a ferramenta buscou por metadados estruturados como JSON-LD e RDFa no código, mas não encontrou. Por fim, quanto à **reutilização**, é preciso que as informações de licença estejam inseridas nos metadados, de tal forma que sejam legíveis por máquina. Outro ponto que pode ser trabalhado é o formato do arquivo, seguindo padrões da comunidade, preferencialmente abertos.

Alguns pontos fortes podem ser destacados: há uma descrição rica dos dados de pesquisa (para aqueles que alcançaram 25% de *FAIRness*); os metadados são fornecidos de forma que os principais mecanismos de pesquisa conseguem inserí-los em seus catálogos; os metadados são acessíveis por meio de protocolos de comunicação padronizados; e os metadados incluem as informações de proveniência dos dados. Nota-se uma tendência entre os repositórios avaliados, sendo alguns princípios mais fáceis de se atingir.

4.5 Repositório de Dados de Pesquisa da UFABC

O repositório institucional da UFABC[7] é destinado exclusivamente a dados de pesquisa e adota o software *Dataverse*. No momento da coleta continha um total de seis conjuntos de dados e nenhum dos *datasets* conseguiu atingir a nota máxima em localizável, acessível, interoperável ou reutilizável. Todos os *datasets* obtiveram exatamente as mesmas notas: em localizável foi 4 de 7; em acessível 1 de 3; em interoperável 2 de 4 e em reutilizável foi 2 de 10. Novamente a reutilização se mostra como um ponto fraco. Quanto à aderência

[7] Disponível em: https://dataverse.ufabc.edu.br/dataverse/ufabc.

geral aos princípios FAIR, todos os conjuntos de dados obtiveram a mesma pontuação: 37%. Ou seja, há um alto grau de consistência entre os pontos fortes e fracos desses conjuntos de dados.

Algumas melhorias podem ser sugeridas: em "**localizável**", a ferramenta não conseguiu validar o *Digital Object Identifier* (DOI) dos dados de pesquisa, apesar de ele estar atribuído no repositório, o que pode indicar algum erro na sintaxe do identificador. Também não conseguiu recuperar as informações relacionadas ao conteúdo dos dados nos metadados, como nome do arquivo, tamanho e tipo. Quanto à **acessibilidade**: é preciso declarar, nos metadados, o nível e as condições de acesso aos dados. A ferramenta também não conseguiu acessar os dados através de um protocolo de comunicação padronizado.

No que se refere à **interoperabilidade**, os conjuntos de dados da UFABC se destacaram. A única sugestão seria aperfeiçoar o uso de recursos semânticos nos metadados, que foi indicado como nível inicial pela ferramenta. O uso de *namespaces* se mostra, novamente, importante. E já com relação à **reutilização** é preciso que as informações de licença estejam inseridas nos metadados, de tal forma que sejam legíveis por máquina. Outro ponto que pode ser trabalhado é o formato do arquivo, seguindo padrões da comunidade, preferencialmente abertos, como já mencionado outras vezes.

Alguns pontos fortes podem ser destacados: é o segundo melhor resultado de aderência ao FAIR após a Unicamp. Para isso, o repositório conta com uma descrição rica dos dados de pesquisa; os metadados são fornecidos de forma que os principais mecanismos de pesquisa conseguem inseri-los em seus catálogos; os metadados são acessíveis por meio de protocolos de comunicação padronizados; há ligação entre os *datasets* e seus recursos relacionados; os metadados são representados usando uma linguagem de representação de conhecimento formal (este foi um ponto fraco para a maioria dos *datasets*); e os metadados incluem as informações de proveniência dos dados.

4.6 Repositório de Dados de Pesquisa Unifesp

O repositório institucional da Unifesp[8] é destinado exclusivamente a dados de pesquisa e adota o software *Dataverse*. No momento da coleta continha um total de 15 conjuntos de dados depositados e nenhum deles conseguiu atingir nota máxima, o que se aplica a todas as quatro facetas do FAIR. Mas, como já visto, foi exatamente o mesmo caso

[8] Disponível em: https://repositoriodedados.unifesp.br/dataverse/unifesp.

de todos os outros repositórios da amostra. Todos os *datasets* obtiveram exatamente as mesmas notas: em localizável foi 2,5 de 7; em acessível foi 1 de 3; em interoperável foi 0 de 4 e em reutilizável foi 0 de 10. Novamente a reutilização se mostra de difícil aderência, junto com a interoperabilidade.

Já quanto à aderência geral aos princípios, todos obtiveram a mesma pontuação: 14% de *FAIRness*. Ou seja, assim como no caso da UFABC, há um alto grau de consistência entre os pontos fortes e fracos dos conjuntos de dados do repositório. Isso pode estar ligado ao auto arquivamento guiado por instruções da instituição, levando à padronização.

Algumas melhorias podem ser sugeridas: em "**localizável**", a F-UJI não conseguiu validar o DOI dos *datasets*, apesar de ele estar atribuído no repositório, o que pode indicar algum erro na sintaxe do identificador. Além disso, a F-UJI não conseguiu recuperar as informações relacionadas ao conteúdo dos dados nos metadados e é necessário que haja uma descrição mais rica dos *datasets*. Quanto à **acessibilidade**, é preciso declarar, nos metadados, o nível e as condições de acesso aos dados. A ferramenta também não conseguiu acessar os dados através de um protocolo de comunicação padronizado.

A **interoperabilidade** precisa ser aperfeiçoada, uma vez que nenhum dos conjuntos de dados de pesquisa conseguiu pontuar. Ou seja, seria aconselhável investir em melhorias para: 1) fazer uso de recursos semânticos como *namespaces;* 2) criar *links* entre os *datasets* e seus recursos relacionados e, por fim, 3) investir em metadados estruturados e analisáveis (JSON-LD, RDFa) incorporados no código XHTML/HTML da página. Quanto à **reutilização**, também há pontos a serem melhorados: 1) os metadados precisam especificar o conteúdo dos dados (tipo, formato, tamanho do recurso); 2) as informações de licença devem ser inseridas nos metadados, de tal forma que sejam legíveis por máquina; 3) as informações de proveniência (origem) dos dados precisam estar declaradas nos metadados e 4) o formato do arquivo deve seguir padrões da comunidade, preferencialmente abertos.

Alguns pontos fortes podem ser destacados: os metadados são fornecidos de forma que os principais mecanismos de pesquisa conseguem inserí-los em seus catálogos e os metadados são acessíveis por meio de protocolos de comunicação padronizados.

4.7 Comparações

Como visto anteriormente, existem algumas semelhanças e diferenças entre os resultados encontrados para os dados de pesquisa depositados nos seis repositórios da

amostra. Para uma visualização melhor, no Quadro 1 abaixo tem-se as médias (por repositório) das pontuações dos conjuntos de dados de pesquisa para cada faceta do FAIR: localizável, acessível, interoperável e reutilizável. Percebe-se, por exemplo, que o melhor resultado para "localizável" foi no repositório da Unicamp; para "acessível" foi nos repositórios da UFSCar e UNESP; para "interoperável" foi no repositório da Unicamp; e para "reutilizável" foi no repositório da UFSCar.

É importante relembrar que alguns pontos elencados dizem mais respeito à ação direta do repositório (exemplo: atribuir identificadores persistentes aos dados de pesquisa); e outros dizem mais respeito aos pesquisadores (exemplo: optar por um formato de arquivo que segue os padrões da sua comunidade, preferencialmente aberto). Mas, de qualquer forma, é essencial que o *feedback* seja passado para os repositórios que têm contato direto com os depositantes dos dados de pesquisa, indicando as melhorias que podem ser feitas no depósito. Vale também citar que a gestão adequada dos dados de pesquisa nas etapas antes do seu depósito pode contribuir para maiores níveis de *FAIRness*, o que está diretamente ligado aos pesquisadores que coletam e tratam os dados. Eles também podem se guiar por este estudo para se alinhar às boas práticas internacionais.

Quadro 1 – Média das pontuações atingidas para cada faceta do FAIR por repositório

MÉDIA DAS PONTUAÇÕES ATINGIDAS PARA CADA FACETA DO FAIR POR REPOSITÓRIO				
	Localizável (máx. 7)	Acessível (máx. 3)	Interoperável (máx. 4)	Reutilizável (máx. 10)
Repositório institucional da UFSCar (15 *datasets*)	3,2	1,5	0,5	2,3
Repositório institucional da UNESP (44 *datasets*)	3,5	1,5	0,07	1
Repositório institucional da Unicamp (68 *datasets*)	6	1	2,9	2
Repositório institucional da USP (118 *datasets*)	2,7	1	0	2
Repositório institucional da UFABC (6 *datasets*)	4	1	2	2
Repositório institucional da Unifesp (15 *datasets*)	2,5	1	0	0

Fonte: elaborado pelas autoras (2023).

Algumas **semelhanças** podem ser apontadas como: nenhum dos conjuntos de dados de pesquisa conseguiu atingir a pontuação máxima em nenhuma das facetas do FAIR. Em localizável foi possível encontrar pontos em comum como problemas com a validação do identificador persistente. A UFSCar, a USP e a UNESP, por exemplo, usam o *Handle System*, mas a F-UJI só validou a persistência do identificador no caso da UNESP. Já a Unicamp, a Unifesp e a UFABC adotam o DOI, mas ele só foi validado para a Unicamp. Outro ponto comum quanto à localização dos dados de pesquisa é a necessidade de descrições mais ricas nos casos da UNESP, USP e Unifesp. Em acessibilidade também há um ponto em comum: a F-UJI não conseguiu acessar os dados através de um protocolo de comunicação padronizado, o que vale para todos os *datasets* avaliados neste trabalho.

A interoperabilidade foi o ponto fraco da maioria dos dados de pesquisa: é preciso investir em recursos semânticos e em metadados estruturados e analisáveis (JSON-LD, RDFa) incorporados no código XHTML/HTML da página. Reutilização foi outro ponto fraco comum: é preciso que as informações de licença estejam inseridas nos metadados, de tal forma que sejam legíveis por máquina, o que não foi validado para a maioria dos conjuntos de dados (só dois *datasets* da UFSCar pontuaram nesse aspecto). Outro ponto que pode ser trabalhado é o formato do arquivo, seguindo padrões da comunidade, preferencialmente abertos. Isso pode ser solicitado nos manuais de auto arquivamento do repositório, indicando alguns formatos preferenciais como o CSV, XML e PNG.

Algumas **diferenças** também foram observadas: três repositórios adotam o software *DSpace* (UFSCar, USP e UNESP) enquanto três adotam o *Dataverse* (Unicamp, UFABC e Unifesp). Mas, em comum, todos os repositórios adotam o *Dublin Core* (DC). Por este ser o padrão pré-definido do *DSpace*, era um resultado esperado. Vale destacar, entretanto, que os três repositórios *Dataverse* (Unicamp, UFABC e Unifesp) permitem a extração dos metadados em padrões adicionais como o *Data Documentation Initiative* (DDI). Estes dados permitem traçar um perfil inicial das soluções que vêm sendo adotadas quando o assunto são repositórios de dados de pesquisa no contexto de São Paulo.

5 CONSIDERAÇÕES FINAIS

Como já esperado, a aderência dos conjuntos de dados dos repositórios da amostra foi baixa. A maior pontuação geral de *FAIRness* foi de 50% (Unicamp), seguida de 37% (UFABC) e 35% (UFSCar). A menor aderência foi encontrada no repositório da Unifesp, onde

todos os conjuntos de dados obtiveram 14% de *FAIRness*. Os dados de pesquisa da USP e da UNESP obtiveram pontuações mais parecidas, variando entre 22 a 29% de aderência.

Quatro dos seis repositórios são dedicados exclusivamente aos dados de pesquisa, enquanto os outros dois (UFSCar e UNESP) são híbridos: dados e publicações são depositados no mesmo ambiente. Há também uma diferença quanto ao software adotado: três optaram pelo *DSpace* (UFSCar, USP e UNESP) enquanto outros três optaram pelo *Dataverse* (Unicamp, UFABC e Unifesp). Em comum, o padrão de metadados adotado é o *Dublin Core*. Mas os três repositórios *Dataverse* permitem a extração dos metadados em padrões adicionais como o DDI.

Foi possível identificar que, de fato, a interoperabilidade e a reutilização foram as facetas mais difíceis de aderir dentro da amostra. O único repositório que se destacou quanto a uma delas foi o da Unicamp, onde a maioria dos *datasets* obtiveram pontuação 3 de 4 em interoperável. Os dados de pesquisa da Unicamp também se destacaram quanto à sua localização: todos receberam nota 6 de 7. Apesar do ótimo resultado, esses mesmos conjuntos de dados receberam notas baixas em reutilizável, indicando um ponto fraco comum entre todos os seis repositórios avaliados.

A melhor aderência foi em "localizável", onde a média de pontuações para os conjuntos de dados foram: 3,2 (UFSCar); 3,5 (UNESP); 6 (Unicamp); 2,7 (USP); 4 (UFABC) e 2,5 (Unifesp) de 7. Mas um ponto fraco comum foi a falta de validação pela ferramenta dos identificadores persistentes. Outro ponto que pode ser revisado é a declaração nos metadados das condições de acesso aos dados, além das licenças para a reutilização dos dados, em formatos legíveis por máquina. Em "acessível" as notas variaram: em dois repositórios (UFSCar e UNESP) todos os conjuntos de dados obtiveram nota 1,5 de 3. Mas a maioria dos *datasets* obteve pontuação 1 de 3 em acessível (Unicamp, USP, UFABC e Unifesp).

Como já dito, é difícil que os *datasets* alcancem 100% de aderência, ainda mais por estarem depositados em repositórios multidisciplinares, mas o objetivo é alcançar níveis cada vez maiores de *FAIRness*, tornando os dados mais acionáveis por máquina. Vale lembrar que os resultados aqui expostos se restringem ao que foi entregue automaticamente pela ferramenta auxiliar F-UJI, que apresenta suas limitações. Os resultados não podem ser generalizados, posto que a amostra investigada se refere a um pequeno recorte do todo, ou seja, da regionalização dos objetos estudados.

6 AGÊNCIA DE FOMENTO

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Processo: 2021/04469-0.

REFERÊNCIAS

BORGMAN, C. L; SCHARNHORST, A.; GOLSHAN, M. S. Digital data archives as knowledge infrastructures: mediating data sharing and reuse. **Journal of the Association for Information Science and Technology**, v. 70, n. 8, 2019. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/epdf/10.1002/asi.24172. Acesso em: 17 jun. 2022.

DUNNING, A.; SMAELE, M.; BÖHMER, J. **Are The Fair Data Principles Fair**? 2017. Disponível em: https://zenodo.org/record/321423. Acesso em: 27 jul. 2022.

EUROPEAN COMMISSION. **Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020**. 2017. Disponível em: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf. Acesso em: 15 abr. 2022.

FERREIRA, V. B. E-science. *In:* **E-science e políticas públicas para ciência, tecnologia e inovação no Brasil** [online]. Salvador: EDUFBA, 2018, p. 13-30. Disponível em: https://books.scielo.org/id/bc84k/pdf/ferreira-9788523218652.pdf. Acesso em: 10 jan. 2023.

GUANDALINI, C. A.; FURNIVAL, A. C. M.; ARAKAKI, A. C. S. Boas práticas científicas na elaboração de planos de gestão de dados. **RDBCI**: Revista Digital de Biblioteconomia e Ciência da Informação, Campinas, SP, v. 17, 2019. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8655895. Acesso em: 21 jan. 2021.

HENNING, P. C. *et al*. Desmistificando os princípios fair: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados fair. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, v. 14, n. 3, p. 175-192, 2019a. Disponível em: https://brapci.inf.br/index.php/res/v/150613. Acesso em: 20 abr. 2022.

RODRIGUES, M. M.; DIAS, G. A.; LOURENÇO, C. A. Repositórios de dados científicos na américa do sul: uma análise da conformidade com os princípios fair. **Em Questão**, Porto Alegre, v. 28, n. 2, p. 113057, 2022. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/113057. Acesso em: 28 jul. 2022.

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. Disponível em: https://direct.mit.edu/dint/article/2/1-2/238-245/10004. Acesso em: 27 jul. 2022.

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. Disponível em:

http://www.uel.br/revistas/uel/index.php/informacao/article/view/27939. Acesso em: 01 nov. 2019.

SOUZA, M.; ALMEIDA, F. G. O comportamento do termo dado na ciência da informação. **Ciência da Informação em Revista**, v. 8, n. 2, p. 39-54, 2021. Disponível em: https://www.seer.ufal.br/index.php/cir/article/view/11764. Acesso em: 22 fev. 2022

WILKINSON, M. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Sci Data**, n. 3, 2016. Disponível em: https://www.nature.com/articles/sdata201618. Acesso em: 20 jan. 2021.