UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

LETÍCIA GUARANY BONETTI

METADADOS E PRINCÍPIOS FAIR: UM ESTUDO DOS REPOSITÓRIOS DE DADOS DE PESQUISA EM SÃO PAULO

SÃO CARLOS, SP
2023

LETÍCIA GUARANY BONETTI

METADADOS E PRINCÍPIOS FAIR: UM ESTUDO DOS REPOSITÓRIOS DE DADOS DE PESQUISA EM SÃO PAULO

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de São Carlos como requisito parcial para a obtenção título de Mestre em Ciência da Informação.
Área de concentração: Conhecimento, Tecnologia e Inovação
Linha de pesquisa: Tecnologia, Informação e Representação
Orientadora: Profa. Dra. Ana Carolina Simionato
Financiamento: Processo FAPESP 2021/04469-0

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO

grupo de pesquisa dados e metadados

SÃO CARLOS, SP
2023

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Ciência da Informação

Folha de Aprovação

Defesa de Dissertação de Mestrado da candidata Letícia Guarany Bonetti, realizada em 02/03/2023.

Comissão Julgadora: Documento assinado digitalmente gov.br ANA CAROLINA SIMIONATO ARAKAKI Data: 03/03/2023 17:13:01-0300 Verifique em https://verificador.iti.br

Profa. Dra. Ana Carolina Simionato Arakaki (UFSCar)

Profa. Dra. Ariadne Chloe Mary Furnival (UFSCar)

Profa. Dra. Luana Farias Sales Marques (UFRJ)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa d
Pós-Graduação em Ciência da Informação.

AGRADECIMENTOS

Agradeço primeiramente a minha família e amigos, que me apoiaram no meu sonho de seguir a área acadêmica e fazer pós-graduação. Toda essa jornada teria sido muito mais difícil sem os conselhos e incentivos que recebi. Se eu consigo sonhar alto hoje em dia é porque vocês foram a minha base.

Agradeço também a minha orientadora, Professora Ana Carolina, que sempre me deu todo apoio para realizar essa pesquisa e acreditou no meu trabalho desde o começo. Todas as sugestões, conselhos e alterações foram essenciais para essa dissertação. Sua paciência e dedicação me ensinaram muito e me lembraram o quanto anseio pela docência, mesmo com todas as dificuldades que envolvem a profissão no Brasil.

Agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo que acreditou no potencial dessa pesquisa e a financiou durante todos os meses de Mestrado.

Agradeço também a todos os professores incríveis que guiaram meu caminho até aqui, aumentando ainda mais minha paixão pela área acadêmica. Assim como os demais profissionais do Departamento de Ciência da Informação, que me deram todo o suporte necessário.

Agradeço aos amigos que fiz durante esses dois anos de Mestrado, principalmente o Vinicius e a Luiza. Fofocar e desabafar com vocês fez toda essa trajetória mais leve e divertida, mesmo com todos os momentos difíceis.

Agradeço também à banca examinadora, composta pelas professoras Adriadne Chloe Furnival e Luana Sales, duas profissionais que eu admiro e que contribuíram imensamente para a melhora desse trabalho.

Por fim, mas não menos importante, agradeço à Veronica, minha namorada, que esteve comigo desde o momento de nervosismo que me inscrevi no programa de pós da UFSCar até o segundo que finalmente defendi minha dissertação. Ela sabe o quanto essa conquista significa para mim e me apoiou em cada etapa, por mais difícil que fosse.

RESUMO

No contexto da *e-Science*, tem-se uma grande produção de dados por parte dos pesquisadores. Os dados de pesquisa demandam um alto grau de contextualização e precisam de gestão adequada em repositórios, para assim maximizar seus benefícios, como o aumento da visibilidade para o pesquisador e para a instituição. Mas para isso não basta disponibilizá-los na web, é preciso se atentar às boas práticas internacionais que visam aumentar a integridade e o valor dos dados, como o caso dos princípios FAIR. Por meio de uma pesquisa exploratória e descritiva de abordagem quali-quantitativa, este estudo teve como objetivo avaliar o nível de conformidade dos dados de pesquisa depositados nos repositórios institucionais quanto aos princípios FAIR. A amostra é composta pelos repositórios do Estado de São Paulo mapeados no metabuscador da Fundação de Amparo à Pesquisa do Estado de São Paulo. Os dados foram coletados com o auxílio da ferramenta F-UJI, depois compilados em planilhas para análise. Foi possível constatar que as facetas interoperável e reutilizável foram as mais difíceis de aderir, com a maioria dos *datasets* obtendo nota 0 ou 1 de no máximo 4, e nota 1 ou 2 de no máximo 10, respectivamente. A faceta encontrável foi a que teve maior aderência pelos *datasets,* que alcançaram notas como 6, 4 e 3,5 de no máximo 7. Os conjuntos de dados mais bem avaliados estavam depositados em repositórios que utilizam o *software Dataverse* (Unicamp e UFABC), e todos os seis repositórios da amostra utilizavam o *Dublin Core*. No geral, as pontuações recebidas pelos *datasets* da amostra foram baixas, com a maior aderência geral sendo igual a 50% e a menor a 14%. Foram entregues *feedbacks* para todas as seis instituições, que poderão se nortear para implementar melhorias quanto aos princípios FAIR, almejando conjuntos de dados mais encontráveis, acessíveis, interoperáveis e reutilizáveis. A partir dos resultados, foi também possível sugerir um conjunto mínimo de 15 elementos de metadados descritivos para a representação dos dados de pesquisa nos repositórios.

Palavras-chave: Princípios FAIR. *e-Science*. Gestão de dados de pesquisa. Repositórios de dados de pesquisa. Metadados.

ABSTRACT

In the context of e-Science, there is a large production of data by researchers. Research data demand a high degree of contextualization and need proper management in repositories, in order to maximize its benefits, such as increased visibility for the researcher and for an institution. But to do so, it is not enough to make them available on the web, it is necessary to pay attention to international good practices that aim to increase the integrity and value of data, as is the case with the FAIR principles. Through an exploratory and descriptive research with a qualiquantitative approach, this study aims to assess the level of compliance of research data deposited in institutional repositories regards to the FAIR principles. The sample consists of repositories of the State of São Paulo mapped in the metasearch engine of the Fundação de Amparo à Pesquisa do Estado de São Paulo. Data were collected with the help of the F-UJI tool, then compiled into Google Sheets for analysis. It was possible to verify that the interoperable and reusable facets were the most difficult to adhere to, with most datasets obtaining a score of 0 or 1 out of a maximum of 4, and a score of 1 or 2 out of a maximum of 10, respectively. The findable facet was the one with the greatest adherence by the datasets, which reached scores such as 6, 4 and 3.5 out of a maximum of 7. The best evaluated datasets were deposited in repositories using Dataverse (Unicamp e UFABC), and all six repositories in the sample used the Dublin Core. Overall, the scores received by the sample datasets were low, with the highest adherence equal to 50% and the lowest 14%. Feedback was given to all six institutions, which could be guided to implement the FAIR principles, aiming for more findable, accessible, interoperable and reusable datasets. From the results, it was possible to suggest a minimum set of 15 descriptive metadata elements for the representation of research data.

Keywords: FAIR principles. e-Science. Research data management. Research data repositories. Metadata.

SUMÁRIO

1 INTRODUÇÃO

Os avanços relacionados com a inserção das Tecnologias Digitais de Informação e Comunicação (TDICs) levaram a mudanças na forma como é feita a ciência atualmente, com transformações cada vez mais rápidas. As tecnologias disponíveis remodelam o paradigma científico e a comunicação científica, num cenário em que os dados produzidos pelos pesquisadores têm protagonismo.

Bertin, Visoli e Drucker (2017, p. 35) afirmam que os "[...] pesquisadores têm produzido uma quantidade de dados sem precedentes, muitos dos quais são subutilizados ou pouco explorados em seu potencial para o avanço científico e tecnológico". As tecnologias existentes, o uso intensivo de computação, a colaboração entre cientistas, a preocupação com o amplo acesso e o imenso volume de dados produzidos leva a esse novo paradigma científico, fortemente baseado em dados e denominado como *e-Science* (SAYÃO; SALES, 2015). Assim, o compartilhamento e a gestão adequada dos dados de pesquisa ganham um papel de destaque.

Esse movimento em direção ao aumento do compartilhamento de dados levou a vários debates nas comunidades científicas sobre os padrões e as boas práticas, garantindo a qualidade dos repositórios de dados e dos próprios dados depositados (AUSTIN *et al.,* 2016). As discussões e possíveis soluções estão principalmente relacionadas com a melhor forma de lidar com as grandes quantidades de dados e metadados associados em todos os seus formatos. Isso é fundamental para que o benefício máximo possa ser extraído, permitindo, inclusive, o processamento automático por máquinas.

A publicação de dados pode fornecer resultados mensuráveis e citáveis, acelerando assim a mudança de paradigma. Segundo Guandalini, Furnival e Arakaki (2019), nesse cenário, as instituições de ensino e as agências financiadoras se atentam cada vez mais com boas práticas ligadas aos dados de pesquisa. Borgman, Scharnhorst e Golshan (2019) acrescentam que muitas partes interessadas já estão envolvidas nas infraestruturas associadas aos dados de pesquisa como: os acadêmicos e equipes que produzem os dados; as agências de financiamento; as instituições de pesquisa onde as investigações são conduzidas; os formuladores de políticas de pesquisa em organizações públicas e privadas; os usuários atuais e potenciais desses dados e as bibliotecas e arquivos que podem adquirir e administrá-

los. Uma das opções de infraestrutura amplamente utilizadas para o compartilhamento e a gestão dos dados de pesquisa são os repositórios de dados.

Considerando a relevância dos dados de pesquisa no paradigma da *e-Science*, surgiram vários estudos e boas práticas para potencializar os benefícios extraídos de seu compartilhamento. Os princípios FAIR, reconhecidos mundialmente como elementos-chave para boas práticas em todos os processos de gestão de dados, são um exemplo de suma importância no cenário atual (SALES *et al*., 2020). Eles permitem aprimorar a capacidade das máquinas de processar os dados automaticamente, aumentando a probabilidade de serem encontráveis, acessíveis, interoperáveis e reutilizáveis, trazendo diversos benefícios para a ciência.

Entretanto, os princípios FAIR não foram devidamente detalhados e explicados na sua publicação original (WILKINSON *et al.*, 2016), sendo mais estabelecidos como conceitos. Isso pode acabar levando a níveis de aderência baixos devido à dificuldade das instituições de aplicarem os princípios aos seus dados de pesquisa. Portanto, com essa hipótese em mente, tem-se como objetivo investigar e responder a seguinte questão de pesquisa: **o quão alinhados aos princípios FAIR estão os conjuntos de dados depositados nos repositórios institucionais de dados de pesquisa do Estado de São Paulo?**

1.1 Objetivos

O objetivo geral deste estudo é avaliar a aderência dos conjuntos de dados depositados nos repositórios institucionais de dados de pesquisa do Estado de São Paulo quanto aos princípios FAIR. Como objetivos específicos busca-se:

- Avaliar os níveis de aderência geral e individual dos dados de pesquisa depositados nos repositórios quanto aos princípios FAIR;
- Apresentar *feedbacks* para as equipes gestoras dos repositórios de dados de pesquisa;
- Indicar um conjunto mínimo de metadados para a representação dos dados de pesquisa em repositórios;
- Investigar a padronização no esquema de metadados adotado entre os repositórios de dados de pesquisa.

1.2 Justificativa

A proposta de pesquisa é motivada pelos estudos durante a Iniciação Científica e o Trabalho de Conclusão de Curso (BONETTI, 2019), que avaliou serviços de gestão de dados de pesquisa em bibliotecas universitárias brasileiras.

A pesquisa constatou que as bibliotecas universitárias da amostra ainda se encontravam em estágios iniciais quanto à oferta de serviços de gestão de dados de pesquisa, principalmente no que diz respeito aos serviços técnicos, que envolvem o preparo, o depósito e a preservação dos dados de pesquisa nos repositórios institucionais. Ou seja, há a necessidade de maiores estudos que identifiquem os pontos fracos e as melhorias necessárias nos repositórios quanto às boas práticas, permitindo sua otimização.

É importante que haja investimentos nesses repositórios porque seu uso traz diversos benefícios como: preservação digital, créditos ao autor, memória científica e transparência, segurança dos dados, indicação da qualidade e produtividade da instituição, dentre outros (SAYÃO; SALES, 2016). Há, inclusive, um movimento global em torno dos repositórios e de dados de pesquisa, com a criação de diversas diretrizes e estruturas para boas práticas. Alguns exemplos são os princípios FAIR[1], o *Core Trustworthy Data Repositories Requirements*[2], o *The TRUST Principles for digital repositories*[3], o *PLOS "Criteria that Matter"*[4], Boas práticas para dados na web[5] e o *COAR Community Framework for Good Practices in Repositories*[6].

Somado a isso, Araújo (2017) em estudo de tendências e perspectivas da Ciência da Informação apontou a curadoria digital como uma prática que vem sendo amplamente difundida devido à importância da certificação de confiabilidade, da evolução dos formatos e do problema da obsolescência. A curadoria digital está intimamente ligada com a criação e manutenção de repositórios, demonstrando novamente a importância de estudos contemporâneos sobre a temática no campo da Ciência da Informação, buscando aperfeiçoar os serviços atualmente oferecidos.

---

[1] Disponível em: https://www.nature.com/articles/sdata201618.
[2] Disponível em: https://www.coretrustseal.org.
[3] Disponível em: https://www.nature.com/articles/s41597-020-0486-7.
[4] Disponível em: https://theplosblog.plos.org/2019/11/request-for-comments-on-data-repository-selection-criteria-that-matter/.
[5] Disponível em: https://www.w3.org/Translations/DWBP-pt-BR/.
[6] Disponível em: https://www.coar-repositories.org/coar-community-framework-for-good-practices-in-repositories/.

No contexto da *e-Science,* muito se discute sobre a implementação e o impacto dos princípios FAIR no uso e reuso dos dados, mas ainda são poucos os estudos que avaliam a implementação dos princípios em casos concretos (DIAS; ANJOS; RODRIGUES, 2019). Há, portanto, uma lacuna que indica a necessidade de maiores investigações, ainda mais no cenário atual em que financiadores, editores e o próprio governo, de acordo com McQuilton *et al*. (2020), já esperam que os pesquisadores compartilhem seus dados usando padrões em conformidade com o cenário global, além de armazená-los em repositórios sustentáveis.

É importante ressaltar ainda que vários eventos e publicações já indicam a importância e atualidade do tema dentro da rede internacional de pesquisa: o *FAIR Festival 2021* realizado pela *GO FAIR*, iniciativa apoiada pela Europa, Estados Unidos, Austrália e África; o *Open Repositories 2021*, que tem membros de instituições como *Indiana University, British Library, DataCite, University of California* e do Instituto Brasileiro de Informação em Ciência e Tecnologia no comitê permanente; e o *The Dataverse Community Meeting*, organizado pelo *Dataverse Project* filiado à Universidade de *Harvard*. Assim, é essencial que o Brasil se alinhe com esse cenário.

Sendo assim, o presente estudo contribui no desenvolvimento de novos conhecimentos acerca dos estágios de conformidade dos dados de pesquisa depositados em repositórios regionais com relação aos princípios FAIR. Assim é possível trazer *feedbacks* para as instituições, sugerindo melhorias para a otimização dos repositórios e atuando em prol de uma ciência mais eficaz, que atende "[...] de maneira ágil a demandas específicas, decorrentes das definições políticas, governamentais e de outras implementações" (HENNING *et al*., 2019b, p. 403). Além disso, indica um conjunto mínimo de metadados descritivos para a representação dos dados de pesquisa em repositórios, buscando auxiliar na criação de boas descrições para esses recursos.

Justifica-se no contexto atual e internacional de busca por garantia das boas práticas em repositórios, uma vez que infraestruturas que assegurem o máximo de confiabilidade, interoperabilidade e acessibilidade aos dados de pesquisa possibilitam que eles sejam encontrados, reutilizados e citados (MCQUILTON *et al.*, 2020), dando maior visibilidade para a produção nacional e aumentando seu prestígio. Esse processo também leva à maior agilidade no ciclo da ciência, além da economia de recursos e do retorno social.

Por isso, estudos mais aprofundados nessa temática, que ainda se encontram em estágios iniciais no Brasil quanto à aplicação (DIAS; ANJOS; RODRIGUES, 2019), apresentam importantes contribuições para a sociedade e a inovação. Os repositórios são infraestruturas tecnológicas de grande importância dentro do cenário da *e-Science*, fortemente baseada em dados, buscando a aproximação da sociedade e da ciência, com maior transparência e integridade.

1.3 Estrutura do trabalho

A dissertação apresenta, além do presente capítulo correspondente à introdução, com justificativa e objetivos, a seguinte estruturação:

O **Capítulo 2 – E-SCIENCE E DADOS DE PESQUISA** define e apresenta uma contextualização da relação entre o paradigma da *e-Sience* e os dados de pesquisa, que para serem devidamente geridos demandam mais que apenas seu depósito em repositórios, mas também a atenção às boas práticas internacionalmente adotadas.

O **Capítulo 3 – REPOSITÓRIOS DE DADOS E BOAS PRÁTICAS** define os repositórios de dados de pesquisa e suas características, além da sua relação com os metadados e boas práticas, mais precisamente os princípios FAIR.

O **Capítulo 4 – PROCEDIMENTOS METODOLÓGICOS** oferece uma descrição detalhada dos procedimentos realizados para a execução deste trabalho, como a escolha da amostra e as etapas.

O **Capítulo 5 – METADADOS, REPOSITÓRIOS DE DADOS E PRINCÍPIOS FAIR** apresenta os dados coletados por meio da ferramenta F-UJI e trás as análises dos estágios de conformidade com os princípios FAIR de cada repositório da amostra e seus respectivos conjuntos de dados. Além disso, indica um conjunto mínimo de metadados descritivos para a representação de dados de pesquisa em repositórios.

O **Capítulo 6 – CONCLUSÃO** apresenta as conclusões do estudo, fazendo uma breve contextualização e sintetizando os resultados obtidos no trabalho.

2 E-SCIENCE E DADOS DE PESQUISA

2.1 *e-Science*

A forma como se faz ciência modifica-se constantemente devido às transformações cada vez mais rápidas ocorridas no cenário atual (SILVEIRA *et al.*, 2021). As tecnologias remodelam os métodos científicos, com a produção e uso intensivo de dados num ambiente de ampla colaboração que ficou conhecido como *e-Science* (CÓRDULA; ARAÚJO, 2019). Há ainda uma dificuldade formal de se definir o termo devido à falta de consenso no meio científico, mas uma boa forma de compreender o que a *e-Science* representa é através da sua origem.

O termo foi introduzido por John Taylor, diretor geral do Conselho de Pesquisa do Gabinete de Ciência e Tecnologia - *Office of Science and Technology* (OST) do Reino Unido, em 1999. A partir de sua experiência anterior como chefe de laboratórios de pesquisa europeus e de sua experiência em seu cargo atual, Taylor percebeu que muitas áreas estavam se tornando cada vez mais dependentes de novas formas de trabalho colaborativo e multidisciplinar (HEY; TREFETHEN, 2003), tirando os pesquisadores do isolamento. E, de fato, é possível observar atualmente o papel essencial dos grupos de pesquisa para a ciência.

Quando combinado a um conjunto de termos, a partícula “e” implica uma transformação para as redes *on-line* e a utilização das tecnologias de informação (TI), o que também engloba a ciência (KOSCHTIAL, 2021). Ferreira (2018) acrescenta que o “e” começou significando *eletronic* (eletrônico), mas atualmente representa melhor *enhanced* (melhorada) ou *enabled* (habilitada). O conceito que é discutido principalmente na Alemanha e na Grã-Bretanha sob o termo *e-Science* corresponde nos Estados Unidos ao conceito de infraestruturas cibernéticas e na Austrália ao conceito de *e-research* (KOSCHTIAL, 2021).

Ferreira (2018, p. 13) afirma que “A necessidade de uma tecnologia adicional que suportasse o desenvolvimento das pesquisas e aliviasse o isolamento do pesquisador foi o impulso extra para o aparecimento da *e-Science*”. A *e-Science* trata-se, portanto, da colaboração global em áreas-chave da ciência e da geração de infraestrutura que a possibilitará. Mustafee *et al.* (2019) acrescentam que requer a utilização de quantidades incomuns de recursos de computação e conjuntos de dados massivos para realizar pesquisas científicas, com dados que podem exigir análise de

especialistas pertencentes a várias organizações e especialistas em diferentes domínios de conhecimento.

A *e-Science* é, então, um paradigma da ciência e, segundo Córdula e Araújo (2019), está atrelado a outros termos como ciência orientada para dados, computação fortemente orientada para dados, E-infraestrutura e ciberinfraestrutura. Esse último sendo, para alguns autores, o meio para se atingir a *e-Science*. Já Ferreira (2018) afirma que a *e-Science* resgata outros termos que foram percussores do contexto atual da ciência, sendo eles: *Big Science*, *Cyberscience*, *Big Data*, *Open Data* e o *Open Science.* Independente do termo utilizado, é importante salientar que a *e-Science* origina-se num cenário de políticas governamentais do Reino Unido, buscando "[...] fortalecer o desenvolvimento da alta computação e da tecnologia de informação e comunicação" (FERREIRA, 2018, p. 18).

Mustafee *et al*. (2019, p. 1, tradução nossa) explicam que para o crescimento da *e-Science* é "[...] fundamental o conjunto integrado de tecnologias conhecidas como e-Infraestruturas ou ciberinfraestruturas (termos que surgiram simultaneamente na Europa e na América do Norte no final dos anos 2000)". Isso inclui redes de comunicação de pesquisa de alta velocidade, recursos computacionais poderosos (computadores de alto desempenho, *clusters*), *grid*, tecnologia de nuvem, infraestruturas de dados que fornecem acesso a fontes de dados, sensores e suporte para diferentes formas de acesso.

O potencial da *e-Science* se expande, redesenhando tarefas, ligada ao surgimento de organizações virtuais e à importância cada vez maior da colaboração. Ela é fortemente relacionada a uma dimensão tecnológica, sendo o ponto em que a TI encontra os cientistas, e tem passado por mudanças e amadurecimento nos últimos tempos (KOSCHTIAL, 2021). Em seu "[...] bojo identifica-se a preocupação crescente pela captura, curadoria e análise dos dados" (FERREIRA, 2018, p. 13) e, hoje em dia, já existem várias iniciativas internacionais de projetos *e-Science* que são, inclusive, fomentados por agências e conselhos como a *National Science Foundation* (NSF), nos Estados Unidos; o *Research Councils UK* (RCUK), no Reino Unido; e a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), no Brasil. Há ainda os centros de *e-Science* na Holanda, Suécia e Austrália (FERREIRA, 2018). Nota-se, portanto, que a *e-Science* já é uma realidade e traz um grande potencial para a ciência, principalmente quando se pensa no grande volume de dados a serem processados.

Ainda segundo Ferreira (2018), a *e-Science* surge justamente dessa necessidade de enfrentar o “dilúvio de dados”, demandando o uso de computação intensiva. A infraestrutura da *e-Science* destina-se a capacitar e auxiliar os pesquisadores a realizarem suas pesquisas de modo mais ágil e eficiente, superando o obstáculo do grande volume de dados, e a Ciência da Informação também se insere nesse contexto. Sales e Sayão (2015, p. 31) explicam que:

> A inserção da biblioteca de pesquisa nos ambientes virtuais que permeiam as metodologias e as dinâmicas de comunicação da *eScience* pressupõe não somente a disponibilização de novos serviços de informação, mas uma parceria mais sofisticada, que compreendem intervenções nos diversos estágios de processamento dos dados de pesquisa – no planejamento, na formação de coleções, no controle de qualidade, na visualização e no apoio aos novos modelos de publicação acadêmica e de comunicação científica.

Ainda de acordo com os autores, a curadoria dos dados de pesquisa parece ser uma extensão natural das funções da biblioteca, uma vez que eles representam ativos competitivos na ciência e são fruto dos estudos realizados pelos pesquisadores da instituição. Entretanto, a maioria das bibliotecas e dos bibliotecários “[...] são menos familiares com as fases iniciais do ciclo de comunicação científica – por exemplo, com as fases de concepção, modelagem e planejamento – e estão mais próximos com as atividades de pós-pesquisa, como de reportar, comunicar e publicar” (SALES; SAYÃO, 2015, p. 42). Isso exige e exigirá dos bibliotecários, principalmente aqueles ligados às bibliotecas universitárias, o desenvolvimento de novas habilidades ligadas à gestão dos dados, auxiliando os pesquisadores (REIS; CENAS, 2021).

Será preciso, ainda, que exista uma colaboração próxima com grupos de pesquisa, testando e difundindo padrões, tecnologias e boas práticas na gestão dos dados de pesquisa, essenciais no paradigma da *e-Science*. Nesse cenário, inclusive, “[...] novas ferramentas e outras abordagens sobre a geração e preservação de dados são inseridas e, consequentemente, surge a Biblioteconomia de Dados (*e-Science Librarianship*)” (REIS; CENAS, 2021, p. 54). De acordo com os autores, é uma discussão que vem sendo realizada desde 1960 na América do Norte e Europa, enfatizando a criação de serviços ligados aos dados, repositórios e à própria gestão dos dados de pesquisa.

Wilkinson *et al.* (2016) explicam que a e-*Science* contemporânea exige que os dados sejam cada vez mais encontráveis, acessíveis, interoperáveis e reutilizáveis a

longo prazo, objetivos esses que estão rapidamente se tornando expectativas de agências e editores, como já visto. Pensar no processamento automático por máquinas é, então essencial, uma vez que a quantidade massiva de dados cria essa demanda. Mas para que os dados de pesquisa sejam legíveis por máquina, se alinhando à *e-Science*, é preciso um tratamento adequado deles, pensando em boas práticas, e os profissionais da informação têm um papel importante nisso. Sendo assim, considerando o que foi exposto, é de suma importância que os profissionais entendam o que são esses dados, para então realizar o tratamento adequado.

2.2 Dados de pesquisa

O conceito de "dado" dentro da Ciência da Informação ainda não encontra uma uniformidade, "[...] justamente por ser uma área relativamente nova e interdisciplinar, que recebe diferentes contribuições em seu arcabouço teórico" (SOUZA; ALMEIDA, 2021, p. 40). Pomerantz (2015) afirma que dado é apenas informação em potencial, bruto e não processado, antes de alguém poder se informar por meio dele. Traz em seu livro a tríade dado-informação-conhecimento, em que o dado se apresenta como o menor elemento, o que é coletado por instrumentos ou mecanismos para então ser processado e transformado em informação.

Em um mesmo sentido 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. Sayão e Sales (2020, p. 32), em concordância, o descrevem como uma “[...] dádiva, oferta ou algo reconhecido e usado como base para cálculos, é o vínculo primordial com os fenômenos de amplo espectro em que estamos imersos". Nota-se que são definições mais amplas e abstratas.

Sob uma perspectiva mais restritiva, Amaral (2016, p. 3) afirma que “[...] dados são fatos coletados e normalmente armazenados”. Borgman, Scharnhorst e Golshan (2019, p. 2, tradução nossa) corroboram ao dizer que "[...] dados assumem muitas formas e podem se originar de observações, experimentos, escavações, espécimes físicos ou outros métodos". Logo, determinar o que são dados é uma tarefa complexa. Não existe um consenso absoluto sobre a definição do termo "dado", podendo variar de acordo com cada área do conhecimento ou do contexto em que é usado.

Entretanto, Santos e Sant'Ana (2013) argumentam que é importante que o dado seja compreendido como elemento básico nos fluxos informacionais, mesmo que o termo e o dado em si sejam usados de formas diferentes por cada área do conhecimento. Os autores trazem ainda a diferença entre dados estruturados e não estruturados, sendo os estruturados aqueles que possuem estrutura interna explícita (como um registro bibliográfico), e os não estruturados aqueles que dependem de um interpretador externo (como um livro digital). Santos e Sant'Ana (2013, p. 205) concluem que:

> [...] dado é uma unidade de conteúdo necessariamente relacionada a determinado contexto e composta pela tríade entidade, atributo e valor, de tal forma que, mesmo que não esteja explícito o detalhamento sobre contexto do conteúdo, ele deverá estar disponível de modo implícito no utilizador, permitindo, portanto, sua plena interpretação.

Percebe-se uma ligação intrínseca com os metadados, que possibilitam essa plena interpretação. Na Ciência da Informação, quando se fala sobre representação de recursos, os metadados têm essa função de destaque, descrevendo e identificando os recursos no sistema de informação. Arakaki e Arakaki (2020, p. 37) definem metadados como "[...] uma informação estruturada para as ações de identificação, descoberta, seleção, uso, acesso e gerenciamento.". Eles fornecem contexto, identificam relações com outros recursos do sistema e garantem pontos de acesso para os usuários.

Arakaki e Arakaki (2020, p. 43), trabalhando a relação entre dados e metadados, explicam que "[...] a depender da interpretação e do contexto que são tratados, os dados podem ser entendidos como metadados ou documento". Mas sabe-se que os metadados exercem um papel importante no cenário da recuperação dos dados em sistemas, facilitando a descoberta, garantindo a preservação e agregando valor aos mais diversos tipos de dados.

Atualmente fala-se muito sobre um tipo mais específico: o dado de pesquisa, ou dado científico, que são, de maneira simplificada, aqueles coletados por pesquisadores durante suas atividades científicas. A ciência contemporânea restituiu-os ao seu protagonismo histórico (SALES *et al.*, 2020), sendo considerados como a moeda mais valiosa da ciência (DAVIS; VICKERY, 2007).

Os dados de pesquisa, segundo a *Organisation for Economic Co-operation and Development* (OECD, 2007, p. 13, tradução nossa), podem ser definidos como "[...]

registros factuais usados como fontes primárias na pesquisa científica, e que são geralmente aceitos na comunidade científica como sendo necessários para validar os resultados de pesquisa". Ou seja, há uma relação com o embasamento da pesquisa, a possibilidade de reprodutibilidade e de transparência científica.

Já a *European Commission* (2017) 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. No mesmo sentido, Sayão e Sales (2020, p. 32) afirmam que os dados de pesquisa:

> [...] podem ser dados brutos coletados diretamente por um instrumento ou um sensor e agregados a partir de múltiplas fontes; ou podem ser produtos de um modelo teórico, simulação ou visualização; ou de experimentos conduzidos na bancada de um laboratório; ou ainda podem ser textos, bibliotecas de imagens digitais e modelos em 3D, tais como os usados para a reconstrução de sítios históricos e mitológicos.

Borgman (2015, p. 29, tradução nossa) também oferece uma definição para os dados de pesquisa como “[...] entidades usadas como evidência de fenômenos para fins de pesquisa ou erudição”. Costa e Leite (2018, p. 89) trazem uma contribuição sintetizada, definindo-os como aqueles dados “[...] produzidos e/ou utilizados para o desenvolvimento de uma pesquisa”, o que, como visto, pode incluir números, imagens, textos, vídeos, áudio, *software*, algoritmos, equações, modelos, simulações, observações, dentre outros, dependendo da área de conhecimento.

Segundo Sayão e Sales (2015, p. 7), os dados de pesquisa "[...] podem ser caracterizados de várias formas, por exemplo, de acordo com sua natureza, origem ou de acordo com seu status no fluxo de trabalho da pesquisa". Quanto à sua origem eles foram categorizados pelo *National Science Board* da *National Science Foundation* como experimentais, computacionais e observacionais, e têm um papel fundamental nas decisões operacionais sobre que conjuntos de dados devem ser preservados e por qual período em repositórios (infraestruturas estas que serão foco deste estudo).

Quadro 1 – Categorias dos dados segundo a origem

CATEGORIAS SEGUNDO A ORIGEM (***NATIONAL SCIENCE BOARD***)			
	Dados experimentais	**Dados computacionais**	**Dados observacionais**
Conceito	São dados obtidos por meio de observações diretas, que podem ser associadas a lugares e tempo específicos	São resultados da execução de modelos computacionais ou de simulações	São provenientes de situações controladas em bancadas de laboratórios
Exemplos	Desempenho de um motor ou taxas de reação química	Resultados da execução de um modelo de computador ou simulação	Observações diretas da temperatura do oceano em uma data específica
Preservação	**Passíveis de reprodução com precisão**: não precisam ser armazenados de maneira permanente **Não passíveis de reprodução com precisão**: preservação a longo prazo deve ser garantida	Se houver informações abrangentes sobre o modelo, a preservação em um repositório de longo prazo pode não ser necessária, pois podem ser reproduzidos	Arquivados de maneira permanente por se tratar de dados de observações únicas

Fonte: Adaptado de *National Science Board* (2005).

Nota-se por essas categorias que os dados de pesquisa variam de acordo com os pesquisadores que os produzem e os domínios de conhecimento, possuindo características diversas e sendo classificados de diferentes formas. Sayão e Sales (2020, p. 32) defendem que "[...] dado de pesquisa permanece um conceito ambíguo, exigindo que os arquivos, repositórios, centros de dados, se adaptem às novas formas de dados na medida em que eles aparecem".

Há essa noção de heterogeneidade que demanda dos repositórios um planejamento detalhado para que os dados sejam corretamente representados. “O reconhecimento dessa idiossincrasia torna-se crucial quando se estabelecem as opções gerenciais e tecnológicas para o arquivamento persistente e para a curadoria digital” (SAYÃO; SALES, 2014, p. 81). Ou seja, tudo isso precisa ser levado em conta no contexto de repositórios de dados, definindo, por meio de políticas, o que deve ser preservado ou não, além da necessidade constante de atualização por parte dos gestores dos repositórios e dos pesquisadores, levando em conta o caráter abrangente do termo "dado". Sayão e Sales (2020, p. 32) acrescentam que:

> Essa diversidade que vai sendo delineada pelas especificidades de cada disciplina, suas condicionantes metodológicas, protocolos *workflows* e seus objetivos, se torna um desafio, mesmo para o pesquisador, pelo alto grau de contextualização necessário.

Ainda é preciso levar em conta que o termo "dado" evolui rapidamente junto com as tendências sociais e técnicas, não existindo uma uniformidade para as ciências. É importante, nesse contexto, aproximar o ciclo de vida da pesquisa ao ciclo de vida dos dados. Isso porque as particularidades desses dados precisam ser levadas em conta no momento da gestão, principalmente em repositórios multidisciplinares, como é o caso dos repositórios de dados foco deste estudo.

A gestão de dados de pesquisa em repositórios, conforme apontam Tartarotti, Dal'Evedove e Fujita (2019), é hoje um dos principais recursos para o avanço científico, e deve ir além do armazenamento seguro e do acesso a esses dados; deve estar fortemente relacionado à adoção de boas práticas ao longo do ciclo de vida da pesquisa (SALES *et al.*, 2020).

> A gestão eficaz se dá por meio da estruturação adequada dos dados, da adoção de metadados de qualidade, do potencial de interoperabilidade entre eles, do atendimento a questões legais e éticas (por exemplo, licenciamento adequado) e outras práticas fortemente associadas ao que foi designado pela comunidade científica como os princípios FAIR (SALES *et al*., 2020, p. 239, tradução nossa).

A gestão dos dados de pesquisa se mostra ainda mais fundamental quando se pensa que muitos dos dados coletados por pesquisadores não podem ser substituídos em casos de perda ou destruição como, por exemplo, pela obsolescência das mídias. Eles são únicos e por isso seu depósito e preservação são o que garantem que eles serão acessados e reutilizados. Além disso, Austin *et al.* (2016) afirmam que há uma preocupação crescente em torno da reprodutibilidade da pesquisa publicada, ou seja, é importante que os resultados possam ser reproduzidos usando os dados, códigos, ferramentas e métodos empregados pelo pesquisador, garantindo a transparência da pesquisa e a economia de recursos.

Conforme explicam Araújo, Dias e Autran (2021), a gestão e compartilhamento dos dados que subsidiam as pesquisas é uma questão de suma importância na atualidade, com a relevância desses dados crescendo cada vez mais no contexto da *e-Science*, visto anteriormente. Mas o volume em que eles são produzidos e coletados

excede a capacidade humana de análise e interpretação, tornando imprescindível o uso de computadores avançados que possibilitam a integração e análise de quantidades massivas de dados, revelando padrões invisíveis a "olho nu" (SAYÃO; SALES, 2020).

Atualmente os dados têm um enorme potencial para solucionar problemas e apoiar tomadas de decisões assertivas, tudo com a análise automática dos sistemas de computadores. Com isso em mente, percebe-se a razão do protagonismo dos dados de pesquisa no quarto paradigma da ciência (*e-Science*), que precisa lidar com a quantidade cada vez maior de dados sendo produzidos.

Entretanto, para que os dados possam ser interpretados e analisados pelas máquinas, é preciso uma representação com metadados de qualidade, que fornecem a semântica necessária. Hoje é possível agilizar o ciclo científico, replicar pesquisas e desenvolver novos estudos com a economia de recursos graças ao compartilhamento e reutilização dos dados coletados pelos pesquisadores, representados de acordo com boas práticas em repositórios. Logo:

> [...] é imprescindível a conscientização entre os pesquisadores sobre a importância da disponibilização dos seus dados científicos, a criação de políticas que incentivem essa prática, o desenvolvimento tecnológico e a capacitação dos pesquisadores com ferramentas que facilitem o processo (ARAÚJO; DIAS; AUTRAN, 2021, p. 102).

A Ciência da Informação tem um papel fundamental nesse contexto, contribuindo para que o acesso e o uso dos dados se deem da melhor forma possível, identificando fatores que possibilitem a máxima otimização no uso dos dados de pesquisa, como é o caso das boas práticas internacionais. Os dados estão no foco do quarto paradigma da ciência, e os bibliotecários e profissionais da Ciência da Informação "[...] podem contribuir de forma vital com a curadoria de dados, a preservação, e com habilidades em arquivamento para garantir custódia segura da produção de pesquisa" (RODRIGUES, 2020, p. 34). Para isso é essencial as competências com repositórios de dados. É por meio desses sistemas de informação que os dados podem ser compartilhados, representados e acessados pelos pesquisadores ao redor do mundo, sendo então reutilizados, agilizando o ciclo da ciência.

3 REPOSITÓRIOS DE DADOS E BOAS PRÁTICAS

3.1 Repositórios de dados

Os repositórios de dados têm um papel importante na ciência contemporânea quando se fala na gestão dos dados de pesquisa, que demandam uma representação adequada por meio de metadados e identificadores. Não basta disponibilizá-los na *web*, é preciso contextualizá-los e garantir sua preservação, para que eles sejam então interpretados e reutilizados. É nesse sentido que se pode afirmar que "A gestão tem como infraestrutura central, conforme já afirmado, o repositório digital de dados de pesquisa " (SAYÃO; SALES, 2016, p. 95).

Sanchez, Vidotti e Vechiato (2017, p. 3) conceituam os repositórios de dados como aqueles que "[...] buscam organizar, estruturar, permitir acesso, disseminar e preservar todos os dados gerados por meio de pesquisas realizadas em sua maioria por Instituições de Ensino e Pesquisa". Eles costumam ser pautados pelo autoarquivamento, garantindo o armazenamento, gestão, acesso e preservação dos conjuntos de dados dos pesquisadores.

Já 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". Para Monteiro, Sant'Ana e Pérez (2019, p. 161), os repositórios de dados são "[...] ambientes digitais implementados nas universidades com infraestrutura para dar suporte aos pesquisadores na gestão e na disponibilização de dados científicos". Ou seja, os autores também fazem essa conexão direta dos repositórios de dados com as instituições de ensino e pesquisa.

Sales e Sayão (2016) acrescentam que os repositórios de dados podem ser divididos, de acordo com a literatura, em quatro tipos: institucionais, disciplinares, multidisciplinares e orientados por projetos. Os repositórios analisados nessa dissertação dizem respeito aos institucionais que, segundo os autores, são aqueles gerenciados e mantidos no âmbito de uma instituição acadêmica “[...] como universidades ou institutos de pesquisa, e são voltados para arquivar dados que são, geralmente, provenientes unicamente das atividades acadêmicas dessas instituições” (SALES; SAYÃO, 2016, p. 101). É preciso citar que, como as universidades abrangem um amplo escopo de cursos, de diferentes domínios, os repositórios institucionais

também são multidisciplinares, lidando com uma grande variedade de dados de pesquisa heterogêneos, o que pode dificultar seu trabalho com a consistência da representação desses recursos.

Conforme já visto, os dados de pesquisa são diferentes das publicações acadêmicas que falam por si só (SAYÃO; SALES, 2016). Os dados de pesquisa precisam de contextualização para serem interpretados, o que é feito por meio da documentação apropriada e da descrição em repositórios, com os devidos metadados e as ferramentas usadas para criá-los, armazená-los, adaptá-los e analisá-los. Isso porque:

> Sem uma descrição minuciosa do contexto tecnológico dos arquivos de dados, do contexto no qual os dados foram criados ou coletados, das medidas que foram feitas, dos detalhes espaciais e temporais, dos instrumentos usados, dos parâmetros e unidades e da qualidade dos dados e da sua proveniência, é improvável que os dados possam ser descobertos, interpretados, gerenciados e efetivamente usados e reusados (SAYÃO; SALES, 2015, p. 20).

Os autores ainda afirmam que esses conjuntos de dados de pesquisa precisam de ações específicas que levem em conta todo o seu ciclo de vida, uma vez que isso os permite "[...] revelarem e transmitirem conhecimento no tempo e no espaço, e a partir daí serem interpretadas, sintetizadas e reanalisadas em contextos diversos – e diferentes para os quais foram geradas e coletadas originalmente" (SAYÃO; SALES, 2016, p. 91), demonstrando a importância do compartilhamento em repositórios com descrições ricas.

Nesse sentido, Sayão e Sales (2015) argumentam que os repositórios de dados se incorporam rapidamente à infraestrutura mundial de informação científica, ainda mais quando se fala dos inúmeros benefícios trazidos pela gestão de dados em repositórios como: o aumento do impacto e da visibilidade da pesquisa; a inovação e novos usos potenciais aos dados; maximização da transparência; melhoria e validação dos métodos de pesquisa; redução de custos, dentre outros (FELIPE; SANTOS, 2022). Uma vantagem especialmente importante para os pesquisadores com relação ao compartilhamento de seus dados é o aumento da visibilidade e do impacto de sua pesquisa. A contagem de citações é frequentemente usada em decisões de financiamento e promoção de pesquisas, e esse compartilhamento pode trazer vantagens competitivas ao pesquisador.

Para visualização do benefício de forma prática, Piwowar e Vision (2013) fizeram um estudo que demonstrou que os trabalhos que disponibilizaram os dados

em um repositório público receberam 9% (intervalo de confiança de 95%: 5% a 13%) mais citações do que estudos semelhantes para os quais os dados não foram disponibilizados. Ou seja, os repositórios de dados trazem uma contribuição significativa para o alcance e disseminação das pesquisas, gerando maior impacto.

Apesar dos benefícios citados, muitos pesquisadores ainda se sentem relutantes em compartilhar seus dados. Em estudo de Veiga *et al.* (2019) para avaliar a percepção de pesquisadores da Fiocruz quanto ao compartilhamento de seus dados, 53% deles afirmou concordar plenamente que o compartilhamento de dados abertos contribui para o aumento da credibilidade da pesquisa.

Entretanto, 45% responderam que o motivo para não compartilharem seus dados era o receio deles serem mal interpretados e mal utilizados por outros pesquisadores. Esse é um problema que pode ser ligado, dentre outros fatores, à representação e contextualização dos dados nos repositórios. Dados têm um alto grau de abstração e, portanto, descrições ricas, com licenças bem declaradas, auxiliam os pesquisadores a interpretar e reutilizar os dados, fazendo sua citação adequada. Por isso é importante se pensar não só na infraestrutura tecnológica em si, mas também na representação dos dados de pesquisa, com um conjunto mínimo de elementos de metadados para assegurar as informações necessárias ao seu acesso e reutilização. Isso será visto em mais detalhes no capítulo de discussão de resultados.

Mas nem todos os pesquisadores possuem as competências necessárias para a gestão adequada de seus dados, o que foi visto na pesquisa com 52% dos respondentes afirmando que gostariam que a instituição oferecesse serviços de compartilhamento e acesso aos dados; e com 50% afirmando que gostaria de formação e consultoria na gestão de seus dados (VEIGA *et al.,* 2019). Nota-se um interesse em serviços institucionais que apoiem o compartilhamento e gestão dos dados em todo o ciclo de vida da pesquisa. Isso porque a maioria dos pesquisadores compreende a importância dessa iniciativa, garantindo a transparência, reprodutibilidade e avanço científico, mas há ainda algumas barreiras:

> A desinformação sobre as formas e normas de citação de dados e do uso dos identificadores persistentes (que facilitam a citação dos dados) também afasta o pesquisador do compartilhamento de dados. A má utilização ou interpretação dos dados também preocupa o pesquisador, mas ele desconhece a documentação que deve acompanhar os dados compartilhados, minimizando a utilização ou

> interpretação inadequada por outros pesquisadores (VEIGA *et al.*, 2019, p. 327).

Além da desinformação com relação às normas, às políticas e à documentação necessária para garantir o melhor aproveitamento de seus dados, grande parte dos pesquisadores afirma não conhecer repositórios para fazer o compartilhamento. De acordo com Sayão e Sales (2015, p. 49), "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".

Uma solução para essa questão são os diretórios que mapeiam repositórios de dados ao redor do mundo, promovendo uma cultura de compartilhamento, acesso e visibilidade dos dados de pesquisa, como o *Re3data.org.* Ele é um diretório global de repositórios de dados de pesquisa que abrange todas as disciplinas acadêmicas. Apresenta repositórios para armazenamento permanente e acesso a conjuntos de dados de pesquisa para pesquisadores, órgãos de financiamento, editores e instituições acadêmicas (STRECKER *et al*., 2021).

O *Re3data.org* promove uma cultura de compartilhamento, maior acesso e visibilidade de dados de pesquisa, e é um serviço parceiro da *DataCite*, uma organização global sem fins lucrativos que está ativamente envolvida em várias iniciativas para melhorar a disponibilidade e citação de resultados de pesquisa (STRECKER *et al*., 2021). O projeto começou a indexar repositórios de dados de pesquisa em 2012 e oferece aos pesquisadores, organizações financiadoras, bibliotecas e editores uma visão sistemática do cenário heterogêneo dos repositórios de dados de pesquisa. Em junho de 2022, o *Re3data.org* listava 2.874 repositórios.

O principal objetivo do *Re3data.org* é oferecer orientação aos pesquisadores no cenário heterogêneo de repositórios de dados de pesquisa, tanto em seu papel como produtores de dados quanto como usuários de dados de seus pares. Outros grupos-alvo são financiadores de pesquisa, *data centers* e bibliotecas universitárias. Além disso, o *Re3data.org* visa contribuir para o estabelecimento de um ecossistema de repositórios de dados mais coerente e integrado.

Sendo assim, o diretório ajuda os pesquisadores a encontrarem repositórios apropriados para o armazenamento e acesso de seus dados. Uma pesquisa realizada por país, em setembro de 2022, no *Re3data.org*, mostrou que o Brasil tinha cadastrado 17 repositórios de dados. Em 2018, o diretório mapeava nove repositórios brasileiros (PAGANINE; AMARO, 2020). Nota-se um aumento importante em quatro

anos, apesar de ainda não representar 100% o cenário nacional. Apenas dois repositórios da amostra (UNESP e Unicamp), por exemplo, estavam mapeados no diretório. Todos os demais (UFSCar, USP, Unifesp e UFABC) não estavam presentes no catálogo do Brasil, afetando sua visibilidade.

Apesar de 17 ser um número ainda baixo perto de países como Estados Unidos (1.154 repositórios de dados), Canadá (315 repositórios de dados), Alemanha (481 repositórios de dados), França (124 repositórios de dados) e Espanha (48 repositórios de dados), quando comparado ao número de repositórios de dados de outros países da América Latina, o Brasil se mostra na dianteira. A Argentina, por exemplo, apresenta oito repositórios mapeados, a Colômbia 11, o Peru três e o Chile dois.

Os diretórios ajudam na descoberta e seleção de ambientes para o compartilhamento seguro dos dados de pesquisa, mapeando inclusive iniciativas de diversas instituições de ensino superior brasileiras em criar seus próprios repositórios de dados. Mas a própria FAPESP criou um metabuscador dos repositórios de dados do Estado de São Paulo, aumentando a visibilidade deles. Isso facilita o compartilhamento seguro dos dados, que podem ser depositados dentro da própria instituição do pesquisador, contando com o apoio das equipes responsáveis.

Mas além da falta de conhecimento sobre iniciativas ligadas à gestão dos dados e a relutância dos pesquisadores quanto a esse movimento, um estudo de Rodrigues, Dias e Lourenço (2022, p. 327) apontou que, no âmbito dos profissionais da informação, ainda é preciso investir em capacitação quando se trata de dados e projetos para a implementação de repositórios de dados de pesquisa. Isso porque é preciso "[...] entendimento das divergentes necessidades entre comunidades, pelo conhecimento técnico computacional exigido a tais práticas, e idealmente, pela busca da padronização e manutenção desses serviços", ou seja, implementação de boas práticas.

Não se pode esquecer que os dados de pesquisa são extremamente heterogêneos, o que configura um desafio para as equipes gestoras dos repositórios. Isso demanda competências e habilidades específicas ligadas à captura, catalogação, arquivamento e preservação desses dados. É esperado, portanto, que haja uma certa dificuldade por parte das equipes de gerir e manter os repositórios de dados de suas instituições, principalmente considerando que muitos deles são ainda recentes e surgiram para suprir uma demanda que muitos profissionais ainda não estavam preparados.

Mas é importante que os profissionais responsáveis pelo repositório e até mesmo os pesquisadores estejam a par das habilidades necessárias para a gestão dos dados, porque quando descritos com metadados ricos, fornecendo a contextualização necessária, o uso dessas infraestruturas tecnológicas traz inúmeros benefícios para a sociedade, para os próprios pesquisadores e para a ciência, como créditos ao autor, memória científica e transparência, segurança dos dados, indicação da qualidade e produtividade da instituição, dentre outros (SAYÃO; SALES, 2016). O conhecimento sobre o papel dos metadados, que há anos estão ligados ao trabalho dos profissionais da informação, é, portanto, essencial para maximizar o potencial dos repositórios e dos próprios dados de pesquisa depositados.

3.2 Metadados e dados de pesquisa

A relação entre os metadados e os repositórios é indiscutível e extremamente importante no contexto dos dados de pesquisa que, como visto, dependem de descrições ricas para serem interpretados. É preciso fornecer dados que falem sobre os dados, trazendo contexto e semântica para eles. Os metadados fornecem informações adicionais que ajudam os usuários dos dados a entenderem melhor seu significado, sua origem, sua estrutura e esclarecer outras questões como direitos e termos de licença, a organização que os gerou, os métodos de acesso e o cronograma de atualização dos conjuntos de dados. Logo, os metadados são essenciais na contextualização, interpretação e reutilização dos dados de pesquisa.

A ideia de "metadado", segundo Pomerantz (2015), remonta à época das primeiras bibliotecas. Metadados são declarações sobre os dados e, linguisticamente, o prefixo "meta" vem do grego e significa "além de". Ou seja, de forma simplista, os metadados acrescentam dados sobre os dados, facilitando a organização, a descoberta e a recuperação de itens.

Bibliotecários trabalharam com metadados ao longo de centenas de anos, chamando-os de "informação no catálogo da biblioteca", que buscava ajudar os usuários a encontrarem materiais na coleção da instituição (POMERANTZ, 2015). Um caso a ser citado é o *Pinakes*, considerado por historiadores como o primeiro catálogo de biblioteca, criado na época da biblioteca de Alexandria por volta de 245 a.C. Sabe-se que as obras da coleção eram listadas por alguns critérios como gênero, título, autor, resumo, etc. Décadas depois, as bibliotecas ainda utilizam os mesmos pedaços

de informação nos catálogos (adicionando novos), e a natureza e o objetivo da descrição de entidades portadoras de informação permanece praticamente igual (ZENG; QIN, 2016).

Percebe-se que a ideia do metadado já existia há anos, mas o termo como hoje é conhecido foi criado no final da década de 60 e se popularizou entre as décadas de 80 e 90, com a criação do padrão *Dublin Core* (metadados descritivos para recursos da *web*). Pomerantz (2015) afirma que os metadados se referem a uma declaração, uma afirmação sobre um objeto potencialmente informativo. De forma semelhante, Zeng e Qin (2016) definem metadados como dados estruturados e codificados que descrevem características de entidades portadoras de informações. Fala-se sobre uma declaração emparelhada de propriedade-valor, descrevendo um recurso.

Adicionalmente, Gilliland (2016) afirma que todos os objetos informacionais, independente do formato físico ou intelectual, possuem três dimensões: 1) conteúdo, que diz respeito ao que o objeto contém ou sobre o que ele é; 2) contexto, que pode ser traduzido pelas perguntas "quem?", "o que?", "por quê?", "onde?" e "como?" associadas com a criação do objeto; e 3) estrutura, ou o conjunto formal de associações dentro ou entre objetos de informação individuais. Todas essas dimensões precisam ser contempladas pelos metadados dentro do cenário dos repositórios de dados de pesquisa.

Zeng e Qin (2016) dividem os metadados em cinco categorias: os administrativos, descritivos, de preservação, técnicos e de uso. Os administrativos são aqueles usados na gestão de coleções e recursos informacionais (exemplos: aquisição, direitos e reprodução, informações de localização, *etc*.). Os descritivos são aqueles usados para identificar e descrever coleções e recursos informacionais relacionados (exemplos: catalogação de registros, informações de curadoria, *etc*.). Os de preservação dizem respeito aos metadados ligados à preservação das coleções e dos recursos (exemplos: documentação das condições físicas do recurso, ações tomadas para a preservação, versões digitais do recurso, *etc*.). Já os metadados técnicos são os relacionados com como o sistema funciona ou os metadados se comportam (exemplos: informação sobre os requisitos de *hardware* e *software*, digitalização, autenticação e segurança dos dados, *etc*.). Por fim, os metadados de uso são aqueles que se referem ao nível e tipo de uso das coleções e recursos (exemplos: registros de circulação, histórico de uso, *etc*.). Nesse sentido, Pavão *et al.* (2015, p. 104) acrescentam que:

> Os metadados são usados para definir permissões, direitos de acesso, compartilhamento, reutilização, redistribuição e políticas, bem como os requisitos técnicos para visualização, acesso ou preservação de objetos digitalizados ou concebidos originalmente em formato digital.

As entidades portadoras de informação podem ser de vários tipos, e exigem diferentes metadados para descrevê-los, permitindo sua descoberta, identificação, acesso e uso (ZENG; QIN, 2015). É preciso levar em conta as especificidades de cada domínio, principalmente quando se fala nos dados de pesquisa enquanto objetos digitais em repositórios institucionais. Como já visto, esses dados são heterogêneos e se mostram como um desafio pelo alto grau de contextualização necessário. Não se pode esquecer que a *web* é um espaço de informação aberto, onde a ausência de um contexto específico, como o sistema de informação interno de uma organização, faz com que a disponibilização de metadados seja um requisito fundamental.

Apesar das particularidades intrínsecas aos dados de pesquisa, é preciso pensar também em uma padronização da representação, que permite a recuperação eficaz da informação e garante a interoperabilidade dos sistemas. Henning *et al*. (2019b) afirmam que, dentre as questões relativas à gestão de dados de pesquisa, uma das mais preocupantes é a falta de padronização. A homogeneização das descrições dos dados de pesquisa é essencial no processo de interoperabilidade transparente entre as instituições, evitando esforços de integração adicionais.

Inclusive, um dos objetivos da interoperabilidade é ajudar os usuários a encontrar e acessar recursos que são distribuídos entre domínios e instituições, fundamental no contexto da *web*. De acordo com Gomes (2015), ainda há a carência de ações que buscam a padronização dos metadados de objetos digitais que são disponibilizados em ambientes digitais como os repositórios, e por isso é importante reafirmar sua importância.

Nesse sentido, Sousa (2012) afirma que os bibliotecários têm um papel central nos repositórios institucionais, justamente na análise e produção de metadados para assegurar a qualidade e padronização das representações dos recursos. Pavão *et al*. (2015, p. 109) argumentam que é essencial "[...] estabelecer os requisitos de descrição de cada elemento e promover a padronização, normalização e enriquecimento dos metadados para fortalecer a qualidade dos registros". Não basta disponibilizar os dados de pesquisa, é preciso pensar na consistência para que os

metadados sejam mais efetivos em suas funções. É onde entram os padrões de metadados que:

> [...] estabelecem regras para a definição de atributos (metadados) de recursos de informacionais, para a) obter coerência interna entre os elementos por meio de semântica e sintaxe; b) promover necessária facilidade para esses recursos serem recuperados pelos usuários; c) permitir a interoperabilidade dos recursos de informação. (ALVES, 2010, p. 47).

Os metadados são elementos primordiais para a gestão dos dados, e por isso é tão importante que a instituição mantenedora do repositório considere quais padrões são comumente usados entre organizações semelhantes, pois usar o mesmo melhora a interoperabilidade entre as coleções e sistemas. Mas definir padrões de metadados não é uma tarefa tão simples, justamente porque os repositórios institucionais contam com recursos de diferentes domínios, em diferentes formatos, levando a instituição a avaliar se um único padrão de metadados atende a todas as coleções ou se será preciso combiná-los (PAVÃO *et al.*, 2015).

Além da escolha do padrão, ainda é preciso pensar na escolha dos elementos de metadados que serão utilizados na representação dos dados de pesquisa. Lóscio, Burle e Calegari (2017) citam alguns metadados descritivos para conjuntos de dados como: título; descrição; palavras-chave; data de publicação; entidade responsável por disponibilizar o conjunto de dados; cobertura espacial; período temporal coberto; data da última modificação e temas/categorias cobertos pelo conjunto de dados.

Esses metadados citados permitem que os usuários humanos interpretem a natureza do conjunto de dados, e que os agentes de *software* descubram automaticamente conjuntos de dados. É importante, portanto, que as equipes gestoras dos repositórios de dados de pesquisa busquem incentivar descrições mais ricas, indicando alguns elementos de metadados essenciais para o acesso, interpretação e reutilização de dados. Mas isso ainda pode esbarrar no problema da heterogeneidade dos dados, que podem exigir elementos de descrição diferentes, um dos maiores desafios para a interoperabilidade dos sistemas. Assim, visando diminuir a falta de padronização para alcançar maiores níveis de interoperabilidade e consistência, surgem as boas práticas.

3.3 Boas práticas para publicação de dados de pesquisa em repositórios

Guandalini, Furnival e Arakaki (2019, p. 3) afirmam que o termo “boas práticas” “[...] pode ser entendido como condutas adotadas para uma maior divulgação, disseminação e desenvolvimento da ciência”. Já a publicação se refere à liberação de dados de pesquisa, dos metadados associados, da documentação que os acompanha e do código de *software* (nos casos em que os dados brutos foram processados ou manipulados) para reutilização e análise de forma que possam ser descobertos na *web* de forma única e persistente (AUSTIN *et al.,* 2016).

A publicação de dados ocorre por meio de repositórios de dados e/ou periódicos de dados que garantem que os objetos de pesquisa publicados sejam bem documentados, curados, arquivados a longo prazo, interoperáveis, citáveis, com qualidade garantida e detectáveis – todos os aspectos da publicação de dados que são importantes para reutilização. Considerando então a relevância dos dados de pesquisa e dos repositórios de dados, é preciso falar das boas práticas associadas a eles. E, como pode ser visto, os metadados estão no centro dessas discussões.

Austin *et al*. (2016) explicam que os mandatos de editores e agências de financiamento para tornar acessíveis os dados subjacentes às publicações estão mudando a conversa de “Os pesquisadores devem publicar seus dados?” para “Como podemos publicar dados de forma confiável?”. Agora já é possível observar requisitos de abertura e transparência e um impulso para considerar os dados como um resultado de pesquisa de primeira classe. Logo, considerando o valor desses dados, é de suma importância que as partes interessadas se atentem às boas práticas para extrair o máximo de benefício deles.

Um exemplo a ser citado são os seis princípios dos “bons metadados” publicados pela *National Information Standards Organization* (NISO, 2007, p. 61, tradução nossa), com "bom" se referindo à exigência de interoperabilidade, reutilização, persistência, verificação, documentação e suporte para direitos de propriedade intelectual. São eles:

1) "bons metadados" estão em conformidade com os padrões da comunidade de maneira apropriada aos materiais, usuários e usos atuais e potenciais da coleção;
2) oferecem suporte à interoperabilidade;

3) usam controle de autoridade e padrões de conteúdo para descrever objetos e relacioná-los;
4) incluem uma declaração clara das condições e termos de uso do objeto digital;
5) apoiam a curadoria e preservação a longo prazo de objetos em coleções;
6) são objetos em si e, portanto, devem ter as qualidades de bons objetos, incluindo autoridade, autenticidade, arquivamento, persistência e identificação única.

Os princípios listados demonstram a importância dada internacionalmente aos "bons metadados", ou metadados de qualidade. A NISO (2007) afirma que a identificação dos objetos digitais na *web* é um dos maiores desafios dentro desse cenário e os metadados são fundamentais para contorná-lo. É a existência de metadados descritivos pesquisáveis que aumenta a probabilidade do conteúdo digital ser descoberto e utilizado, além de garantir a interoperabilidade dos repositórios, um dos aspectos essenciais no ambiente da *web* atualmente. Pavão *et al.* (2015, p. 110) ainda acrescentam que:

> Ao mesmo tempo em que adotam diretrizes comuns, os repositórios devem procurar aprimorar os metadados de descrição de conteúdo considerando a missão de manter um padrão de alta qualidade, em conformidade com as mais recentes tecnologias, e desenvolver recursos, a fim de que os conteúdos possam ser inequivocamente identificados tendo em vista sua ampla disseminação.

Em concordância, a NISO (2007) orienta que metadados de qualidade devem ser coerentes, significativos e úteis em contextos globais além daqueles em que foram criados. Isso significa que eles devem incluir todas as informações pertinentes sobre o objeto digital, pois suposições sobre o contexto no qual ele é acessado localmente podem não ser mais válidas no ambiente de rede. É preciso pensar os recursos de uma coleção dentro de um cenário além do próprio repositório. "Ações cooperativas e a observância das normas e padrões na definição de esquemas e perfis de metadados para descrição de objetos digitais são imprescindíveis para proporcionar a uniformidade na descrição e a interoperabilidade" (PAVÃO *et al.*, 2015, p. 115).

Felipe e Santos (2022, p. 16) acrescentam que os metadados "[...] condicionam a recuperação, o acesso, o uso e o reuso dos dados", ou seja, há uma dependência entre os elementos "dado" e "metadado". Os metadados, quando pensados no contexto das boas práticas, garantem a contextualização e interpretação dos conjuntos de dados depositados nos repositórios. E, além da interpretação humana dos recursos, há também o processamento automático pelas máquinas, tópico que vem sendo amplamente debatido e estudado, principalmente no contexto dos princípios FAIR.

Torino e Vidotti (2021, p. 8), em estudo que analisa boas práticas propostas pela *World Wide Web Consortium* (W3C) para dados na *web*, afirmam que um dos grandes desafios relacionado a sua disponibilização é "[...] fornecer metadados compreensíveis a usuários humanos e agentes computacionais, para que possam ser compreensíveis e processáveis, quer seja apresentando-os como parte de uma página HTML ou em um arquivo adicional.". A W3C forneceu uma estrutura de 35 boas práticas para dados, com abordagens de implementação, exemplos, formas de testar a aplicação, benefícios, dentre outras características que auxiliam na aderência às boas práticas.

É recomendado que os agentes computacionais "[...] tenham acesso à uma estrutura padronizada e semanticamente formal, que privilegia a descoberta, a interoperabilidade e o desenvolvimento de aplicações para o consumo automático dos dados" (TORINO; VIDOTTI, 2021, p. 8), e a representação adequada e rica deles é fundamental para isso. É preciso ter em mente as vantagens de permitir que as máquinas acessem, usem e interpretem automaticamente os conjuntos massivos de dados, permitindo análises e detectando padrões antes impossíveis pelo homem, principalmente no contexto já citado da *e-Science*.

Levando em conta a importância de dados legíveis por máquina, com estruturas padronizadas e enriquecidas semanticamente, permitindo aos agentes computacionais a interpretação automática, cita-se a iniciativa FAIR. Felipe e Santos (2022, p. 2) acrescentam que essa iniciativa "[...] apresenta requisitos que tendem a promover melhores práticas de gestão de dados mediante ações e tecnologias voltadas a essa gestão", buscando uma melhoria da qualidade dos dados de pesquisa. Dentro do cenário da *web* e da busca pela padronização na gestão dos dados é fundamental conhecer e se atentar aos princípios FAIR.

Os princípios FAIR, um acrônimo para "*Findable*", "*Accessible*", "*Interoperable*" e "*Reusable*", que em português significa "Encontrável", "Acessível", "Interoperável" e "Reutilizável", são diretrizes internacionalmente conhecidas. Os princípios foram estabelecidos como resultado da conferência internacional '*Jointly designing the data FAIRPORT*' realizada em janeiro de 2014 no *Lorentz Center* em Leiden, Holanda (HENNING *et al*., 2021). A conferência reuniu 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,* buscando criar uma infraestrutura global para dados.

Mas a disseminação mais ampla dos princípios FAIR começou em março de 2016, com a sua publicação no *Nature's Journal Scientific Data* e "[...] tiveram sua aplicação consolidada em 2017, quando a Comissão Europeia passou a exigir a adoção de plano de gestão de dados, com base nesses princípios, por projetos financiados por seus recursos" (HENNING *et al.*, 2019a, p. 175). Ainda de acordo com Mons *et al*. (2020), a principal contribuição que os princípios FAIR devem trazer para a pesquisa do século 21 é apoiar máquinas e humanos em seu trabalho colaborativo.

Ser legível por máquinas é uma das principais bandeiras defendidas pelos princípios FAIR, em busca do processamento automático da grande massa de dados disponíveis na *web,* que diz respeito à *e-Science*. Esse volume massivo de dados ultrapassa a capacidade humana de processamento, e por isso a importância de se pensar em dados legíveis por máquinas. Os princípios FAIR são divididos da seguinte forma:

Quadro 2 - Princípios FAIR

PRINCÍPIOS FAIR	
Para ser **encontrável** (***findable***):	E1 - um identificador globalmente exclusivo e persistente deve ser atribuído aos (meta)dados
	E2 - os dados são descritos com metadados enriquecidos
	E3 - os metadados incluem de forma clara e explícita o identificador dos dados que descrevem
	E4 - os (meta)dados são registrados ou indexados em um recurso que permite pesquisas
	A1 - (meta)dados são recuperáveis por seu identificador usando um protocolo de comunicação padronizado

Para ser **acessível (*accessible*)**:	A1.1 - o protocolo é aberto, gratuito e universalmente implementável
	A1.2 - o protocolo permite um procedimento de autenticação e autorização, quando necessário
	A2 - metadados são acessíveis, mesmo quando os dados não estão mais disponíveis
Para ser **interoperável (*interoperable*):**	I1 - (meta)dados usam uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento
	I2 - (meta)dados usam vocabulários que seguem os princípios FAIR
	I3 - (meta)dados fazem referências qualificadas a outros (meta)dados
Para ser **reutilizável (*reusable*)**:	R1 - (meta)dados são descritos de forma rica com uma pluralidade de atributos precisos e relevantes
	R1.1 - (meta)dados são publicados com uma licença de uso de dados clara e acessível
	R1.2 - (meta)dados estão associados a uma proveniência detalhada
	R1.3 - (meta)dados atendem aos padrões da comunidade relevantes para o domínio

Fonte: Adaptado de Wilkinson *et al.* (2016).

Wilkinson *et al*. (2016) explicam que os princípios FAIR enfatizam o aprimoramento da capacidade das máquinas de encontrar e usar os dados automaticamente, além de apoiar sua reutilização por indivíduos. Isso porque o ecossistema digital existente em torno da publicação de dados de pesquisa impede a extração do benefício máximo dos investimentos. A explosão de dados criados por pesquisadores através de avanços cada vez maiores em automação e instrumentação gera a necessidade de “máquinas” como assistentes analíticos. Mas para que os dados possam ser interpretados pelas máquinas, é preciso torná-los os mais eficientes possíveis (MONS *et al*., 2020).

Sendo assim, é necessário mais do que apenas disponibilizar e armazenar os dados de pesquisa na *web*, uma vez que seu potencial de reuso "[...] 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.*, 2019a, 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, o que demanda uma gestão adequada deles, pensando nas boas práticas internacionalmente adotadas.

Quando os dados não são devidamente estruturados, descritos com metadados ricos e seguindo padrões, impõem-se limitações às suas contribuições para novos ciclos de pesquisa, o que também impõe barreiras ao avanço científico. O "reutilizável" é justamente um dos princípios defendidos pelo FAIR, mas volta a tocar na problemática da má interpretação dos dados por outros pesquisadores quando não há a documentação adequada, permitindo até mesmo às máquinas fazer a interpretação e reutilizar os dados já coletados.

Wilkinson *et al*. (2016) exemplificam essa questão ao afirmar que frequentemente é gasto um período longo de semanas (e até meses) de esforço técnico especializado para reunir os dados necessários para uma investigação científica. O motivo não é a falta de tecnologia apropriada, mas sim a falta de um preparo desses dados, para que eles possam ser acionáveis por máquina, como é defendido pelos princípios FAIR. De acordo com Sales *et al*. (2020), os princípios FAIR já são reconhecidos mundialmente como elementos-chave para boas práticas em todos os processos de gerenciamento de dados, mas há ainda barreiras que impedem a ampla implementação de dados FAIR no cenário acadêmico.

Apesar disso, a adesão a esses princípios vem sendo defendida e ampliada por várias comunidades ligadas à pesquisa internacionalmente (HENNING *et al.*, 2021). A *Australian Research Data Commons* (ARDC), por exemplo, o órgão máximo da Austrália para dados de pesquisa, defende que dados FAIR maximizam o impacto do investimento, incluindo a obtenção de mais citações para os conjuntos de dados. Já o *Office of Data Science Strategy* do *National Institutes of Health* esclarece que todos os dados de pesquisa biomédica devem aderir aos princípios FAIR. É importante lembrar que novas tecnologias e mudanças culturais no modo de fazer ciência costumam depender de políticas mandatórias de agências e do governo para que haja um movimento de conscientização e capacitação.

Logo, fica nítido que os pesquisadores precisam de todo um novo arcabouço teórico e prático, com serviços institucionais de apoio para se adaptarem a um novo modelo que demanda a gestão de dados de acordo com boas práticas. Entretanto, é preciso citar que os princípios FAIR, por exemplo, quando “[...] colocados em prática, podem gerar equívocos devido às dificuldades no seu entendimento e na falta de experiências na sua aplicação e uso.” (HENNING *et al.,* 2019b, p. 404).

Os princípios FAIR não são e não exigem padrões ou ferramentas específicas. Seu objetivo é "[...] contextualizar e apontar para direção de maior utilidade e melhores serviços de dados, oferecendo suporte à sua reutilização e, assim, facilitar a escolha de quais padrões podem ser utilizados" (HENNING *et al.*, 2019a, p. 179). Além disso, "[...] os elementos dos princípios FAIR estão relacionados, mas são independentes e separáveis e podem ser implementados em qualquer combinação" (SILVA; SANTAREM SEGUNDO; SILVA, 2018, p. 5201), considerando as necessidades e investimentos de cada instituição, que evoluem em busca de níveis cada vez maiores de conformidade com os princípios FAIR.

É importante citar que o FAIR não é binário (FAIR/não FAIR), mas sim um espectro ao longo do qual vários graus de *FAIRness* são possíveis (HIGMAN; BANGERT; JONES, 2019). Ou seja, existem diferentes ordens e meios de implementar os princípios FAIR, o que leva a uma abertura e liberdade para que pesquisadores e instituições decidam como querem implementá-los. Isso, de certa forma, pode ser considerado positivo, uma vez que cada país, instituição ou área de pesquisa pode investir no que se adequa melhor ao seu contexto e necessidades.

Entretanto, isso cria desafios quanto à aplicação dos princípios, uma vez que eles não foram devidamente detalhados e explicados na sua publicação original (WILKINSON *et al.*, 2016), sendo mais estabelecidos como conceitos. Isso pode estar diretamente ligado com a dificuldade de instituições aplicarem os princípios FAIR aos seus dados de pesquisa, uma vez que a própria publicação original é recente e abre margem para interpretações ambíguas.

Na prática, "[...] os princípios FAIR estão sujeitos a diferentes interpretações, levando a diferentes aplicações e implementações e, consequentemente, reduzindo sua eficácia" (HENNING *et al*., 2021, p. 724, tradução nossa). Além do fato de que o "[...] que é considerado FAIR em uma comunidade pode não ser para outra comunidade" (HENNING *et al.*, 2019a, p. 181). A flexibilidade quanto à aplicação se mostra, portanto, uma via de mão dupla, que pode acabar dificultando a implementação e diminuindo sua eficácia pela falta de padronização.

Com isso em mente, Devaraju *et al*. (2021) propõem que os pesquisadores tenham o apoio do repositório na hora do depósito de seus dados. Os repositórios podem fornecer uma lista de verificação manual, automática ou semiautomática adaptada à prática de curadoria de dados. Podem ainda implementar um recurso para verificar automaticamente certos aspectos como parte do fluxo de trabalho de

depósito. Além de garantir dados mais FAIR, o repositório também garantiria uma maior padronização dos conjuntos de dados depositados em sua coleção.

Ainda nesse sentido, Hodson *et al.* (2018), um grupo de especialistas europeus, expuseram recomendações e práticas em um documento intitulado *Turning FAIR into reality*[7]. De acordo com os autores, os princípios FAIR "[...] adotam uma abordagem centrada em dados, listando aquilo que precisa ser feito e os atributos que os dados precisam ter para possuírem maior valor e usabilidade, por humanos e máquinas" (HODSON *et al*., 2018, p. 15, tradução nossa).

Entretanto, surgem dúvidas sobre o significado exato de certos princípios e como eles se aplicam a diferentes disciplinas de pesquisa, e por isso é importante, ao considerar a implementação dos princípios FAIR, definir passos para se aplicar adequadamente. No relatório os autores agrupam recomendações primárias em quatro etapas para tornar os dados FAIR uma realidade, sendo elas:

Quadro 3 - Recomendações primárias para aplicação dos princípios FAIR

RECOMENDAÇÕES PRIMÁRIAS PARA APLICAÇÃO DOS PRINCÍPIOS FAIR	
Definir e aplicar FAIR adequadamente	FAIR não se limita aos seus quatro elementos constituintes: ele também deve incluir a abertura apropriada, a acessibilidade aos dados, a administração de longo prazo e outras características relevantes
	O mandato de Dados Abertos para pesquisas com financiamento público deve ser explícito em todas as políticas. É importante que a máxima “tão aberto quanto possível, tão fechado quanto necessário” seja aplicada proporcionalmente com os melhores esforços genuínos para compartilhar dados
	A implementação do FAIR requer um modelo para dados FAIR que, por definição, tenha um identificador persistente vinculado a diferentes tipos de metadados essenciais, incluindo proveniência e licenciamento. O uso de padrões da comunidade e o compartilhamento de código também são fundamentais para a interoperabilidade e a reutilização
Desenvolver e	A aplicação de dados FAIR depende, no mínimo, dos seguintes componentes essenciais: políticas, plano de gestão de dados, identificadores, padrões e repositórios. É necessário haver registros catalogando cada componente do ecossistema e fluxos de trabalho automatizados entre eles
	Os componentes do ecossistema FAIR precisam ser mantidos em um serviço profissional com financiamento sustentável

[7] Disponível em: https://doi.org/10.5281/zenodo.1285272.

apoiar um ecossistema de dados FAIR sustentável	Os financiadores de serviços de dados de pesquisa devem consolidar e aproveitar os investimentos existentes em infraestrutura e ferramentas. O financiamento deve estar vinculado a esquemas de certificação à medida que se desenvolvem para cada um dos componentes do ecossistema FAIR
Garantir dados FAIR e serviços certificados	É preciso dar apoio às comunidades de pesquisa para que elas desenvolvam e mantenham suas estruturas de interoperabilidade disciplinar: princípios e práticas para gerenciamento e compartilhamento de dados, acordos comunitários, formatos de dados, padrões de metadados, ferramentas e infraestrutura de dados
	As estruturas de interoperabilidade devem ser articuladas de maneira comum e adotar padrões globais sempre que possível para permitir a pesquisa interdisciplinar
	Um conjunto de métricas para dados FAIR deve ser desenvolvido e implementado, a partir do núcleo comum básico de metadados descritivos, identificadores persistentes únicos e acesso
	Os repositórios precisam ser incentivados e apoiados para obter a certificação *CoreTrustSeal*
	Esquemas de certificação são necessários para avaliar todos os componentes do ecossistema de dados FAIR. Assim como o *CoreTrustSeal*, eles devem abordar aspectos de gerenciamento e sustentabilidade de serviços, em vez de se basearem apenas nos princípios FAIR, que são articulados principalmente para dados e objetos
Incorporar uma cultura de FAIR na prática de pesquisa	Qualquer projeto de pesquisa deve incluir o gerenciamento de dados como elemento central necessário para a entrega de seus objetivos científicos, abordando isso em um plano de gestão de dados
	É preciso tomar medidas para desenvolver duas coortes de profissionais para apoiar os dados FAIR: cientistas de dados inseridos nos projetos de pesquisa que precisam deles e administradores de dados que garantirão a gestão e a curadoria de dados FAIR
	Os dados FAIR devem ser reconhecidos como um produto central da pesquisa e incluídos na avaliação das contribuições da pesquisa e progressão na carreira. O fornecimento de infraestrutura e serviços que permitem dados FAIR também devem ser reconhecidos e recompensados de acordo

Fonte: Adaptado de Hodson *et al*. (2018).

Definir o que é FAIR, desenvolver e apoiar um ecossistema sustentável, garantir que os dados sejam FAIR por meio de certificações e incorporar uma cultura

de dados FAIR no dia a dia da pesquisa são algumas das recomendações quando se pensa em tornar essa aplicação uma realidade palpável. O relatório acima se concentra nas ações necessárias em termos de cultura e tecnologia de dados de pesquisa, mas sabe-se que a implementação de um ecossistema FAIR depende de várias iniciativas e estudos, uma mudança cultural na prática da pesquisa e a capacitação dos pesquisadores. Muitos sequer sabem por onde começar ou onde depositar seus dados, além do receio do mau uso dos dados depositados.

Por isso é importante ver todos os componentes aqui citados como um grande sistema que envolve dados-metadados-boas práticas como um caminho a ser seguido em prol de dados de pesquisa com máxima extração de benefícios. De acordo com Hodson *et al*. (2018), os dados de pesquisa devem ser disponibilizados (e FAIR) o mais rápido possível. Isso é fundamental para garantir que as comunidades de pesquisa possam colaborar de forma eficaz e aumentar a velocidade da resposta e de novas descobertas.

Mais importante é lembrar que os dados de pesquisa, ao contrário de artigos e livros, por exemplo, não falam por si mesmos, contendo um alto grau de abstração. Por isso os metadados desempenham um papel tão importante, porque garantem a contextualização necessária para que os dados sejam legíveis por humanos e por máquinas. Essa segunda sendo uma tarefa mais complexa, que exige o cuidado com o uso de vocabulários, padrões internacionais e a adoção de boas práticas, em busca da integridade científica, da interoperabilidade e da credibilidade dos dados, do repositório, do pesquisador e da instituição.

Sendo assim, os princípios FAIR buscam maximizar o benefício que pode ser extraído dos dados e trazem diretrizes que podem ser aplicadas para que os dados de pesquisa sejam encontráveis, acessíveis, interoperáveis e reutilizáveis. Esses princípios, como visto, se aplicam aos metadados e, segundo Henning *et al*. (2019b, p. 398) "[...] devem ser aceitos pela comunidade de produtores e consumidores de dados de pesquisa, com a finalidade de incorporar boas práticas para a publicação e compartilhamento de dados científicos".

Considerando então todo esse contexto e a importância das boas práticas para pesquisadores, agências de financiamento, universidades, formuladores de políticas de pesquisa e os usuários dos dados, no Capítulo 5 é possível observar a avaliação dos níveis de conformidade que se encontram os dados de pesquisa depositados nos repositórios da amostra quanto a cada um dos princípios FAIR. Isso permitirá

visualizar em mais detalhes como todos os conceitos e diretrizes citadas estão sendo aplicadas no contexto regional de São Paulo, culminando em *feedbacks* para as instituições da amostra. Dessa forma será possível apontar alguns caminhos para a melhoria do cenário nacional quanto aos princípios FAIR, maximizando o potencial dos nossos dados.

4 PROCEDIMENTOS METODOLÓGICOS

A pesquisa caracterizou-se como exploratória, que busca "[...] proporcionar maior familiaridade com o problema, tornando-o explícito", e descritiva, que "[..] expõe as características de uma determinada população ou fenômeno, demandando técnicas padronizadas de coleta de dados" (PRODANOV; FREITAS, 2013, p. 127). Por meio de uma abordagem quali-quantitativa buscou avaliar os dados de pesquisa depositados nos repositórios do Estado de São Paulo à luz dos princípios FAIR, podendo então trazer um *feedback* para as instituições da amostra.

A amostra foi definida a partir do metabuscador de dados de pesquisa da FAPESP, que reúne repositórios de dados de pesquisa de instituições no Estado de São Paulo. O metabuscador, em 21 de março de 2022, mostrava nove instituições com seus respectivos repositórios de dados, somando um total de 588 conjuntos de dados de pesquisa. Entretanto, para a amostra, foram considerados apenas os repositórios institucionais diretamente ligados às instituições de ensino superior, excluindo, portanto, o repositório da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) e da FAPESP COVID-19 *Data Sharing*/BR. O repositório de dados 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 dados depositados poderiam se encontrar em diferentes estágios de "*FAIRness*" (WILKINSON *et al*., 2016). Para a verificação da aderência dos dados aos princípios FAIR foi utilizada uma ferramenta auxiliar, a *F-UJI Automated FAIR Data Assessment Tool*[8], um serviço *web* para avaliar programaticamente o nível de aderência dos dados. A ferramenta é baseada em *Representational State Transer* (REST) para a avaliação programática e automatizada do "*FAIRness*" dos conjuntos de dados de pesquisa. O uso da F-UJI foi importante para a avaliação à luz do FAIR porque, como explicam Henning *et al.* (2019a, p. 190):

> Colocá-los em prática ainda é uma tarefa árdua em função do seu alto grau de subjetividade e complexidade, uma questão que vem sendo tratada pelas métricas FAIR. Essas métricas são consideradas

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

> ferramentas úteis para medir o nível de *FAIRness* de um recurso digital e, portanto, garantir a aderência do recurso digital a esses princípios.

A ferramenta F-UJI foi desenvolvida pela *Fostering Fair Data Practices in Europe* – FAIRsFAIR, que visa fornecer soluções práticas para o uso dos princípios FAIR ao longo do ciclo de vida dos dados de pesquisa. O projeto desenvolve padrões globais para a certificação FAIR de repositórios e dos dados neles depositados, sendo um projeto de extrema relevância no contexto deste estudo e no cenário internacional (DEVARAJU *et al.*, 2021), contando com financiamento da *European Union's Horizon 2020*. Além da questão citada da credibilidade da ferramenta, o fato dela realizar a avaliação de forma automática foi fundamental para a sua escolha, uma vez que a amostra continha um total de 264 conjuntos de dados, o que dificultaria a avaliação manual.

Através do *Uniform Resource Locator* (URL) do conjunto de dados, a ferramenta foi capaz de avaliar o nível (inicial, moderado, avançado ou incompleto) em que ele se encontrava quanto a cada um dos princípios FAIR. Além de entregar também uma porcentagem geral de *FAIRness*. Todos esses dados (percentual e nível de aderência) foram entregues automaticamente pela ferramenta F-UJI, com métricas baseadas nos indicadores propostos pelo *RDA FAIR Data Maturity Model Working Group* e no *WDS/RDA Assessment of Data Fitness for use checklist*. As métricas, os métodos, as escalas e o código podem ser consultados em detalhes no *site* da ferramenta F-UJI.

Dessa forma, todos os 264 conjuntos de dados que estavam depositados nos seis repositórios foram individualmente avaliados pela ferramenta auxiliar, no período entre janeiro e março de 2022. Após a avaliação, os dados coletados através da ferramenta foram compilados em planilhas para análises e comparações, o que possibilitou apontar alguns pontos fortes e fracos dos repositórios. Sendo assim, o escopo deste trabalho limitou-se aos resultados encontrados a partir da ferramenta citada, contribuindo para ampliar os estudos no tema e formulando recomendações para os repositórios da amostra. Assim será possível traçar estratégias para melhorar suas infraestruturas e atender às demandas dos princípios FAIR.

Numa última fase, foi feita uma verificação adicional dos padrões de metadados adotados pelos repositórios de dados de pesquisa, bem como os elementos de metadados utilizados na descrição dos recursos. Buscou-se: 1) verificar se o *Dublin*

Core era, de fato, o padrão mais adotado para a descrição de dados de pesquisa, conforme citam Simionato (2017) e Sanchez, Silva e Vechiato (2019) em estudo que levantou os padrões de metadados mais utilizados mundialmente em repositórios de dados de pesquisa; e 2) propor um conjunto mínimo de metadados para a boa representação dos dados de pesquisa em repositórios.

Para a fundamentação teórica, esta pesquisa se pautou em buscas na: 1) Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação (BRAPCI); na 2) Biblioteca Eletrônica Científica Online (SciELO); nos 3) periódicos da área de Ciência da Informação no Portal de Periódicos da CAPES; e no 4) Google Acadêmico.

Os termos de busca foram “dados de pesquisa/*research data*”, “gestão de dados de pesquisa/*research data management*”, "*e-Science*", “repositório de dados de pesquisa/*research data repository*”, "princípios FAIR/*FAIR principles*", "boas práticas/*good practices*". Para todos os termos derivados do inglês "*research data*" foi feita uma busca adicional usando "dados científicos", uma vez que a literatura nacional o adota como sinônimo de "dados de pesquisa". As buscas foram realizadas entre maio e setembro de 2021, sem o uso de filtros adicionais. O material selecionado consistiu, principalmente, em publicações nacionais e internacionais a partir do ano de 2014, quando foi realizada a conferência internacional '*Jointly designing the data FAIRPORT*', marco importante das discussões quanto às boas práticas ligadas aos dados de pesquisa.

Para a execução da pesquisa, foi desenvolvido um plano de trabalho. Assim, as sete etapas podem ser vistas, de forma resumida, no Quadro 4 abaixo. Buscou-se contemplar as principais questões estabelecidas nos objetivos, para que pudessem ser localizadas as contribuições científicas pertinentes ao tema.

QUADRO 4 – Plano de trabalho

ETAPAS	ATIVIDADES
1ª etapa: Levantamento bibliográfico e seleção do material	Seleção das publicações para o embasamento teórico da temática, por meio do levantamento bibliográfico realizado em nível nacional e internacional
2ª etapa: Análise das publicações selecionadas e elaboração do referencial teórico	Leitura, interpretação, análise e sistematização das informações

3ª etapa: Avaliação dos dados de pesquisa depositados nos repositórios institucionais quanto aos princípios FAIR	Avaliação individual de todos os conjuntos de dados de pesquisa depositados nos repositórios institucionais de São Paulo mapeados no metabuscador da FAPESP. Avaliação automática realizada pela ferramenta F-UJI, que entrega os resultados para análise
4ª etapa: Sistematização e análise dos resultados da 3ª etapa	Sistematização e análise dos resultados entregues pela ferramenta F-UJI em planilhas no Google Planilhas, permitindo criar gráficos, comparar números e médias para cada um dos princípios. Nessa etapa houve também a manutenção dos resultados obtidos para futuro compartilhamento dos dados coletados
5ª etapa: Verificação da padronização na escolha do esquema de metadados entre os repositórios de dados de pesquisa	Consulta manual no próprio repositório, nos metadados dos itens, para a identificação do esquema de metadados utilizado
6ª etapa: Indicação de um conjunto mínimo de metadados para a representação dos dados de pesquisa	Análise de conjunto de dados que obteve nível avançado de descrição (metadados) pela F-UJI para sintetizar os elementos utilizados na representação deles e criar um quadro indicando metadados mínimos para a representação
7ª etapa: Elaboração da redação final e divulgação da pesquisa	Desenvolvimento do relatório textual para divulgação à comunidade científica dos resultados obtidos, além da entrega dos relatórios de *feedback* para as instituições

Fonte: Elaborado pela autora (2022).

A primeira etapa foi o levantamento bibliográfico e a seleção do material para o embasamento teórico da pesquisa. Após o levantamento, seguiu-se para a análise das publicações selecionadas e para a elaboração do referencial teórico em si. Isso deu origem aos capítulos dois e três, onde buscou-se detalhar o que são dados de pesquisa, repositórios de dados, metadados e boas práticas no cenário da *e-Science*, introduzindo os termos ao leitor.

Em seguida, foram feitas as avaliações dos dados de pesquisa depositados nos repositórios da amostra quanto aos princípios FAIR, tendo como base a ferramenta auxiliar F-UJI, que entregou os dados para análise. Esses dados foram

todos inseridos em planilhas para futuro compartilhamento, além de servir como base para as comparações e discussões trazidas no capítulo cinco. A sistematização e análise dos resultados da 3ª etapa buscou apontar pontos fortes e fracos de cada repositório em detalhes, trazendo um *feedback* para as instituições. O objetivo foi, portanto, auxiliá-las no processo de melhorias em prol de dados cada vez mais FAIR, seguindo o cenário internacional.

Por fim, houve duas etapas adicionais para trazer maiores contribuições no estado da arte dos metadados ligados aos dados de pesquisa. Primeiro foi feita uma verificação da padronização na escolha do esquema de metadados entre os repositórios de dados de pesquisa. O objetivo foi, justamente, identificar se todos os repositórios da amostra adotavam o *Dublin Core*. Esse dado foi importante para, inclusive, a próxima etapa, que dizia respeito à indicação de um conjunto de elementos de metadados mínimos para a descrição de dados de pesquisa. Essas questões também foram abordadas no capítulo cinco.

5 METADADOS, REPOSITÓRIOS DE DADOS E PRINCÍPIOS FAIR

5.1 Aderência aos princípios FAIR

Os objetivos propostos pelos princípios FAIR já se constituem como expectativas de agências e editoras internacionalmente (WILKINSON *et al*., 2016), revelando a importância de sua adoção no cenário dos dados de pesquisa. Além da comunidade científica, comunidades governamentais e iniciativas privadas de diversos países estão se juntando ao grupo que defende a disseminação e implementação dos princípios FAIR (SALES *et al*., 2019), o que demonstra a importância de se buscar adotar melhorias nos serviços nacionais de dados de pesquisa para se adequar a essa nova tendência internacional.

Entretanto, Henning *et al*. (2019a) explicitam que os princípios são recentes e podem gerar equívocos e interpretações ambíguas quando colocados em prática, levando a um grau baixo de aderência ao FAIR. Os autores afirmam que:

> [...] existe uma desigualdade nas adesões ao FAIR, entre pesquisadores, disciplinas e universidades. Isto se dá decorrente de diferentes usos disciplinares das práticas de pesquisa e diferenças culturais relacionados à abertura e ao compartilhamento dos dados. Existe, ainda, uma falta de conformidade em torno do nível de *FAIRness* (HENNING *et al*., 2019a, p. 186).

Sendo assim, este trabalho busca averiguar em que nível de conformidade com os princípios FAIR se encontram os *datasets* depositados nos repositórios institucionais da amostra. A tendência é que ainda haja vários pontos a serem melhorados, 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 (SILVA, RODRIGUES, 2021). Mas é importante que as partes interessadas invistam nesses princípios, buscando níveis cada vez maiores de *FAIRness,* já que esse investimento está ligado com dados legíveis por máquina. No contexto da *e-Science*, essa qualidade é de suma importância.

Com isso em mente, todos os conjuntos de dados depositados foram avaliados pela ferramenta F-UJI, que apresentou, na parte de resumo (Figura 1), uma porcentagem geral do nível de *FAIRness* do conjunto de dados, junto com uma nota para encontrável, acessível, interoperável e reutilizável. Essas notas (por exemplo: 6 de 7 em encontrável) determinam o nível de aderência, que varia numa escala de "incompleto, inicial, moderado e avançado" e aparece logo ao lado da nota. Todas as escalas e notas máximas foram definidas pela ferramenta F-UJI e podem ser vistas em mais detalhes em seu *site*.

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

Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

Já o relatório (Figura 2) mostra em detalhes quais aspectos foram considerados dentro de cada uma das facetas (exemplo: em encontrável, a ferramenta testa o aspecto FsF-F1-02D - foi atribuído um identificador persistente), o que possibilita entender em quais pontos os conjuntos de dados precisam de melhorias para se tornarem mais FAIR.

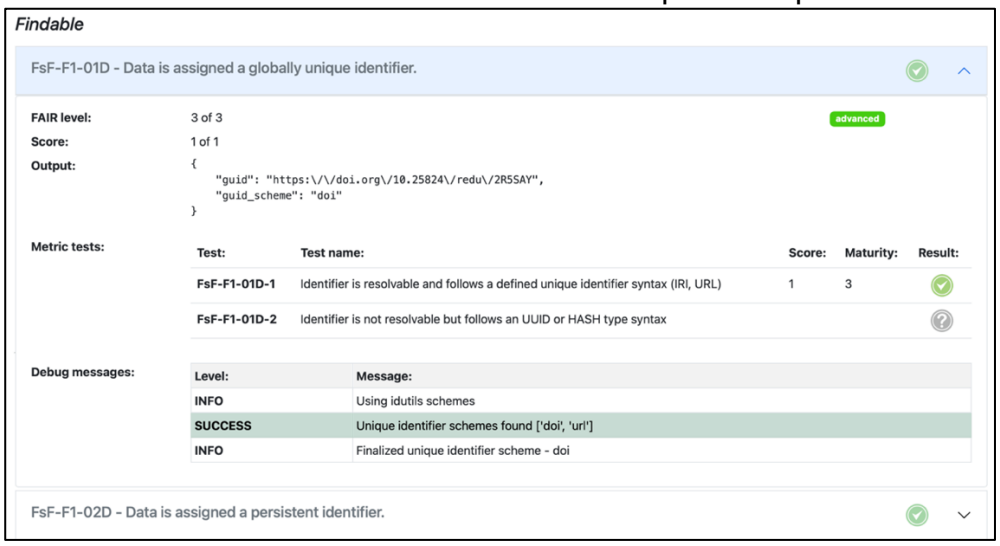
Figura 2 – Relatório da aderência aos princípios FAIR fornecido pela ferramenta F-UJI

Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

No relatório, ao clicar em um aspecto (exemplo: FsF-F1-02D - foi atribuído um identificador persistente), também é possível verificar o nível de *FAIRness* (incompleto, inicial, moderado, avançado) em que o conjunto de dados se encontra e quais métricas foram testadas para chegar àquele resultado entregue pela ferramenta F-UJI (Figura 3). Isso garante a transparência sobre o que está sendo avaliado e permite entender o que precisa ser melhorado para se atingir níveis mais altos de *FAIRness*.

Devido a esse nível de detalhamento, permitindo tanto extrair uma pontuação geral de aderência para comparação das instituições; como também um relatório das métricas testadas, tomou-se a decisão de utilizar a ferramenta como base para as avaliações presentes neste trabalho.

Figura 3 – Nível de *FAIRness* e métricas testadas para o aspecto FsF-F1-01D



Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

Portanto, é importante salientar que os resultados mostrados abaixo dizem respeito aos dados entregues pela ferramenta, que apesar do nível de detalhamento, ainda é baseada em dados e códigos preliminares que estão em desenvolvimento, buscando sempre melhorias. Isso pode levar a resultados diferentes ao longo do tempo, uma vez que as métricas podem ser atualizadas, levando a mudanças nas pontuações dos *datasets*. Além disso, a ferramenta avalia todos os conjuntos de dados de maneira equivalente, sem levar em conta especificidades. Henning *et al.*

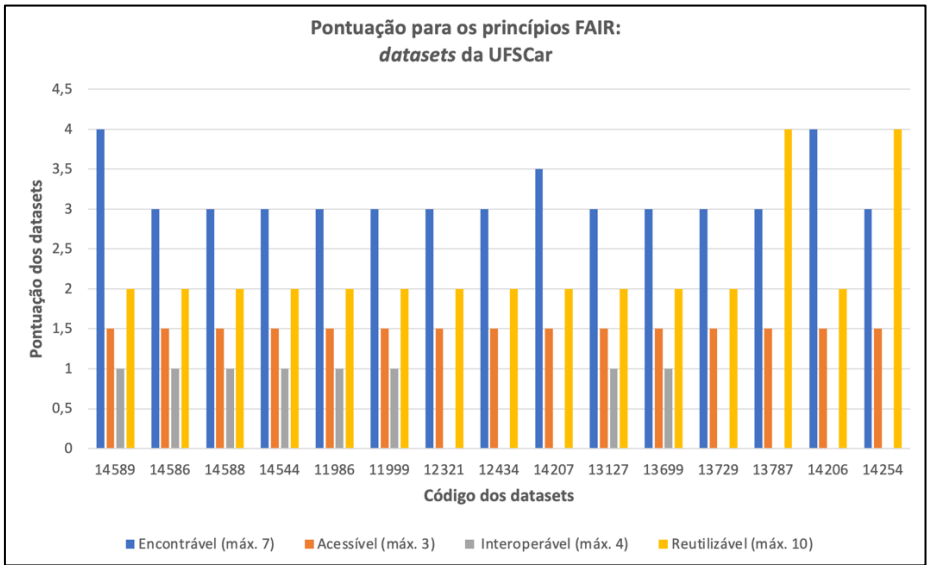
(2019a, p. 190) alertam que "[...] o grande desafio está em respeitar as particularidades de cada área do conhecimento, sendo capaz de acomodar os requisitos de dados específicos de cada domínio, ao invés de utilizar as métricas como um padrão único".

Abaixo será possível verificar e comparar os resultados para os conjuntos de dados depositados em cada um dos seis repositórios da amostra. É importante frisar que algumas recomendações podem não ser aplicáveis ao contexto do repositório e cabe à equipe gestora determinar quais pontos elencados neste trabalho são relevantes para futuras melhorias, pensando nas necessidades da sua coleção e comunidade. O objetivo é, portanto, trazer um *feedback* para ajudar as equipes a identificar onde seus repositórios podem implementar mudanças em busca de dados de pesquisa mais FAIR.

5.1.1 Repositório institucional da Universidade Federal de São Carlos (UFSCar)

A primeira avaliação diz respeito aos 15 conjuntos de dados (*datasets*) depositados no repositório institucional da UFSCar. Cada princípio possui uma nota máxima, conforme o número indicado na legenda do gráfico: encontrável (máximo sete), acessível (máximo três), interoperável (máximo quatro) e reutilizável (máximo dez). As notas, conforme citado anteriormente, foram estipuladas pela ferramenta F-UJI com base nas suas métricas.

Gráfico 1 - Pontuação para os princípios FAIR: *datasets* da UFSCar



Fonte: elaborado pela autora (2022).

É possível observar, pelo Gráfico 1, que nenhum dos *datasets* do repositório da UFSCar conseguiu atingir nota máxima, o que se aplica à encontrabilidade, acessibilidade, interoperabilidade e reutilização. A maior nota obtida em encontrável foi 4 de 7, em acessível 1,5 de 3, em interoperável 1 de 4 e em reutilizável foi 4 de 10. As pontuações máximas foram todas definidas pela ferramenta F-UJI.

Todos os *datasets* obtiveram baixas pontuações, mas principalmente quanto às facetas da interoperabilidade e reutilização, revelando áreas que podem ser priorizadas pela instituição. Dunning, Smaele e Böhmero (2017) corroboram esse resultado ao afirmar que, em avaliação feita em 38 repositórios da Holanda, as facetas "interoperável" e "reutilizável" foram, em particular, as mais difíceis de aderir.

Os 15 *datasets* da UFSCar foram também avaliados em uma escala de 0 a 100% quanto à aderência geral aos princípios FAIR, sendo a maior pontuação igual a 35%. Apenas três dos 15 *datasets* alcançaram essa pontuação mais alta, e é possível observar que dois deles (*datasets* 13787 e 14206) pontuaram de forma equivalente, se destacando na reutilização. Nesses dois conjuntos de dados a ferramenta auxiliar conseguiu identificar nos metadados uma licença clara (sendo ela *Creative Commons*), fundamental quando se pensa no reuso. Já o terceiro *dataset* (14589) que alcançou 35% de *FAIRness* se destacou na encontrabilidade, conforme mostra o Gráfico 1.

Os metadados legíveis por máquina são essenciais para o alcance de pontuações mais altas como desses três conjuntos de dados. Não só porque permitem que ferramentas automáticas como a F-UJI avaliem e validem as características desejadas para maiores níveis de *FAIRness*, mas também porque o objetivo do FAIR é enfatizar o processamento automático das máquinas, e para isso algumas sintaxes e padrões precisam ser adotados. É importante, portanto, que as equipes gestoras dos repositórios avaliem essa prática.

Em contrapartida, a menor pontuação alcançada dentre os 15 *datasets* foi de 27% de aderência, verificada para três conjuntos de dados (*datasets* 12321, 12434 e 13729). Em comum, nenhum deles possuía metadados que incluíssem *links* entre os conjuntos de dados e suas entidades relacionadas, prejudicando a interoperabilidade. Ou seja, não havia menção explícita aos recursos relacionados nos metadados, o que é um ponto importante quando se pensa na Web Semântica, por exemplo, com dados conectados fornecendo informações para as máquinas. Logo, num âmbito geral, percebe-se que o percentual alcançado quanto ao nível de *FAIRness* pelos dados de

pesquisa da UFSCar é baixo, indo de acordo com o que é defendido por Henning *et al*. (2019a) quanto à baixa conformidade esperada.

Indo para uma análise mais minuciosa, a ferramenta apresenta o nível (inicial, moderado, avançado ou incompleto) de conformidade quanto a cada princípio FAIR. Dessa forma é possível verificar em quais implementações o repositório pode investir e o que pode ser feito, na prática, para melhorar essas pontuações. No Quadro 5 tem-se os resultados para a **encontrabilidade** dos dados avaliados.

Quadro 5 - Aderência quanto à Encontrabilidade (UFSCar)

Código	FsF-F1-01D	FsF-F1-02D	FsF-F2-01M	FsF-F3-01M	FsF-F4-01M
14589	avançado	avançado	moderado	incompleto	avançado
14586	avançado	incompleto	moderado	incompleto	avançado
14588	avançado	incompleto	moderado	incompleto	avançado
14544	avançado	incompleto	moderado	incompleto	avançado
11986	avançado	incompleto	moderado	incompleto	avançado
11999	avançado	incompleto	moderado	incompleto	avançado
12321	avançado	incompleto	moderado	incompleto	avançado
12434	avançado	incompleto	moderado	incompleto	avançado
14207	avançado	inicial	moderado	incompleto	avançado
13127	avançado	incompleto	moderado	incompleto	avançado
13699	avançado	incompleto	moderado	incompleto	avançado
13729	avançado	incompleto	moderado	incompleto	avançado
13787	avançado	incompleto	moderado	incompleto	avançado
14206	avançado	avançado	moderado	incompleto	avançado
14254	avançado	incompleto	moderado	incompleto	avançado

Fonte: elaborado pela autora (2022).

O primeiro aspecto (FsF-F1-01D), que diz respeito a: "são atribuídos identificadores globalmente exclusivos aos dados", foi avaliado como "avançado" para todos os conjuntos de dados da UFSCar. A ferramenta auxiliar verificou inicialmente se o identificador atribuído segue uma sintaxe de identificador único definido (IRI, URL) e, de fato, todos os *datasets* cumprem o requisito. Esse resultado é importante

porque não se pode acessar, interoperar ou reutilizar dados sem conseguir identificá-los (JUTY *et al*., 2020). Logo, é essencial que os *datasets* tenham nomes únicos e sejam encontrados por associação do nome único com um protocolo de recuperação.

Já o segundo aspecto (FsF-F1-02D) diz respeito a: "são atribuídos identificadores persistentes aos dados", em que apenas dois *datasets* obtiveram nota "avançada", um obteve "inicial" e todos os demais obtiveram "incompleto". A ferramenta buscou por esquemas baseados em *process ID* (PID) como o *Digital Object Identifier* (DOI), que foi identificado para os dois *datasets* mais bem avaliados. Era preciso que o identificador do *dataset* seguisse uma sintaxe de identificador persistente definida, o que a ferramenta teve dificuldade de encontrar na maioria dos conjuntos de dados avaliados.

Juty *et al*. (2020) afirmam que existem quatro tipos comuns de identificadores FAIR: DOI, *Archival Resource Key* (ARK), Identifiers.org e *persistent uniform resource locator* (PURL), que não foram encontrados para a maioria dos *datasets* da UFSCar. O DOI, por exemplo, "[...] consiste em uma sequência de caracteres única que identifica uma entidade em um ambiente digital - em outras palavras, identifica o próprio objeto e não o local onde está localizado" (NEUMANN; BRASE, 2014, p. 1036, tradução nossa).

Dessa forma, mesmo que o objeto (*dataset*) seja movido para outro local, ainda será possível identificar e localizar o objeto de forma persistente. Isso é importante porque, conforme explica Sayão (2007, p. 68), "Embora um dado recurso tenha sido transferido ou perdido seu valor no âmbito da organização proprietária, na perspectiva do usuário o mesmo recurso pode continuar a ter valor permanente como referência científica". Entretendo, a atribuição de DOI é paga, e depende do serviço que é adquirido pela instituição, o que pode ser uma barreira para sua adoção nos repositórios da amostra.

Os identificadores persistentes são amplamente debatidos e cobrados internacionalmente na gestão de dados de pesquisa. Hodson *et al*. (2018) afirmam que os identificadores permitem que ligações persistentes sejam estabelecidas entre os dados e metadados e outros materiais relacionados, o que também auxilia na encontrabilidade e interoperabilidade dos recursos. Mas Marín-Arraiza e Heredia (2021, p. 29) explicam que "[...] a persistência está relacionada ao serviço oferecido pelo sistema e não ao identificador em si", e o uso dos PIDs é crucial para a preservação dos dados.

Logo, recomenda-se que o repositório busque investir, com certa prioridade, no uso de identificadores como o DOI, em que os metadados anexos estão disponíveis sob uma licença *Creative Commons* 0, aberta a todos. Isso garante que os conjuntos de dados sejam encontrados, interpretados e devidamente citados (o que aumenta também sua reusabilidade).

O terceiro aspecto avaliado (FsF-F2-01M) se refere a: "metadados incluem elementos essenciais (*'creator', 'title', 'publisher', 'publication_date', 'summary', 'keywords', 'object_identifier', 'object_type'*) para dar suporte à encontrabilidade dos objetos digitais". Os princípios FAIR dão atenção especial aos metadados, e todos os princípios se relacionam com metadados em pelo menos um aspecto.

Juty *et al*. (2020) afirmam que os conjuntos de dados devem ter pelo menos alguns metadados descritivos associados para auxiliar na descoberta e verificação quando o identificador não for conhecido *a priori.* É possível observar, para a maioria dos conjuntos de dados depositados, os criadores, o título, o publicador, a data de publicação, a descrição do recurso, palavras-chave, o seu identificador e o tipo, elementos que foram validados pela ferramenta, conforme mostra a Figura 4.

Figura 4 – *Output* exibido pela ferramenta F-UJI ao avaliar um conjunto de dados da UFSCar no aspecto FsF-F2-01M

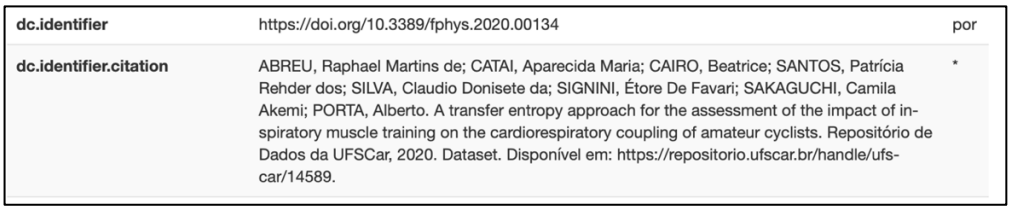
Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

É possível ainda, no registro, identificar metadados adicionais como citação, agência financiadora da pesquisa, língua, licença, programa ligado à publicação (exemplo: Programa de Pós-Graduação em Fisioterapia) e até o currículo *lattes* dos autores. Ou seja, há um bom nível de descrição dos recursos, permitindo que eles sejam encontrados na *web* (e devidamente citados).

Segundo Henning *et al*. (2019a, p. 178), os *datasets* “[...] devem possuir metadados ricos o suficiente para que, uma vez indexados para um mecanismo de busca, esses metadados possam ajudar o usuário dos dados a encontrá-los". E o *DSpace*, infraestrutura de *software* adotada pela UFSCar, permite a pesquisa filtrada (facetada) e a navegação em todos os objetos. É possível pesquisar texto completo e campos de metadados, além de ser indexado no *Google Scholar*, o que aumenta a encontrabilidade dos dados, um bom indicativo em prol de dados FAIR.

Outro bom indicativo é que a ferramenta verificou que os metadados essenciais de citação estavam presentes, o que permite que os *datasets* sejam reutilizados e devidamente citados por outros pesquisadores, tornando o conteúdo científico mais visível. A possibilidade de citá-los é de suma importância na popularização do seu reuso, incentivando os pesquisadores a adotarem a prática sem receio de infringir os direitos autorais. Inclusive, o repositório da UFSCar oferece a citação dos conjuntos de dados já pronta, permitindo que o pesquisador apenas copie, conforme mostra a Figura 5 abaixo:

Figura 5 – Citação pronta disponibilizada nos metadados do *dataset* no repositório da UFSCar



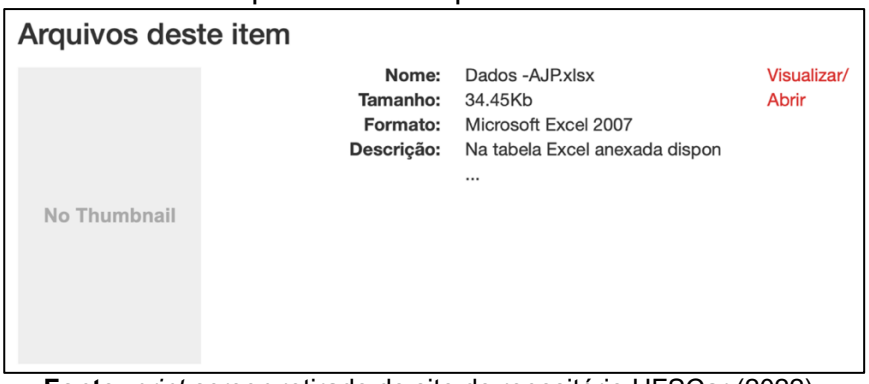
dc.identifier	https://doi.org/10.3389/fphys.2020.00134	por
dc.identifier.citation	ABREU, Raphael Martins de; CATAI, Aparecida Maria; CAIRO, Beatrice; SANTOS, Patrícia Rehder dos; SILVA, Claudio Donisete da; SIGNINI, Étore De Favari; SAKAGUCHI, Camila Akemi; PORTA, Alberto. A transfer entropy approach for the assessment of the impact of inspiratory muscle training on the cardiorespiratory coupling of amateur cyclists. Repositório de Dados da UFSCar, 2020. Dataset. Disponível em: https://repositorio.ufscar.br/handle/ufscar/14589.	*

Fonte: *print screen* retirado do site do repositório da UFSCar (2022).

Apesar disso, a ferramenta não encontrou metadados acessíveis por meio de "*typed links*" ou "*signposting links*". Esses termos dizem respeito às sinalizações e ligações que tornam a *web* acadêmica mais amigável para as máquinas. Como visto anteriormente, ser legível por máquina é um dos pontos mais importantes quando se fala sobre princípios FAIR. Ao visitar portais acadêmicos, os leitores podem facilmente descobrir *links* para registros bibliográficos, autoria, etc. Mas, como os portais usam convenções diferentes para transmitir esses padrões, as máquinas têm dificuldade em se orientar. E por isso a implementação de padrões de sinalização ganha um papel de destaque, permitindo que as máquinas naveguem por portais acadêmicos de maneira uniforme. Por ser um aspecto que demanda conhecimento técnico, pode explicar a falta de aderência dos *datasets*.

O próximo aspecto (FsF-F3-01M) diz respeito a: "os metadados incluem o identificador do dado que descrevem". A ferramenta avaliou duas características: 1) os metadados contêm informações relacionadas ao conteúdo dos dados (nome do arquivo, tamanho, tipo) e 2) os metadados contêm um PID ou URL que indica a localização do conteúdo dos dados para *download*. Mas o identificador dos dados foi tido como ausente e, por isso, todos os *datasets* foram avaliados como "incompleto". O "tamanho do arquivo", por exemplo, não foi recuperado, o que pode estar associado à falta da indicação adequada num metadado legível por máquina, uma vez que é possível identificar essa informação no registro (Figura 6).

Figura 6 – Visualização do nome, tamanho e formato do arquivo de um conjunto de dados depositado no repositório da UFSCar

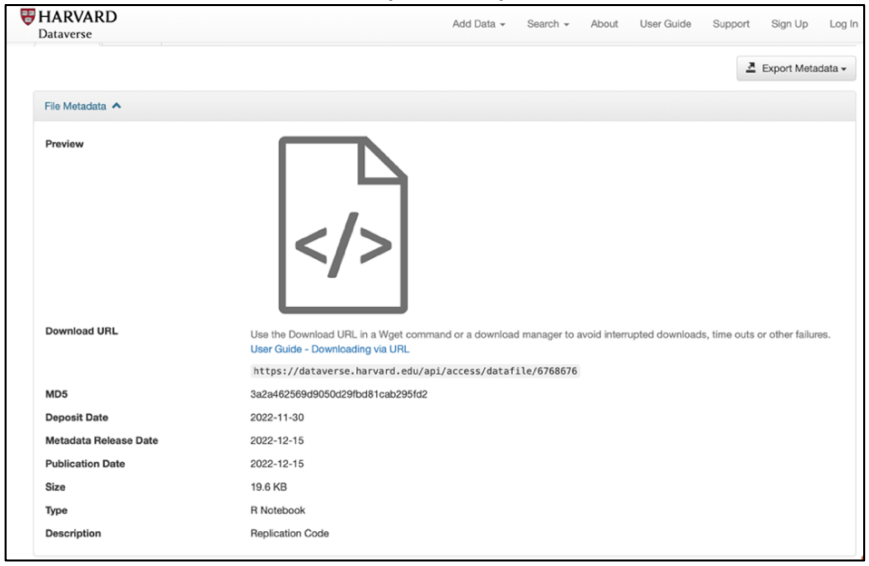


Fonte: *print screen* retirado do site do repositório UFSCar (2022).

Fazendo uma comparação, ao analisar um conjunto de dados (https://doi.org/10.7910/DVN/WR4S9I) depositado no *Harvard Dataverse*, repositório que é citado no documento oficial dos princípios FAIR, é possível observar os arquivos e seus respectivos metadados, conforme mostra a Figura 7.

Esse conjunto de dados do *Harvard Dataverse* obteve uma aderência de 83% aos princípios FAIR quando avaliado pela ferramenta F-UJI, e foi possível recuperar (em nível avançado) o nome, tamanho e o tipo do arquivo, podendo ser uma base de estudo para melhorias nos repositórios da amostra. Apesar do repositório da UFSCar também possuir os arquivos do item, não é possível acessar o registro de metadados do arquivo em si, apenas do *dataset* no geral, diferente do caso do *Harvard Dataverse.*

Figura 7 – Metadados de um arquivo depositado no *Harvard Dataverse*



Fonte: *print screen* retirado do repositório *Harvard Dataverse* (2022).

Para ser encontrável é preciso, portanto, uma descrição clara dos arquivos contidos no conjunto de dados, fornecendo as informações necessárias nos metadados adequados, de tal forma que tanto um usuário humano quanto um agente computacional possam extrair todas essas informações sobre o conjunto de dados e interpretá-las corretamente. Esse é um ponto que o repositório da UFSCar pode aplicar melhorias, mas é preciso verificar até que nível o *DSpace*, *software* utilizado pela instituição, permite esse nível de descrição.

A própria escolha de infraestruturas de *software* voltadas à gestão e curadoria dos dados de pesquisa sob a ótica dos princípios FAIR é de suma importância, e as investidas recentes em repositórios institucionais têm se concentrado no *DSpace* e *Dataverse*. Nesse sentido, Rocha *et al*. (2018, p. 74) afirmam que:

> O *Dataverse* possui recursos para configuração de vários tipos de ambientes de repositório, incluindo hierarquias organizacionais e políticas de gestão distintas para unidades ou grupos, incluindo esquemas de metadados e licenças. Isso é possível em *DSpace*, entretanto, exige adaptações ou configurações, com algumas limitações no controle de versões.

Por fim, no aspecto "FsF-F4-01M", os *datasets* foram testados quanto aos metadados fornecidos de uma maneira que os principais mecanismos de pesquisa pudessem inseri-los em seus catálogos (exemplos: JSON-LD, *Dublin Core*, RDFa), e

todos foram avaliados como “avançado”. Isso é um bom indicativo para sua encontrabilidade e para a visibilidade da instituição.

Para maior elucidação, o *JavaScript Object Notation* (JSON) é um formato compacto, de padrão aberto independente, de troca de dados simples e rápida entre sistemas, no formato atributo-valor. Ele permite a troca de informações no formato texto entre sistemas independente da linguagem de programação e é derivado do *JavaScript*. Ou seja, favorece a encontrabilidade pelas máquinas, auxiliando também na interoperabilidade entre os sistemas. Já o JSON-LD (LD equivale a *Linked Data*) é um formato que permite a ligação de dados (dados vinculados), como preconiza o FAIR, sendo leve e facilmente interpretado por humanos e máquinas, o que é ideal para a interoperabilidade dos dados.

O *Resource Description Framework* (RDF) é um modelo padrão para intercâmbio de dados na *web,* usando *Uniform Resource Identifier* (URI) para nomear o relacionamento entre as coisas, bem como as extremidades do *link*. Assim dados estruturados e semiestruturados podem ser misturados e compartilhados. Mas o JSON e o RDF serão vistos em mais detalhes na seção “Interoperável” deste trabalho.

A ferramenta rodou vários testes para o aspecto "FsF-F4-01M" como: busca por dados estruturados (o que foi validado); busca por um identificador listado no cache de *Google Dataset Search* (não validado)*;* busca por um identificador listado no catálogo *Mendeley Data* (não validado). Após os testes a ferramenta F-UJI alegou que os metadados não foram encontrados através de registros considerados pelo serviço de avaliação, sendo eles ‘*Datacite Search'* e *'DataCite Registry'*, e a equipe gestora pode avaliar a pertinência dessa aplicação.

O *DataCite Search* oferece uma interface de busca integrada, onde é possível pesquisar, filtrar e extrair todos os detalhes de uma coleção de registros. Os metadados (através do DOI) são usados para um grande índice de dados de pesquisa que podem ser consultados diretamente para encontrar dados, obter estatísticas e explorar conexões. Todos os metadados são gratuitos para acesso. Sendo assim, é um portal para encontrar, acessar e reutilizar dados. Esse é, então, outro ponto que garante maior visibilidade aos *datasets* do repositório*,* que poderão ser recuperados por outros mecanismos de busca na *web*. Entretanto, os conjuntos de dados foram avaliados como avançado no aspecto FsF-F4-01M, e pode ser dada prioridade para outros pontos do FAIR.

Quanto à **acessibilidade**, os conjuntos de dados foram avaliados em três aspectos: 1) Os metadados contêm o nível de acesso e as condições de acesso aos dados (FsF-A1-01M); 2) Os dados são acessíveis por meio de um protocolo de comunicação padronizado (FsF-A1-03D) e 3) Os metadados são acessíveis por meio de um protocolo de comunicação padronizado (FsF-A1-02M). Segundo Dias, Anjos e Rodrigues (2019, p.183), a acessibilidade se relaciona ao "[...] fato de que os dados deveriam poder ser acessados por protocolos-padrões". Os resultados podem ser vistos no Quadro 6.

Quadro 6 - Aderência quanto à Acessibilidade (UFSCar)

Código	FsF-A1-01M	FsF-A1-03D	FsF-A1-02M
14589	inicial	incompleto	avançado
14586	inicial	incompleto	avançado
14588	inicial	incompleto	avançado
14544	inicial	incompleto	avançado
11986	inicial	incompleto	avançado
11999	inicial	incompleto	avançado
12321	inicial	incompleto	avançado
12434	inicial	incompleto	avançado
14207	inicial	incompleto	avançado
13127	inicial	incompleto	avançado
13699	inicial	incompleto	avançado
13729	inicial	incompleto	avançado
13787	inicial	incompleto	avançado
14206	inicial	incompleto	avançado
14254	inicial	incompleto	avançado

Fonte: elaborado pela autora (2022).

Nota-se, pelo Quadro 6, que todos os conjuntos de dados aderiram ao primeiro e terceiro aspecto, apesar de em níveis diferentes. Para o aspecto FsF-A1-01M, a ferramenta auxiliar verificou se os *datasets* possuíam informações sobre restrições ou direitos de acesso nos metadados, recuperando uma licença *Creative Commons* (CC)

associada no registro. É de extrema importância os repositórios aderirem a essa prática, inclusive com os dados de pesquisa, pois assim é possível identificar se o acesso é público, embargado ou restrito.

Apesar de localizar a licença CC nos metadados, a ferramenta F-UJI alegou que as informações de acesso aos dados não são legíveis por máquina, um aspecto fundamental para a aderência aos princípios FAIR. Novamente, os dados precisam ser contextualizados tanto para os usuários humanos quanto para os computacionais (que dependem de uma semântica básica para interpretar dados) para que se fale em boas práticas.

Na publicação original dos princípios FAIR, Wilkinson *et al*. (2016) trazem a questão do acesso destacando que o "A" não significa necessariamente 'aberto' ou 'livre'. Na verdade, implica que se deve fornecer as condições exatas de acesso aos dados e, portanto, dados privados também podem ser FAIR. Idealmente, a acessibilidade é especificada de forma que uma máquina possa entender automaticamente os requisitos e, em seguida, executá-los automaticamente ou alertar o usuário sobre eles.

Segundo Hodson *et al*. (2018), é importante buscar sempre uma proteção proporcional que maximize a possibilidade de os dados serem disponibilizados para reutilização em um contexto que avance a pesquisa e a inovação. "Qualquer proteção deve ser ponderada pensando nos benefícios econômicos do compartilhamento de dados e o impacto econômico dos repositórios de dados, para os quais há evidências consideráveis em um amplo número de domínios" (HODSON *et al*., 2018, p. 19, tradução nossa).

Muitas vezes faz sentido solicitar aos usuários que criem uma conta para um repositório, o que é visto em vários repositórios institucionais no Brasil. Isso permite autenticar o proprietário (ou colaborador) de cada conjunto de dados e potencialmente definir direitos específicos do usuário. Portanto, esse critério também precisa ser levado em conta pelas equipes gestoras dos repositórios, decidindo se é possível, por exemplo, publicar dados com embargo ou exigir procedimentos de autenticação e autorização.

Essa questão, inclusive, também é vista nas Boas práticas para dados na Web, da W3C, que defende que as políticas de compartilhamento de dados avaliem o risco de exposição e estabeleçam as medidas de segurança apropriadas para a proteção de dados sensíveis, tais como a autenticação segura e a autorização. Isso porque,

segundo Lóscio, Burle e Calegari (2017), quando se fala em dados sensíveis é preciso lembrar que às vezes seu compartilhamento aberto pode ser seguro, principalmente em ambientes controlados, mas outras vezes esses dados (oriundos de múltiplas fontes) podem ser combinados e permitir a identificação inadvertida dos indivíduos, ferindo a Lei 13.709/2018 – Lei Geral de Proteção de Dados Pessoais (LGPD).

Essa lei afirma que as atividades que tratam de dados pessoais devem observar a "segurança: utilização de medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão". A falta dessa funcionalidade pode levar pesquisadores a temerem depositar seus dados no repositório da instituição, devido à falta de amparo legal.

Mas, no manual de autodepósito de dados de pesquisa da UFSCar, é possível observar a opção de publicar dados com embargo, que consiste no período entre o depósito e a liberação para acesso público. Esse período deve ser indicado no caso de os dados necessitarem de restrição de acesso, conforme explicado. Ou seja, é possível fazer essa proteção no repositório, conforme mostra a Figura 8. Cabe destacar ainda que, mesmo após o *login* como estudante da UFSCar, não foi possível acessar o conjunto de dados, e a seguinte mensagem foi exibida: "Você não possui as permissões para acessar este item hdl:ufscar/14945. Você não possui direitos de acesso para visualizar este item. Por favor contate o administrador do sistema".

Figura 8 – Item restrito no repositório da UFSCar

Fonte: *print screen* retirado do repositório da UFSCar (2022).

Apesar desse ponto positivo para a segurança dos dados de pesquisa depositados, é importante refrisar que a ferramenta F-UJI alegou não encontrar as

informações de acesso aos dados do repositório de forma legível por máquina. Na publicação original dos princípios FAIR é possível examinar essa questão, por exemplo, no A2: “Os metadados devem estar acessíveis mesmo quando os dados não estiverem mais disponíveis”. Lóscio, Burle e Calegari (2017, não paginado) explicam que pode ser fornecido “[...] um documento em HTML – que dá uma explicação legível por pessoas para a indisponibilidade de dados.” ou ainda “[...] podem ser utilizados códigos de status HTTP apropriados, com mensagens personalizadas [...]”, o que permite o processamento por máquina. Logo, a equipe responsável pode avaliar essa implementação para melhorar a acessibilidade de seus recursos e dos próprios metadados (mesmo em itens restritos).

Já em relação ao segundo aspecto (FsF-A1-03D), que diz respeito aos dados serem acessíveis por meio de um protocolo de comunicação padronizado, a ferramenta examinou se os metadados incluíam um *link* para dados com base em protocolos de comunicação da *web* padronizados (*standard_data_protocol*). A maioria dos usuários da Internet recupera dados ‘clicando em um *link*’. Esta é uma interface de alto nível para um protocolo de baixo nível chamado *Transmission Control Protocol* (TCP), que o computador executa para carregar dados no navegador *web* do usuário.

O *Hyper Text Transfer Protocol Secure* (HTTPS) ou *File Transfer Protocol* (FTP), que formam a espinha dorsal da internet moderna, são construídos sobre TCP e tornam a solicitação e o fornecimento de recursos digitais substancialmente mais fáceis do que outros protocolos de comunicação. O “acessível” indica que a recuperação de dados FAIR deve ser mediada sem ferramentas proprietárias de comunicação. Este aspecto se concentra em como os dados e metadados podem ser recuperados de seus identificadores.

Todos os conjuntos de dados foram avaliados como incompletos, mas são identificados por *links* que começam com https://, ou seja, são recuperados usando um protocolo de comunicação padrão, o HTTPS. Essa questão também foi, inclusive, testada para os metadados (FsF-A1-02M), e a ferramenta F-UJI conseguiu localizar o protocolo padrão para acesso aos metadado: HTTPS. Juty *et al.* (2020) explicam que um URI HTTP fornece um protocolo para encontrar e acessar o recurso de informação ao qual se refere. O “acessível” do FAIR está incorporado em identificadores seguros baseados em HTTPS URI.

A ferramenta F-UJI também avaliou os *datasets* quanto à **interoperabilidade**, que segundo Henning *et al.* (2019a), pode representar o maior desafio dentro do FAIR,

pois exige uma compreensão mais sofisticada de linguagens e padrões específicos da disciplina. A necessidade de habilidades técnicas e de entendimento quanto aos padrões que devem ser adotados pode ser um dos fatores que contribui para a dificuldade de se alcançar a interoperabilidade desejada em repositórios institucionais. Com isso em mente, tem-se os resultados (Quadro 7) quanto aos três aspectos avaliados: 1) Os metadados são representados usando uma linguagem de representação de conhecimento formal (FsF-I1-01M); 2) Metadados usam recursos semânticos (FsF-I1-02M) e 3) Os metadados incluem *links* entre os dados e suas entidades relacionadas (FsF-I3-01M).

Quadro 7 - Aderência quanto à Interoperabilidade (UFSCar)

Código	FsF-I1-01M	FsF-I1-02M	FsF-I3-01M
14589	incompleto	inicial	avançado
14586	incompleto	inicial	avançado
14588	incompleto	inicial	avançado
14544	incompleto	inicial	avançado
11986	incompleto	inicial	avançado
11999	incompleto	inicial	avançado
12321	incompleto	inicial	incompleto
12434	incompleto	inicial	incompleto
14207	incompleto	inicial	incompleto
13127	incompleto	inicial	avançado
13699	incompleto	inicial	avançado
13729	incompleto	inicial	avançado
13787	incompleto	inicial	incompleto
14206	incompleto	inicial	incompleto
14254	incompleto	inicial	incompleto

Fonte: elaborado pela autora (2022).

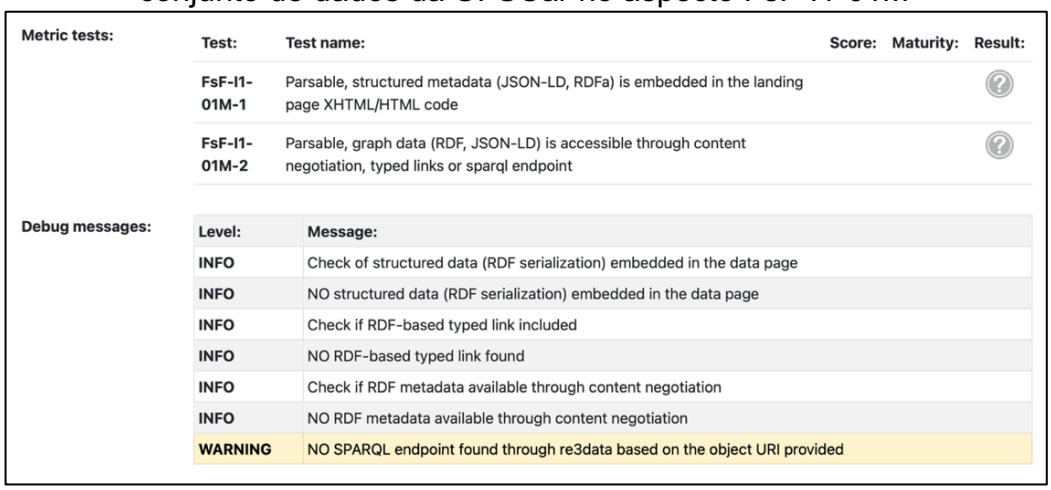
A interoperabilidade busca otimizar a comunicação entre sistemas e a integração de conjuntos de dados diferentes. Segundo Vidotti, Torino e Coneglian (2021, p. 212), para alcança-la “[...] os dados e os metadados precisam ser legíveis e

adequados a padrões e vocabulários reconhecidos, potencializando a ligação com outros padrões e incluir referências qualificadas". Os vocabulários controlados utilizados para descrever os conjuntos de dados precisam ser documentados e resolvíveis, com o uso de identificadores globalmente únicos e persistentes. Essa documentação precisa ser facilmente encontrada e acessível por qualquer pessoa/máquina que use o conjunto de dados.

Nenhum dos conjuntos de dados foi validado quanto ao primeiro aspecto, que diz respeito a uma linguagem de representação de conhecimento formal. Os *datasets* foram testados quanto à existência de metadados estruturados e analisáveis (JSON-LD, RDFa) incorporados ao código XHTML/HTML da página de destino. Segundo Löffler *et al*. (2021), a ideia que vem sendo difundida é adicionar informações descritivas aos dados estruturados, como *Extensible Markup Language* (XML) ou *HyperText Markup Language* (HTML), a fim de aumentar a encontrabilidade e interoperabilidade dos dados de pesquisa.

Segundo Silva, Martins e Siqueira (2019, p. 103), o protocolo *Simple Protocol and RDF Query Language* (SPARQL) "[...] se trata de uma linguagem de consulta, que utiliza um conjunto de especificações que fornecem linguagens e protocolos para consultar e manipular o conteúdo publicado em RDF na Web". Já o RDF, como visto, é um padrão W3C para a descrição conceitual e representação das informações sobre os dados no ambiente *web*, fornecendo uma estrutura comum destinada a representar metadados (SILVA; MARTINS; SIQUEIRA, 2019).

Figura 9 – Métricas e *debug messages* exibidas pela ferramenta F-UJI ao avaliar um conjunto de dados da UFSCar no aspecto FsF-I1-01M



Metric tests:

Test:	Test name:	Score:	Maturity:	Result:
FsF-I1-01M-1	Parsable, structured metadata (JSON-LD, RDFa) is embedded in the landing page XHTML/HTML code			
FsF-I1-01M-2	Parsable, graph data (RDF, JSON-LD) is accessible through content negotiation, typed links or sparql endpoint			

Debug messages:

Level:	Message:
INFO	Check of structured data (RDF serialization) embedded in the data page
INFO	NO structured data (RDF serialization) embedded in the data page
INFO	Check if RDF-based typed link included
INFO	NO RDF-based typed link found
INFO	Check if RDF metadata available through content negotiation
INFO	NO RDF metadata available through content negotiation
WARNING	NO SPARQL endpoint found through re3data based on the object URI provided

Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

Segundo Berners-Lee (2009), o RDF é um padrão fundamental para que os dados sejam conectados na *web* de forma interoperável e processável por máquina. Quando alguém procura por um URI, é preciso fornecer informações úteis usando padrões como RDF e SPARQL. Os *datasets* não atenderam a esses requisitos, revelando uma falta de estruturação quanto aos padrões necessários para se publicar dados FAIR, conforme mostra a Figura 9.

Dias, Anjos e Rodrigues (2019, p. 183) acrescentam que para que exista interoperabilidade entre os conjuntos de dados, "[...] é importante que existam instrumentos para padronizar semanticamente os sistemas envolvidos no processo". Sayão e Sales (2021, p. 229) corroboram ao defender que isso "[...] pressupõe que dados e metadados sejam conceitualizados, expressos e estruturados por meio de padrões de ampla aceitação, publicados, rastreáveis e acessíveis (ou seja, também FAIR)". É nesse sentido que se enquadra o próximo aspecto (FsF-I1-02M), que avaliou os metadados e seus recursos semânticos. Conforme mostra o Quadro 7, todos os conjuntos de dados foram pontuados como nível inicial.

A ferramenta buscou por uso de *namespaces* de vocabulário e de recursos semânticos nos metadados. Foi testado “extrações dos metadados baseados em RDF”, que, como visto, possui recursos que facilitam a combinação de dados, mesmo que os esquemas subjacentes sejam diferentes, e suporta especificamente a evolução dos esquemas ao longo do tempo. O RDF estende a estrutura de *links* da *web* para usar URIs para nomear o relacionamento entre as coisas, e não só as coisas em si. Isso garante que os relacionamentos sejam também legíveis por máquina, fundamental para a interoperabilidade.

Silva, Martins e Siqueira (2019) explicam que o uso de ontologias em repositórios é necessário para que o significado pretendido pelo publicador do *dataset* seja o mesmo do entendido pelo usuário dos dados. Além disso, os autores explicam que “Cada ontologia é descrita por um documento apontado por uma URI ou *namespace* e as classes e as propriedades são construídas a partir da concatenação da URI com o nome da classe ou propriedade desejada” (SILVA; MARTINS; SIQUEIRA, 2019, p. 102). *Namespaces* garantem que determinado conjunto de objetos tenha nomes exclusivos para que possa ser facilmente identificado, então é um bom indicativo que a ferramenta tenha recuperado *namespaces* de vocabulário nos metadados do repositório. Apesar disso, também foi avaliada a existência de

namespaces de recursos semânticos conhecidos nos metadados, o que não foi encontrado.

A semântica é de suma importância uma vez que, conforme dito anteriormente, os princípios FAIR buscam enfatizar o aprimoramento da capacidade das máquinas de encontrar e usar os dados automaticamente. Mas, diferente dos humanos que conseguem interpretar dados por um senso intuitivo de semântica (o significado ou a intenção de um objeto digital), as máquinas precisam de auxílio através de aplicação de tecnologias e padrões de interoperabilidade (KALINAUSKAITĖ, 2017).

Quanto ao último aspecto da interoperabilidade, que diz respeito às ligações entre os dados e suas entidades relacionadas, o resultado se mostrou melhor: nove dos quinze *datasets* obtiveram nível avançado. A ferramenta F-UJI foi capaz de encontrar nesses *datasets* recursos relacionados que são mencionados explicitamente nos metadados e que são indicados por *links* ou identificadores legíveis por máquina. De acordo com o relatório da ferramenta, foi possível extrair do *Dublin Core* o relacionamento "*isRelatedTo*" com recursos relacionados ao *dataset.*

O repositório da UFSCar, como visto, tem aspectos que podem ser trabalhados para uma melhor interoperabilidade do sistema. Criar ligações entre os *datasets* e suas entidades relacionadas e investir em recursos semânticos e em metadados estruturados e analisáveis (JSON-LD, RDFa) incorporados no código XHTML/HTML da página de destino são exemplos que devem ser observados pela instituição, uma vez que a interoperabilidade está ligada à ideia central do FAIR: enfatizar o aprimoramento da capacidade das máquinas de encontrar e usar os dados automaticamente.

Por fim, todos os 15 *datasets* foram avaliados quanto à **reutilização**. De acordo com Vidotti, Torino e Coneglian (2021, p. 212) “A reutilização trata de quão bem estão descritos os dados e os metadados, incluindo informações sobre os direitos de uso, a proveniência e o contexto dos dados”. Logo, para uma boa aderência, eles deveriam ser descritos em um nível de granularidade significativo para que detalhes e atributos fossem representados; e a descrição deveria ser precisa, clara e relevante, tanto para a comunidade quanto para o domínio do repositório de dados.

Os dados foram testados em cinco aspectos pela ferramenta auxiliar: 1) Metadados especificam o conteúdo dos dados (FsF-R1-01MD); 2) Os metadados incluem informações de licença sob as quais os dados podem ser reutilizados (FsF-R1.1-01M); 3) Os metadados incluem informações de proveniência sobre a criação ou

geração dos dados (FsF-R1.2-01M); 4) Os metadados seguem um padrão recomendado pela comunidade de pesquisa-alvo dos dados (FsF-R1.3-01M) e 5) Os dados estão disponíveis em um formato de arquivo recomendado pela comunidade de pesquisa-alvo (FsF-R1.3-02D). Os resultados podem ser vistos no Quadro 8.

Quadro 8 - Aderência quanto à Reutilização (UFSCar)

Código	FsF-R1-01MD	FsF-R1.1-01M	FsF-R1.2-01M	FsF-R1.3-01M	FsF-R1.3-02D
14589	Inicial	incompleto	moderado	inicial	incompleto
14586	Inicial	incompleto	moderado	inicial	incompleto
14588	Inicial	incompleto	moderado	inicial	incompleto
14544	Inicial	incompleto	moderado	inicial	incompleto
11986	Inicial	incompleto	moderado	inicial	incompleto
11999	Inicial	incompleto	moderado	inicial	incompleto
12321	Inicial	incompleto	moderado	inicial	incompleto
12434	Inicial	incompleto	moderado	inicial	incompleto
14207	Inicial	incompleto	moderado	inicial	incompleto
13127	Inicial	incompleto	moderado	inicial	incompleto
13699	Inicial	incompleto	moderado	inicial	incompleto
13729	Inicial	incompleto	moderado	inicial	incompleto
13787	Inicial	avançado	moderado	inicial	incompleto
14206	Inicial	incompleto	moderado	inicial	incompleto
14254	Inicial	avançado	moderado	inicial	incompleto

Fonte: elaborado pela autora (2022).

O primeiro aspecto (FsF-R1-01MD) avaliou pontos como a existência de informações mínimas fornecidas nos metadados sobre o conteúdo dos dados disponíveis, o que foi localizado em nível inicial com o "*object_type*", que recebe como valor “*dataset*”. Entretanto, a ferramenta não conseguiu recuperar informações sobre o tamanho do arquivo, como já visto em encontrabilidade. O repositório até chega a fornecer o tamanho na página do item, na opção de *download*, mas não em um elemento adequado dos metadados. É preciso sempre levar em conta que os princípios FAIR buscam enfatizar o aprimoramento da capacidade das máquinas de

encontrar e usar os dados automaticamente, o que não seria possível nessa situação. É recomendado, portanto, que a equipe gestora avalie essa prática, indicando as informações em elementos de metadados adequados.

A contextualização dos dados, fornecendo o máximo de informação possível sobre suas especificações, é de suma importância para agregar valor, realizar a curadoria e permitir sua reutilização por pesquisadores. Isso porque, segundo Curty (2019, p. 179) “[...] a reusabilidade dos dados depende, em grande parcela, da qualidade associada à documentação dos dados disponibilizados para reuso”.

O segundo aspecto (FsF-R1.1-01M) diz respeito aos metadados incluírem ou não informações de licença sob as quais os dados podem ser reutilizados. Dias, Anjos e Rodrigues (2019, p. 184) explicam que “O que pode ser feito com os dados e quem pode usá-los deve ser especificado de forma bastante clara em um termo de uso.” Essa é uma informação extremamente importante para que pesquisadores possam reutilizar os dados de pesquisa, mas a ferramenta F-UJI não conseguiu recuperá-la em um elemento de metadado apropriado para a maioria dos *datasets* da UFSCar.

Também foi testado se a licença era válida e registrada no *Software Package Data Exchange* (SPDX), que permite a expressão de componentes, licenças, direitos autorais, referências de segurança e outros metadados relacionados ao *software.* Seu objetivo original era melhorar a conformidade de licenças e o SPDX está sob os auspícios da *Linux Foundation*. Mas os metadados de 13 dos 15 *datasets* não forneciam essas informações e, portanto, não foram pontuadas. Torino, Trevisan e Vidotti (2019, p. 45) explicam que:

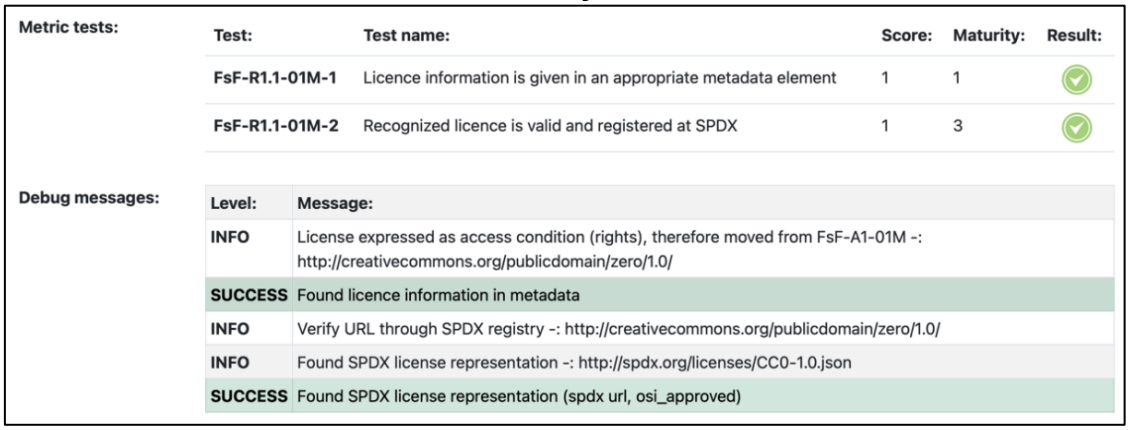
> [...] é necessário estabelecer adequadamente a licença dos dados, não apenas indicando o tipo de licença adotada, mas especificando a versão da licença e utilizando metadados para expressá-la de modo a ser adequadamente compreendida por humanos, por aplicações computacionais e pelo texto jurídico, o que assegura a veracidade e reuso sem qualquer infração à questões legais, mantendo protegidos os fornecedores e consumidores dos dados.

Os dois *datasets* que obtiveram nível avançado quanto às informações de licença tinham algo em comum: ambos indicaram no metadado "*dc.rights*" a licença com a sintaxe "https://creativecommons.org/publicdomain/zero/1.0/", em inglês, em vez de "http://creativecommons.org/licenses/by-nc-sa/3.0/br/", como os demais *datasets*. Segundo Henning *et al*. (2019b, p. 397) “Se houvesse homogeneização e convergência na descrição e licenciamento dos dados, o processo de

interoperabilidade entre as instituições poderia acontecer de forma transparente, evitando esforços de integração adicionais".

Assim, as informações da licença foram encontradas nos metadados e a consulta da ferramenta F-UJI quanto ao SPDX retornou resultado positivo, conforme mostra a Figura 10. Ou seja, recomenda-se que o repositório busque padronizar esta questão, optando por indicar o uso da licença com a sintaxe que é registrada no SPDX. É fundamental que essas informações, que incentivam a reutilização dos dados de forma legal, possam ser interpretadas por computadores e humanos.

Figura 10 – Métricas e *debug messages* exibidas pela ferramenta F-UJI no aspecto FsF-R1.1-01M para o conjunto de dados que foi validado quanto às informações de licença



Metric tests:

Test:	Test name:	Score:	Maturity:	Result:
FsF-R1.1-01M-1	Licence information is given in an appropriate metadata element	1	1	✓
FsF-R1.1-01M-2	Recognized licence is valid and registered at SPDX	1	3	✓

Debug messages:

Level:	Message:
INFO	License expressed as access condition (rights), therefore moved from FsF-A1-01M -: http://creativecommons.org/publicdomain/zero/1.0/
SUCCESS	Found licence information in metadata
INFO	Verify URL through SPDX registry -: http://creativecommons.org/publicdomain/zero/1.0/
INFO	Found SPDX license representation -: http://spdx.org/licenses/CC0-1.0.json
SUCCESS	Found SPDX license representation (spdx url, osi_approved)

Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

O próximo aspecto (FsF-R1.2-01M) avaliou informações de proveniência nos metadados, e todos os 15 *datasets* obtiveram um nível moderado de descrição. Alguns dos elementos que foram extraídos pela ferramenta foram: "*creator", "publication_date", "publisher"* e *"contributor*". Essas informações de proveniência relacionadas à criação dos dados de pesquisa foram encontradas, mas a ferramenta não conseguiu validar a presença, nos metadados, de informações de proveniência usando ontologias formais, o que impossibilitou a pontuação dos *datasets* como avançado.

Quanto ao quarto aspecto (FsF-R1.3-01M), a ferramenta verificou se um padrão de metadados específico da comunidade é detectado usando *namespaces* e se está listado no registro *Re3data*.org. O *Re3data.org* visa contribuir para o estabelecimento de um ecossistema de repositórios de dados mais coerente e integrado. Isso o torna também uma ferramenta para acompanhar as tendências no

desenvolvimento de repositórios de dados de pesquisa (PAMPEL *et al.,* 2013), o que pode auxiliar seus gestores na escolha de padrões de metadados dentro da sua comunidade.

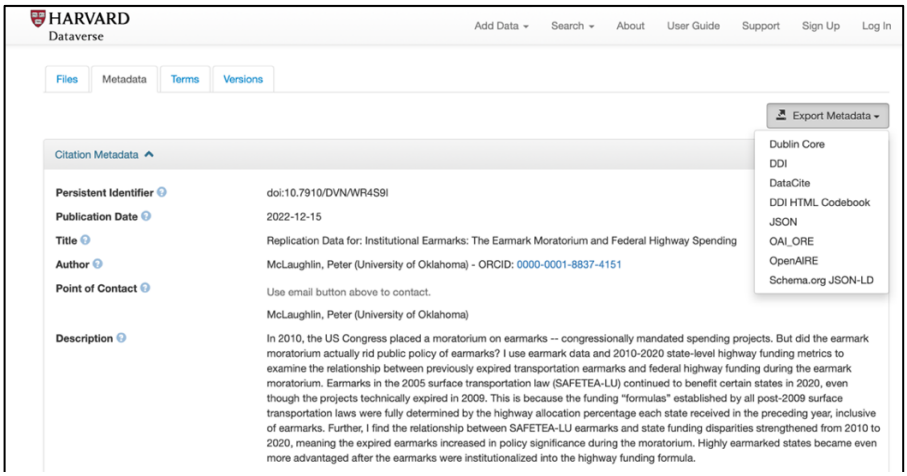
O nível inicial obtido pelos 15 *datasets* no quarto aspecto diz respeito ao uso de um padrão de metadados multidisciplinar, endossado pela comunidade (*RDA Metadata Standards Catalog*), que é detectado através de *namespaces.* A ferramenta F-UJI detectou o padrão *Dublin Core* e, segundo Pavão *et al*. (2015, p. 107):

> A DCMI é provavelmente a iniciativa mundial mais conhecida no que diz respeito a esquemas de metadados para descrição de objetos digitais, principalmente textuais, em repositórios digitais, e proporciona uma base para o desenvolvimento dos mesmos. O Dublin Core (DC) adota a sintaxe do *Resource Description Framework* (RDF) e surgiu como uma alternativa para simplificar e, consequentemente, tornar a descrição e recuperação de objetos digitais na *Web* mais rápida e econômica sem, contudo, deixar de seguir um padrão mínimo, indispensável para a interoperabilidade entre os sistemas.

É mais fácil reutilizar conjuntos de dados se eles forem semelhantes: mesmo tipo de dados, dados organizados de maneira padronizada, formatos de arquivo bem estabelecidos e sustentáveis, documentação (metadados) seguindo um modelo comum e usando vocabulário comum. Se existirem padrões da comunidade ou práticas recomendadas para arquivamento e compartilhamento de dados, eles devem ser seguidos. Ou seja, é um ponto positivo que os *datasets* da UFSCar sejam representados com o uso do DC, um padrão que se adequa a vários domínios (como é o caso em repositórios multidisciplinares) e auxilia na interoperabilidade entre repositórios digitais. Mas é importante salientar que o *DSpace* oferece o *Dublin Core* como esquema de metadados pré-definido, o que é o caso do repositório institucional da UFSCar.

O conjunto de dados anteriormente citado do *Harvard Dataverse* foi pontuado em um nível “moderado” nesse mesmo aspecto (FsF-R1.3-01M), e a ferramenta F-UJI foi capaz de recuperar três padrões: *Dublin Core*, *DataCite Metadata Schema e DDI - Data Documentation Initiative.* Os metadados do *dataset* podem ser exportados em DC, DDI, *DataCite*, JSON, *Schema.org* (Figura 11), dentre outros, representando uma boa prática que pode ser futuramente adotada pelos repositórios institucionais da amostra.

Figura 11 – Opções para exportar metadados de um *dataset* depositado no *Harvard Dataverse*



Fonte: *print screen* retirado do repositório *Harvard Dataverse* (2022).

Por fim, quanto ao último aspecto (FsF-R1.3-02D), a ferramenta verificou se o formato do arquivo de dados é aberto/de longo prazo/científico. Todos os conjuntos de dados foram avaliados como incompletos, ou seja, o formato do arquivo não era fornecido nos metadados/não estava contemplado na lista de formatos de arquivos de longo prazo, formatos de arquivos abertos ou formatos de arquivos científicos.

Por conter apenas 15 *datasets* depositados, foi feito o *download* manual deles para verificar os formatos dos arquivos, e a maioria se tratava de formatos proprietários (.xlsx e .docx), ou seja, elaborados em *softwares* pagos e, portanto, não estão disponíveis sem barreiras de acesso e custo como é defendido pelo movimento da Ciência Aberta. Berners-Lee (2009), inclusive, fez uso de um sistema que classifica em até cinco estrelas a disponibilização dos dados abertos ligados na *Web* (*Linked Open Data*). A terceira estrela diz respeito, justamente, ao uso de formatos não-proprietários, o que é fundamental para fornecer e democratizar o acesso aos dados de pesquisa, possibilitando sua reutilização.

Quadro 9 – Cinco estrelas para dados abertos conectados

Estrelas	Descrição
★	Dados disponíveis na web, em qualquer formato
★ ★	Dados estruturados disponíveis como legíveis por máquina (por exemplo: evitar imagens digitalizadas)
★ ★ ★	Dados estruturados disponíveis em um formato não-proprietário (exemplo: CSV em vez de Excel)
★ ★ ★ ★	Dados publicados usando padrões abertos da W3C (RDF e SPARQL)

★★★★★	Todos os itens acima e links para outros dados abertos conectados

Fonte: Adaptado de Berners-Lee (2006, tradução nossa).

A recomendação não foi aderida pelos *datasets* da amostra, que ficam também à mercê de ferramentas pagas para seu acesso. Essa prática está ligada às políticas de arquivamento da instituição. No manual de autodepósito de dados de pesquisa, disponível no repositório, a instituição afirma que todos os formatos de arquivos são permitidos, mas sugere a utilização da formatação em dados abertos com extensões do arquivo consolidadas como: XML, JSON, *Comma-separated values* (CSV). Supõe-se que os usuários só optariam pelos formatos abertos se houvesse uma política mandatória da instituição, o que pode ser estudado pela equipe para aumentar a reutilização dos *datasets*.

Assim como na interoperabilidade, todos os *datasets* obtiveram notas baixas na reutilização. A maior nota foi, como visto no Gráfico 1, quatro de dez (4 de 10). Essa lacuna abre espaço para a discussão de um ponto importante: para que dados de pesquisa sejam realmente interoperáveis e reutilizáveis, “[...] não devem apenas ser devidamente licenciados, mas os métodos para acessá-los e/ou baixá-los também devem ser bem descritos e, de preferência, totalmente automatizados e usando protocolos bem estabelecidos” (KALINAUSKAITĖ, 2017, p. 22).

É fundamental que os repositórios e os próprios pesquisadores se atentem quanto à reusabilidade dos dados de pesquisa porque, segundo Curty (2019, p. 191), “[...] o compartilhamento de dados de pesquisa de modo sistemático e comprometido com o potencial de reutilização futura deixa de ser opcional e passa a ser imperativo", exigindo a adoção da prática.

Sendo assim, os resultados obtidos por meio da ferramenta F-UJI demonstram uma falta de aderência às quatro facetas do FAIR, com foco em relação à interoperabilidade, em que todas as notas variaram entre 0 e 1 (sendo o máximo 4), e reutilização, em que todas as notas variaram entre 2 e 4 (sendo o máximo 10). Com relação ao acesso, todos os *datasets* da amostra foram pontuados com a mesma nota: 1,5 de 3. Quanto à encontrabilidade, apenas dois *datasets* obtiveram a nota mais alta: 4 de 7, enquanto a maioria recebeu nota 3.

A falta de adoção de um identificador persistente, como o DOI, que é amplamente utilizado para objetos digitais, é uma das lacunas que precisam ser

trabalhadas para a persistência dos dados (essencial para seu acesso e citação). O uso de metadados é outro ponto a ser avaliado e aperfeiçoado pela equipe, buscando aumentar o nível semântico para as máquinas. Mas, devido ao autoarquivamento, talvez seja necessário reformular o próprio manual disponível no repositório para instruir os usuários quanto à uma descrição mais rica dos *datasets,* com as informações sendo fornecidas nos elementos de metadados adequados. É importante frisar que, diferente dos resultados de pesquisa como artigos e dissertações, os dados de pesquisa exigem uma descrição/contextualização muito maior para serem interpretados. Metadados ricos são, portanto, imprescindíveis.

O recomendado é que a equipe gestora do repositório da UFSCar avalie os pontos destacados e verifique quais, de fato, são pertinentes para a instituição, uma vez que “Alguns dos princípios serão triviais a certos domínios de pesquisa e problemáticos a outros; portanto, cada campo de pesquisa precisa definir o que significa ser FAIR e decidir as medidas apropriadas para avaliar isso” (SILVA; RODRIGUES, 2021, p. 129).

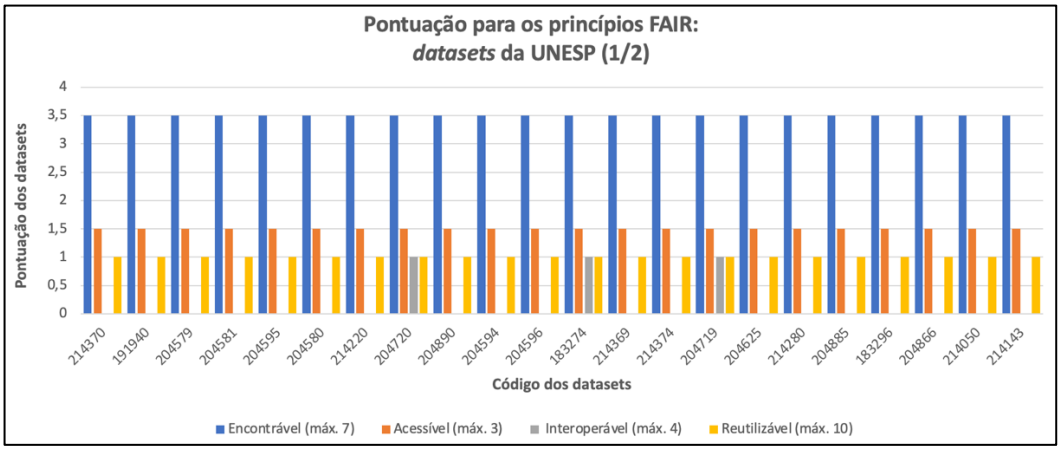
Seria vantajoso realizar essa avaliação enquanto a coleção ainda é pequena e pode ser mais facilmente aperfeiçoada, evitando assim problemas futuros principalmente de interoperabilidade e reuso dos dados depositados. O aperfeiçoamento quanto aos princípios FAIR é um passo inicial para que o repositório possa almejar certificações internacionais de qualidade, trazendo maior visibilidade para a instituição e seus pesquisadores.

5.1.2 Repositório institucional da Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)

O repositório da UNESP continha um total de 44 conjuntos de dados (*datasets*) depositados no momento da coleta. Como já visto, cada faceta do FAIR possui uma nota máxima, conforme o número indicado na legenda do gráfico: encontrável (máximo sete), acessível (máximo três), interoperável (máximo quatro) e reutilizável (máximo dez). É possível observar pelos Gráficos 2 e 3 que nenhum dos *datasets* do repositório da UNESP conseguiu atingir nota máxima. Foi o mesmo caso da UFSCar. A maior nota obtida em encontrável foi 3,5 de 7, em acessível 1,5 de 3, em interoperável 1 de 4 e em reutilizável foi 1 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.

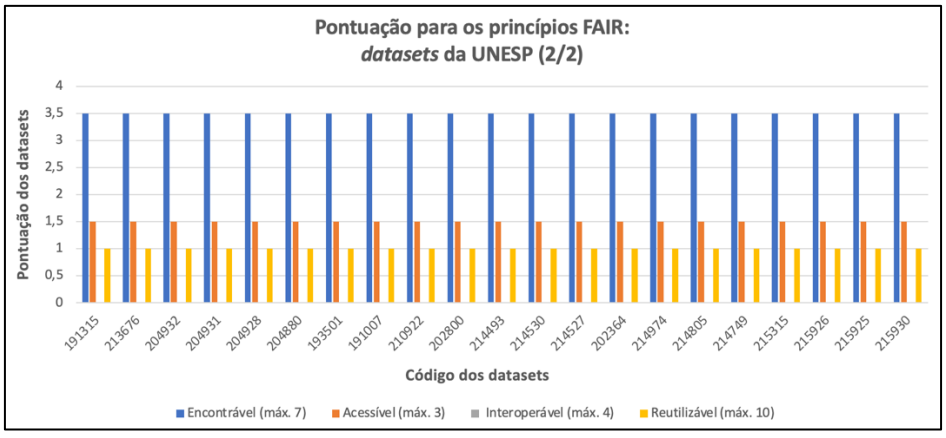
Gráfico 2 - Pontuação para os princípios FAIR: *datasets* da UNESP (1/2)



Fonte: elaborado pela autora (2022).

Como já visto, Dunning, Smaele e Böhmero (2017) encontraram resultado semelhante na Holanda, com as facetas “interoperável” e “reutilizável” sendo as mais difíceis de aderir. É possível também notar uma alta padronização quanto às notas obtidas pelos conjuntos de dados da UNESP, revelando uma consistência nos pontos fracos e fortes deles. Isso pode estar ligado ao autoarquivamento guiado pelo manual institucional, que indica os campos a serem preenchidos, o que também é identificado no caso da UFSCar.

Gráfico 3 - Pontuação para os princípios FAIR: *datasets* da UNESP (2/2)



Fonte: elaborado pela autora (2022).

Os 44 *datasets* da UNESP foram também avaliados em uma escala de 0 a 100% quanto à aderência geral aos princípios FAIR, sendo a maior pontuação igual a 29%. Apenas três alcançaram essa pontuação mais alta. Já a menor pontuação foi 25%, verificada para todos os demais conjuntos de dados depositados no repositório. Percebe-se que o percentual alcançado quanto ao nível geral de *FAIRness* é baixo, indo novamente de acordo com o que é exposto por Henning *et al.* (2019a) quanto à baixa conformidade esperada. Os conjuntos de dados que obtiveram as notas mais altas (29%) foram aqueles que pontuaram na interoperabilidade, tirando um em vez de zero.

Entrando na análise mais minuciosa, que apresenta o nível (inicial, moderado, avançado ou incompleto) de conformidade quanto a cada aspecto testado dentro do FAIR, já é possível ver algumas diferenças e semelhanças quanto ao resultado da UFSCar. No Quadro 10 é possível verificar o nível de aderência para cada um dos 44 conjuntos de dados da UNESP.

Quadro 10 - Aderência quanto à Encontrabilidade (UNESP)

Código	FsF-F1-01D	FsF-F1-02D	FsF-F2-01M	FsF-F3-01M	FsF-F4-01M
214097	avançado	avançado	inicial	incompleto	avançado
214370	avançado	avançado	inicial	incompleto	avançado
191940	avançado	avançado	inicial	incompleto	avançado
204579	avançado	avançado	inicial	incompleto	avançado
204581	avançado	avançado	inicial	incompleto	avançado
204595	avançado	avançado	inicial	incompleto	avançado
204580	avançado	avançado	inicial	incompleto	avançado
214220	avançado	avançado	inicial	incompleto	avançado
204720	avançado	avançado	inicial	incompleto	avançado
204890	avançado	avançado	inicial	incompleto	avançado
204594	avançado	avançado	inicial	incompleto	avançado
204596	avançado	avançado	inicial	incompleto	avançado
183274	avançado	avançado	inicial	incompleto	avançado
214369	avançado	avançado	inicial	incompleto	avançado
214374	avançado	avançado	inicial	incompleto	avançado
204719	avançado	avançado	inicial	incompleto	avançado
204625	avançado	avançado	inicial	incompleto	avançado
214280	avançado	avançado	inicial	incompleto	avançado
204885	avançado	avançado	inicial	incompleto	avançado

183296	avançado	avançado	inicial	incompleto	avançado
204866	avançado	avançado	inicial	incompleto	avançado
214050	avançado	avançado	inicial	incompleto	avançado
214143	avançado	avançado	inicial	incompleto	avançado
191315	avançado	avançado	inicial	incompleto	avançado
213676	avançado	avançado	inicial	incompleto	avançado
204932	avançado	avançado	inicial	incompleto	avançado
204931	avançado	avançado	inicial	incompleto	avançado
204928	avançado	avançado	inicial	incompleto	avançado
204880	avançado	avançado	inicial	incompleto	avançado
193501	avançado	avançado	inicial	incompleto	avançado
191007	avançado	avançado	inicial	incompleto	avançado
210922	avançado	avançado	inicial	incompleto	avançado
202800	avançado	avançado	inicial	incompleto	avançado
214493	avançado	avançado	inicial	incompleto	avançado
214530	avançado	avançado	inicial	incompleto	avançado
214527	avançado	avançado	inicial	incompleto	avançado
202364	avançado	avançado	inicial	incompleto	avançado
214974	avançado	avançado	inicial	incompleto	avançado
214805	avançado	avançado	inicial	incompleto	avançado
214749	avançado	avançado	inicial	incompleto	avançado
215315	avançado	avançado	inicial	incompleto	avançado
215926	avançado	avançado	inicial	incompleto	avançado
215925	avançado	avançado	inicial	incompleto	avançado
215930	avançado	avançado	inicial	incompleto	avançado

Fonte: elaborado pela autora (2022).

O primeiro aspecto (FsF-F1-01D), que diz respeito a: "são atribuídos identificadores globalmente exclusivos aos dados", foi avaliado como "avançado" para todos os conjuntos de dados da UNESP, assim como no caso da UFSCar. A ferramenta auxiliar verificou inicialmente se o identificador atribuído segue uma sintaxe de identificador único definido (IRI, URL) e, de fato, todos os *datasets* cumprem o requisito.

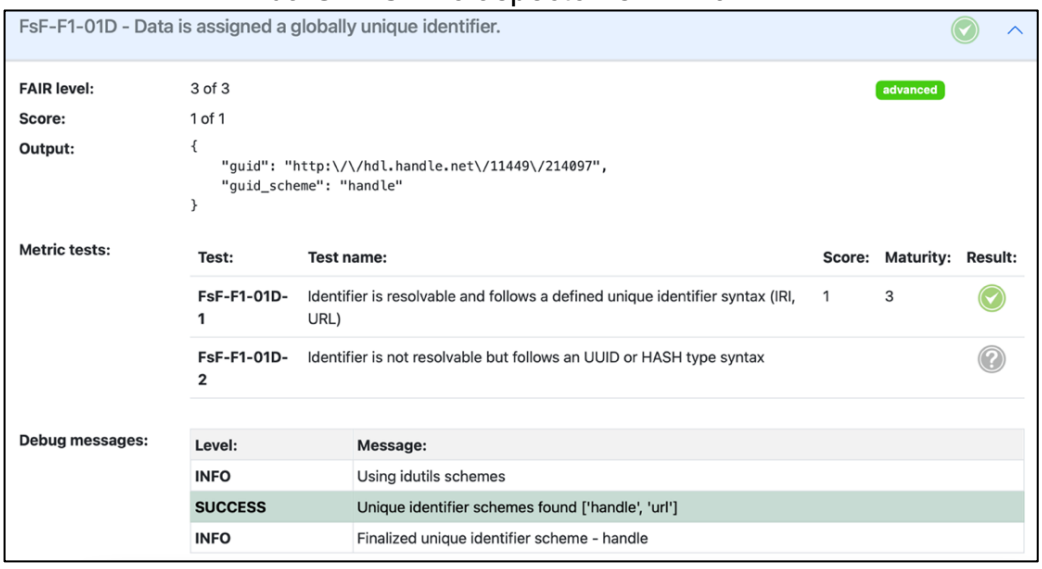
Somado a isso, o repositório da UNESP ainda identifica os autores do *dataset* com seus perfis no Google Acadêmico e *Open Researcher and Contributor ID* (ORCID). Essa é uma funcionalidade não presente no repositório da UFSCar, mas que agrega valor para os dados. De acordo com Marín-Arraiza e Heredia (2021, p. 33)

o ORCID contribui para a aderência dos dados de pesquisa aos princípios FAIR, uma vez que “[...] atua como padrão internacional na identificação persistente de autores”.

Todos os conjuntos de dados da UNESP também foram avaliados como “avançado” no aspecto FsF-F1-02D, que diz respeito a: “são atribuídos identificadores persistentes aos dados". Como já visto, os identificadores persistentes são de suma importância tanto para a descoberta quanto para a reutilização dos dados de pesquisa, uma vez que possuem uma série de metadados associados que são legíveis por máquinas, identificando o objeto e não a localização dele. Diferente dos *datasets* da UFSCar, os da UNESP foram validados pela ferramenta auxiliar, que identificou o uso do esquema *Handle*, conforme mostra a Figura 12. Isso é um excelente indicativo para a encontrabilidade e citação dos dados da instituição.

Segundo Sayão (2007, p.71), o *Handle System* “[...] é um sistema distribuído de computadores concebido para assinalar, armazenar, administrar e resolver identificadores ou nomes persistentes de objetos digitais conhecidos como *handle”.* Ele é adotado pelo *DSpace* e sua sintaxe é composta de duas partes: o “prefixo”, que “[...] identifica a autoridade nomeadora, a unidade administrativa autorizada a criar e manter nomes” e o “sufixo”, que se refere ao “[...] nome do recurso, único dentro do domínio da autoridade nomeadora” (SAYÃO, 2007, p. 72). Como o nome do recurso não possui uma sintaxe pré-estabelecida, é importante que a autoridade local esteja em conformidade com as regras específicas.

Figura 12 – *Output* exibido pela ferramenta F-UJI ao avaliar um conjunto de dados da UNESP no aspecto FsF-F1-01D



Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

Quanto ao terceiro aspecto de “encontrável”, todos os conjuntos de dados foram avaliados pela ferramenta F-UJI como “inicial”. O objetivo era verificar se os metadados incluíam os elementos essenciais (criador, título, identificador, publicador, data de publicação, resumo e palavras-chave) para dar o suporte à encontrabilidade. Foi expresso que nem todos os elementos necessários para a citação dos conjuntos de dados estavam presentes, como o “*publisher”,* o que pode ser aprimorado pela instituição.

De acordo com Hodson *et al*. (2018), os dados precisam ser acompanhados por metadados básicos de descoberta para permitir que sejam encontrados, usados e citados de forma confiável. No modelo para *FAIR Data Objects* que os autores elaboram, os metadados possuem um papel essencial na contextualização. Os metadados básicos permitem a descoberta de dados, mas são necessárias informações e proveniência muito mais ricas para entender como, por que, quando e por quem os dados foram criados. Para permitir a reutilização mais ampla no futuro, os dados devem ser acompanhados por uma 'pluralidade de atributos relevantes'. Ou seja, é importante que as instituições busquem sempre incentivar descrições ricas de seus conjuntos de dados, tornando-os mais (F)AI(R).

Voltando para a análise dos *datasets*, assim como no caso do repositório da UFSCar, o da UNESP também oferece, no registro, a citação pronta do conjunto de dados, o que pode incentivar o reuso dos dados de pesquisa depositados. Mas a ferramenta não encontrou metadados acessíveis por meio de "*typed links*" ou "*signposting links*", assim como no caso da UFSCar. Esses termos, como já visto, dizem respeito às sinalizações e ligações que tornam a *web* acadêmica mais amigável para as máquinas. Ao visitar portais acadêmicos, os leitores podem facilmente descobrir *links* para registros bibliográficos, autoria, etc.

O registro de metadados possuía outros elementos adicionais de descrição como financiamento, formato, língua e o *lattes* dos pesquisadores, mas o nível inicial atribuído pela ferramenta F-UJI é um indicativo de que a UNESP pode procurar investir em descrições mais ricas de seus dados de pesquisa, com uma pluralidade de atributos relevantes, como defendem os princípios FAIR.

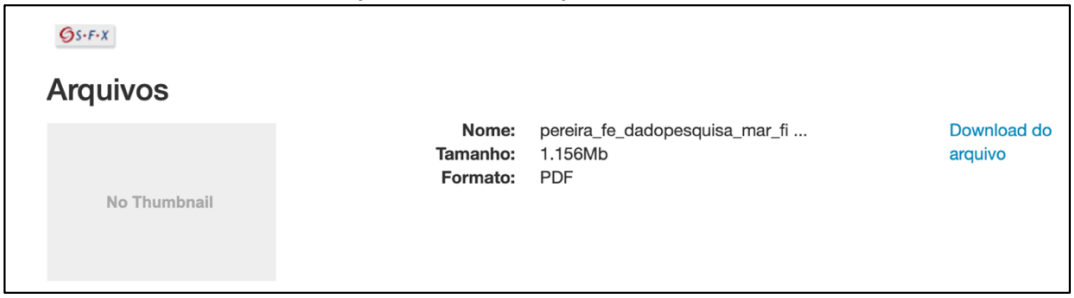
Nesse sentido, pode-se citar uma iniciativa chamada *FAIR Data Point* (FDP), “[...] um dos componentes do *Data FAIRPort*, é a ferramenta voltada para a publicação de metadados de forma FAIR, desenvolvida com tecnologias da *web semântica*,

especificamente *Linked Data* com a linguagem *Resource Description Framework* (RDF)" (HENNING *et al*., 2019a, p. 188). O FDP está sob responsabilidade da iniciativa GO-FAIR e usa o modelo *Data Catalog Vocabulaire* (DCAT) da W3C como base para seu conteúdo de metadados. A equipe gestora pode, então, verificar essa iniciativa para melhor contextualização das descrições de seus conjuntos de dados. Ainda de acordo com Felipe e Santos (2022):

> FAIR *Data Point* consiste em um *software* que trabalha os metadados, em especial para os repositórios de dados de pesquisa, com o emprego de dados relativos à proveniência e à iniciativa FAIR. Objetiva tornar os conjuntos de dados FAIR acessíveis através do uso de metadados legíveis para máquinas e humanos.

O próximo aspecto (FsF-F3-01M) diz respeito a: "os metadados incluem o identificador do dado que descrevem". A ferramenta avaliou duas características: 1) os metadados contêm informações relacionadas ao conteúdo dos dados (nome do arquivo, tamanho, tipo) e 2) os metadados contêm um PID ou URL que indica a localização do conteúdo dos dados para *download*. Mas, assim como no caso dos conjuntos de dados da UFSCar, o identificador dos dados foi tido como ausente e, por isso, todos os *datasets* foram avaliados como "incompleto". O tamanho, nome e tipo do arquivo, por exemplo, não foram recuperados, o que pode estar associado à falta da indicação adequada em metadados legíveis por máquina, uma vez que é possível encontrar essas informações no registro, conforme mostra a Figura 13.

Figura 13 – Visualização do nome, tamanho e formato do arquivo de um conjunto de dados depositado no repositório da UNESP



Fonte: *print screen* retirado do repositório da UNESP (2022).

Por fim, no aspecto FsF-F4-01M, os *datasets* foram testados quanto aos metadados fornecidos de uma maneira que os principais mecanismos de pesquisa pudessem inseri-los em seus catálogos (exemplos: JSON-LD, *Dublin Core*, RDFa), e

todos foram avaliados como “avançado”, mesmo resultado encontrado nos conjuntos de dados da UFSCar. Isso é um bom sinal para sua encontrabilidade e para a visibilidade da instituição, indicando um ponto forte em comum.

Além disso, é interessante citar que o repositório da UNESP, ao contrário do da UFSCar, está indexado no diretório Re3data, aumentando a visibilidade do sistema e da instituição. Essa é uma prática importante quando se fala na encontrabilidade dos dados de pesquisa. Sendo assim, a equipe gestora pode se concentrar em melhorias nos seguintes aspectos de encontrabilidade: metadados descritivos e que incluem o identificador dos dados que descrevem.

Já quanto à **acessibilidade**, os conjuntos de dados foram avaliados em três aspectos: 1) Os metadados contêm o nível de acesso e as condições de acesso aos dados (FsF-A1-01M); 2) Os dados são acessíveis por meio de um protocolo de comunicação padronizado (FsF-A1-03D) e 3) Os metadados são acessíveis por meio de um protocolo de comunicação padronizado (FsF-A1-02M). Vidotti, Torino e Coneglian (2021, p. 212) explicam que esta faceta reflete a capacidade do conjunto de dados de “[...] ser acessado e as especificações para fazê-lo, incluindo a utilização de protocolos de comunicação, autenticação, níveis de acesso e a persistência dos metadados, ainda que os dados não estejam mais disponíveis”.

Quadro 11 - Aderência quanto à Acessibilidade (UNESP)

Código	FsF-A1-01M	FsF-A1-03D	FsF-A1-02M
214097	inicial	incompleto	avançado
214370	inicial	incompleto	avançado
191940	inicial	incompleto	avançado
204579	inicial	incompleto	avançado
204581	inicial	incompleto	avançado
204595	inicial	incompleto	avançado
204580	inicial	incompleto	avançado
214220	inicial	incompleto	avançado
204720	inicial	incompleto	avançado
204890	inicial	incompleto	avançado
204594	inicial	incompleto	avançado
204596	inicial	incompleto	avançado
183274	inicial	incompleto	avançado
214369	inicial	incompleto	avançado
214374	inicial	incompleto	avançado

204719	inicial	incompleto	avançado
204625	inicial	incompleto	avançado
214280	inicial	incompleto	avançado
204885	inicial	incompleto	avançado
183296	inicial	incompleto	avançado
204866	inicial	incompleto	avançado
214050	inicial	incompleto	avançado
214143	inicial	incompleto	avançado
191315	inicial	incompleto	avançado
213676	inicial	incompleto	avançado
204932	inicial	incompleto	avançado
204931	inicial	incompleto	avançado
204928	inicial	incompleto	avançado
204880	inicial	incompleto	avançado
193501	inicial	incompleto	avançado
191007	inicial	incompleto	avançado
210922	inicial	incompleto	avançado
202800	inicial	incompleto	avançado
214493	inicial	incompleto	avançado
214530	inicial	incompleto	avançado
214527	inicial	incompleto	avançado
202364	inicial	incompleto	avançado
214974	inicial	incompleto	avançado
214805	inicial	incompleto	avançado
214749	inicial	incompleto	avançado
215315	inicial	incompleto	avançado
215926	inicial	incompleto	avançado
215925	inicial	incompleto	avançado
215930	inicial	incompleto	avançado

Fonte: elaborado pela autora (2022).

Nota-se, pelo Quadro 11, que todos os conjuntos de dados aderiram ao primeiro e terceiro aspecto, apesar de em níveis diferentes, exatamente igual ao resultado encontrado nos *datasets* da UFSCar. Existe, portanto, essa equivalência de pontos fortes e pontos a serem trabalhados quanto ao acesso aos dados de pesquisa de ambas as instituições.

Para o aspecto FsF-A1-01M, a ferramenta auxiliar verificou se os *datasets* possuíam informações sobre restrições ou direitos de acesso nos metadados,

recuperando a informação “Acesso Aberto” associada aos registros, diferente da UFSCar, que utilizou as licenças CC. Mas a ferramenta F-UJI alegou que a informação de acesso aos dados não é legível por máquina e, portanto, um computador seria incapaz de determinar o nível de acesso em um processamento automático, como é defendido pelos princípios FAIR. Foi o mesmo caso da UFSCar.

Fazendo uma comparação, ao analisar um conjunto de dados (https://doi.org/10.5281/zenodo.7454759) depositado no *Zenodo*, repositório aberto de uso geral desenvolvido sob o programa europeu *OpenAIRE*, a ferramenta conseguiu detectar (em nível avançado) as condições e níveis de acesso do *dataset* (Figura 14), retornado a informação “público”. Olhando para os metadados é possível observar o uso do *dc:rights* com o valor “info:eu-repo/semantics/openAccess”. O *namespace* “info:eu-repo” está registrado em http://info-uri.info e é um *placeholder* autorizado para termos semânticos, vocabulários controlados e identificadores. Ao usar este *namespace* todos os termos usados têm uma "presença na *web*", não sendo uma *string* arbitrária e contendo significado.

Figura 14 – Métricas e *debug messages* exibidas pela ferramenta F-UJI ao avaliar um conjunto de dados do Zenodo no aspecto FsF-A1-01M

Metric tests:

Test:	Test name:	Score:	Maturity:	Result:
FsF-A1-01M-1	Information about access restrictions or rights can be identified in metadata	0.5	1	✓
FsF-A1-01M-2	Data access information is machine readable	0.5	3	✓
FsF-A1-01M-3	Data access information is indicated by (not machine readable) standard terms			?

Debug messages:

Level:	Message:
INFO	Access information specified -: info:eu-repo/semantics/openAccess
INFO	Verify name through SPDX registry -: info:eu-repo/semantics/openAccess
INFO	Standardized actionable access level recognized as -:public
INFO	Found access rights information in dedicated metadata element
SUCCESS	Access level to data could successfully be determined -: public

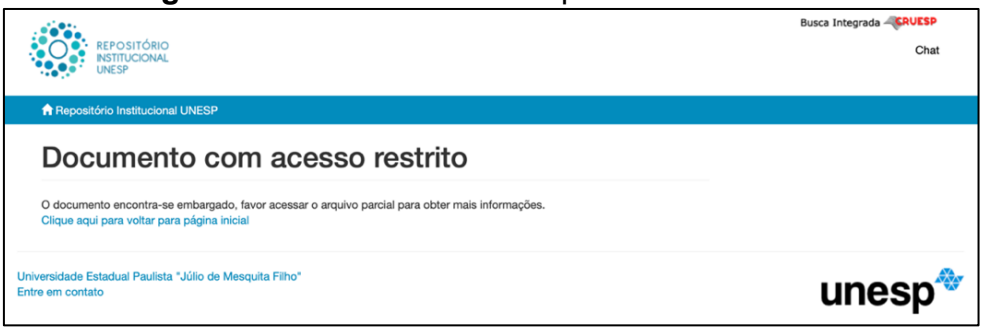
Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

Esse conjunto de dados obteve uma aderência geral de 79% aos princípios FAIR quando avaliado pela ferramenta F-UJI, e pode ser usado como base para estudos e melhorias pelos repositórios da amostra quanto à declaração de acesso em formato legível por máquina.

Como já visto anteriormente, o “A” não significa necessariamente ‘aberto’ ou ‘livre’. Na verdade, implica que se deve fornecer as condições exatas de acesso aos

dados e, portanto, dados privados também podem ser FAIR. E, assim como no caso da UFSCar, existem dados restritos no repositório da UNESP (Figura 15). Seu acesso só é permitido após um período de embargo definido pelo autor.

Figura 15 – Item restrito no repositório da UNESP



Fonte: *print screen* retirado do repositório da UNESP (2022).

Mas, ao contrário do que acontece no repositório da UFSCar, os metadados continuam acessíveis, mesmo não sendo possível acessar o arquivo, como é preconizado nos princípios FAIR. Isso permite que os usuários do repositório avaliem, por meio dos metadados, se aquele *dataset* é, de fato, útil para suas pesquisas, para então entrar em contato com o responsável pelos dados. Além disso, permite que as máquinas reúnam informações sobre o *dataset,* fornecendo análises e relatórios como as linhas de pesquisa em alta na instituição, mesmo sem o acesso ao conjunto de dados em si.

Já em relação ao segundo aspecto (FsF-A1-03D), que diz respeito aos dados serem acessíveis por meio de um protocolo de comunicação padronizado, a ferramenta examinou se os metadados incluíam um *link* para dados com base em protocolos de comunicação da *web* padronizados (*standard_data_protocol*), o que não foi encontrado. Entretanto, todos os *datasets* são identificados por *links* que começam com https://, ou seja, são recuperados usando um protocolo de comunicação padrão, o *Hyper Text Transfer Protocol Secure* (HTTPS). Essa questão também foi, inclusive, testada para os metadados (FsF-A1-02M), e a ferramenta F-UJI conseguiu localizar o protocolo padrão para acesso aos metadado: HTTPS. Logo, é possível verificar uma tendência entre os dois primeiros repositórios avaliados quanto à acessibilidade, o que aponta para melhorias em comum.

Quadro 12 - Aderência quanto à Interoperabilidade (UNESP)

Código	FsF-I1-01M	FsF-I1-02M	FsF-I3-01M
214097	incompleto	inicial	incompleto
214370	incompleto	inicial	incompleto
191940	incompleto	inicial	incompleto
204579	incompleto	inicial	incompleto
204581	incompleto	inicial	incompleto
204595	incompleto	inicial	incompleto
204580	incompleto	inicial	incompleto
214220	incompleto	inicial	incompleto
204720	incompleto	inicial	avançado
204890	incompleto	inicial	incompleto
204594	incompleto	inicial	incompleto
204596	incompleto	inicial	incompleto
183274	incompleto	inicial	avançado
214369	incompleto	inicial	incompleto
214374	incompleto	inicial	incompleto
204719	incompleto	inicial	avançado
204625	incompleto	inicial	incompleto
214280	incompleto	inicial	incompleto
204885	incompleto	inicial	incompleto
183296	incompleto	inicial	incompleto
204866	incompleto	inicial	incompleto
214050	incompleto	inicial	incompleto
214143	incompleto	inicial	incompleto
191315	incompleto	inicial	incompleto
213676	incompleto	inicial	incompleto
204932	incompleto	inicial	incompleto
204931	incompleto	inicial	incompleto
204928	incompleto	inicial	incompleto
204880	incompleto	inicial	incompleto
193501	incompleto	inicial	incompleto
191007	incompleto	inicial	incompleto
210922	incompleto	inicial	incompleto
202800	incompleto	inicial	incompleto
214493	incompleto	inicial	incompleto
214530	incompleto	inicial	incompleto
214527	incompleto	inicial	incompleto
202364	incompleto	inicial	incompleto

214974	incompleto	inicial	incompleto
214805	incompleto	inicial	incompleto
214749	incompleto	inicial	incompleto
215315	incompleto	inicial	incompleto
215926	incompleto	inicial	incompleto
215925	incompleto	inicial	incompleto
215930	incompleto	inicial	incompleto

Fonte: elaborado pela autora (2022).

A ferramenta F-UJI também avaliou os *datasets* quanto à **interoperabilidade** e, segundo Henning *et al.* (2019a), é preciso considerar os problemas relacionados às questões técnicas do FAIR essenciais para a interoperabilidade, o princípio que pode ser visto como o mais desafiador. A interoperabilidade é uma característica essencial no valor e usabilidade dos dados, e não é apenas semântica, mas também técnica. A interoperabilidade técnica significa que os dados e as informações relacionadas são codificados usando um padrão que pode ser lido em todos os sistemas aplicáveis, o que pode ser um desafio para as instituições (HODSON *et al*., 2018). E, de fato, em comparação com o resultado da UFSCar, a UNESP obteve pontuações mais baixas em interoperabilidade.

Como já visto, foram avaliados três aspectos em interoperável: 1) Os metadados são representados usando uma linguagem de representação de conhecimento formal (FsF-I1-01M); 2) Metadados usam recursos semânticos (FsF-I1-02M) e 3) Os metadados incluem *links* entre os dados e suas entidades relacionadas (FsF-I3-01M).

Nenhum dos conjuntos de dados foi validado quanto ao primeiro aspecto, que diz respeito a uma linguagem de representação de conhecimento formal. Os *datasets* foram testados quanto à existência de metadados estruturados e analisáveis (JSON-LD, RDFa) incorporados ao código XHTML/HTML da página de destino. Segundo Löffler *et al*. (2021), a ideia que vem sendo difundida é adicionar informações descritivas aos dados estruturados, como *Extensible Markup Language* (XML) ou *HyperText Markup Language* (HTML), a fim de aumentar a encontrabilidade e interoperabilidade dos dados de pesquisa.

Verificando novamente o conjunto de dados depositado no *Zenodo* (https://doi.org/10.5281/zenodo.7454759) e o conjunto de dados depositado no *Harvard Dataverse* (https://doi.org/10.7910/DVN/WR4S9I), ambos avaliados como

avançado no aspecto FsF-I1-01M, é possível identificar o botão de “exportar metadados”. Dentre as diversas opções disponíveis, tem-se justamente o JSON-LD, aumentando a interoperabilidade dos dados de pesquisa.

Como já visto, ele é um formato recomendado pela W3C, e é baseado em JSON para serializar *Linked Data*. Destina-se principalmente a ser uma maneira de usar *Linked Data* em ambientes de programação baseados na *we*b, para construir serviços interoperáveis e armazenar *Linked Data* em mecanismos de armazenamento baseados em JSON. Logo, tanto a UFSCar quanto a UNESP podem estudar essa implementação para tornar seus conjuntos de dados mais interoperáveis e, consequentemente, mais FAIR.

Já no aspecto FsF-I1-02M a ferramenta buscou por uso de *namespaces* de vocabulário e de recursos semânticos nos metadados, e todos os *datasets* pontuaram como inicial. Como já visto no tópico anterior, o teste buscou extrair *namespaces* dos metadados baseados em RDF*,* que garantem que determinado conjunto de objetos tenha nomes exclusivos para que possa ser facilmente identificado. Henning *et al*. (2019a, p. 183) explicam que “Os padrões de metadados, os vocabulários controlados, as ontologias e os tesauros servem para enriquecer a descrição de dados, desse modo, permitir a transferência e a troca de protocolos e a maior precisão na recuperação dos dados”. Logo, é fundamental que todos os repositórios busquem investir em recursos semânticos como ontologias e vocabulários, melhorando a interoperabilidade de seus sistemas. Sem contar que essa prática facilita o processamento automático por máquinas.

Quanto ao último aspecto da interoperabilidade, que diz respeito às ligações entre os dados e suas entidades relacionadas, apenas três conjuntos de dados conseguiram pontuar como “avançado”, apresentando assim recursos relacionados nos metadados indicados por *links* legíveis por máquina. Esses três conjuntos de dados foram, inclusive, os únicos que alcançaram a maior pontuação geral de aderência: 29%, enquanto os outros obtiveram uma pontuação igual a 25% de *FAIRness*.

Em comum, todos eles apresentavam o elemento “*dc.relation*”, com um *link* associado (*isRelatedTo*). Dois *links* levavam a uma dissertação de mestrado e um levava até um artigo, apontando para os resultados de pesquisa que os dados embasaram. Essa prática é fundamental quando se pensa na Web Semântica, onde os dados são conectados e permitem essas navegações. Todos os outros conjuntos

de dados da UNESP foram avaliados como “incompleto” no aspecto FsF-I3-01M. É recomendado, portanto, que a equipe gestora do repositório busque incentivar o uso do elemento “*dc.relation*”, associando o conjunto de dados depositado com o seu resultado, que muitas vezes já é um trabalho publicado no mesmo repositório institucional ou em revistas de acesso aberto. Isso tornaria os dados mais interoperáveis, sem contar que traria maior contextualização para os dados.

Logo, o repositório da UNESP, como visto, tem aspectos que podem ser trabalhados para uma melhor interoperabilidade do sistema. Criar ligações entre os *datasets* e suas entidades relacionadas e investir em recursos semânticos e em metadados estruturados e analisáveis (JSON-LD, RDFa) incorporados no código XHTML/HTML da página de destino são exemplos que podem ser observados pela instituição, uma vez que a interoperabilidade está ligada à ideia central do FAIR: enfatizar o aprimoramento da capacidade das máquinas de encontrar e usar os dados automaticamente.

Por fim, todos os *datasets* foram avaliados quanto à **reutilização**. De acordo com Henning *et al*. (2019a, p. 177) para alcançar graus maiores de reutilização dos dados é preciso a “[...] adoção de padrões, metadados, vocabulários controlados, ontologias e identificadores persistentes que proporcionam significado preciso aos dados e aos demais objetos a eles vinculados”. Para uma boa aderência, os *datasets* deveriam ser descritos em um nível de granularidade significativo para que detalhes e atributos fossem representados, e a descrição deveria ser precisa, clara e relevante, tanto para a comunidade quanto para o domínio do repositório de dados, com o uso de licenças explícitas e claras.

Os dados de pesquisa foram testados em cinco aspectos pela ferramenta F-UJI: 1) Metadados especificam o conteúdo dos dados (FsF-R1-01MD); 2) Os metadados incluem informações de licença sob as quais os dados podem ser reutilizados (FsF-R1.1-01M); 3) Os metadados incluem informações de proveniência sobre a criação ou geração dos dados (FsF-R1.2-01M); 4) Os metadados seguem um padrão recomendado pela comunidade de pesquisa-alvo dos dados (FsF-R1.3-01M) e 5) Os dados estão disponíveis em um formato de arquivo recomendado pela comunidade de pesquisa-alvo (FsF-R1.3-02D). O Quadro 13 mostra a consistência entre os resultados obtidos por todos os 44 *datasets* avaliados, que receberam as mesmas pontuações em todos os aspectos de reutilização.

Quadro 13 - Aderência quanto à Reutilização (UNESP)

Código	FsF-R1-01MD	FsF-R1.1-01M	FsF-R1.2-01M	FsF-R1.3-01M	FsF-R1.3-02D
214097	incompleto	incompleto	moderado	inicial	incompleto
214370	incompleto	incompleto	moderado	inicial	incompleto
191940	incompleto	incompleto	moderado	inicial	incompleto
204579	incompleto	incompleto	moderado	inicial	incompleto
204581	incompleto	incompleto	moderado	inicial	incompleto
204595	incompleto	incompleto	moderado	inicial	incompleto
204580	incompleto	incompleto	moderado	inicial	incompleto
214220	incompleto	incompleto	moderado	inicial	incompleto
204720	incompleto	incompleto	moderado	inicial	incompleto
204890	incompleto	incompleto	moderado	inicial	incompleto
204594	incompleto	incompleto	moderado	inicial	incompleto
204596	incompleto	incompleto	moderado	inicial	incompleto
183274	incompleto	incompleto	moderado	inicial	incompleto
214369	incompleto	incompleto	moderado	inicial	incompleto
214374	incompleto	incompleto	moderado	inicial	incompleto
204719	incompleto	incompleto	moderado	inicial	incompleto
204625	incompleto	incompleto	moderado	inicial	incompleto
214280	incompleto	incompleto	moderado	inicial	incompleto
204885	incompleto	incompleto	moderado	inicial	incompleto
183296	incompleto	incompleto	moderado	inicial	incompleto
204866	incompleto	incompleto	moderado	inicial	incompleto
214050	incompleto	incompleto	moderado	inicial	incompleto
214143	incompleto	incompleto	moderado	inicial	incompleto
191315	incompleto	incompleto	moderado	inicial	incompleto
213676	incompleto	incompleto	moderado	inicial	incompleto
204932	incompleto	incompleto	moderado	inicial	incompleto
204931	incompleto	incompleto	moderado	inicial	incompleto
204928	incompleto	incompleto	moderado	inicial	incompleto
204880	incompleto	incompleto	moderado	inicial	incompleto
193501	incompleto	incompleto	moderado	inicial	incompleto
191007	incompleto	incompleto	moderado	inicial	incompleto
210922	incompleto	incompleto	moderado	inicial	incompleto
202800	incompleto	incompleto	moderado	inicial	incompleto
214493	incompleto	incompleto	moderado	inicial	incompleto
214530	incompleto	incompleto	moderado	inicial	incompleto
214527	incompleto	incompleto	moderado	inicial	incompleto
202364	incompleto	incompleto	moderado	inicial	incompleto

214974	incompleto	incompleto	moderado	inicial	incompleto
214805	incompleto	incompleto	moderado	inicial	incompleto
214749	incompleto	incompleto	moderado	inicial	incompleto
215315	incompleto	incompleto	moderado	inicial	incompleto
215926	incompleto	incompleto	moderado	inicial	incompleto
215925	incompleto	incompleto	moderado	inicial	incompleto
215930	incompleto	incompleto	moderado	inicial	incompleto

Fonte: elaborado pela autora (2022).

O primeiro aspecto (FsF-R1-01MD) avaliou pontos como a existência de informações mínimas fornecidas nos metadados sobre o conteúdo dos dados disponíveis, e todos os *datasets* pontuaram como “incompleto”. Ou seja, a ferramenta não conseguiu localizar, nos metadados, informações como tipo e tamanho dos arquivos. Como já visto em “encontrável”, isso pode estar associado à falta da indicação adequada em metadados legíveis por máquina, uma vez que é possível encontrar facilmente essas informações no registro.

Assim como no caso da UFSCar, o “tipo” de recurso é “dado de pesquisa”, mas a ferramenta não conseguiu extrair essa informação. Considerando que o repositório da UNESP também engloba outros documentos produzidos no âmbito da instituição, como livros, teses e dissertações, essa informação pode ajudar a contextualizar o recurso e criar a organização do sistema. A ferramenta também não conseguiu extrair o tamanho do arquivo, e assim como no caso da UFSCar, essa informação aparece na opção de “*download* do arquivo”, que não é um metadado apropriado. Essa informação volta a aparecer na lista de arquivos, como visto na Figura 13, mas a ferramenta também não conseguiu localizar.

De forma adicional, há o uso do elemento “*dc.format*”, que oferece mais informações sobre o conjunto de dados. A UFSCar também oferece o formato do arquivo, mas não em um elemento adequado no registro de metadados. Alguns exemplos de valores encontrados para esse elemento são: ‘*Portable Document Format* (PDF)’, ‘mp4’, ‘Excel’, ‘digital’ e ‘arquivos diversos’, de uma forma ampla. A instituição pode avaliar uma maior consistência de preenchimento desse campo, padronizando as opções de valores.

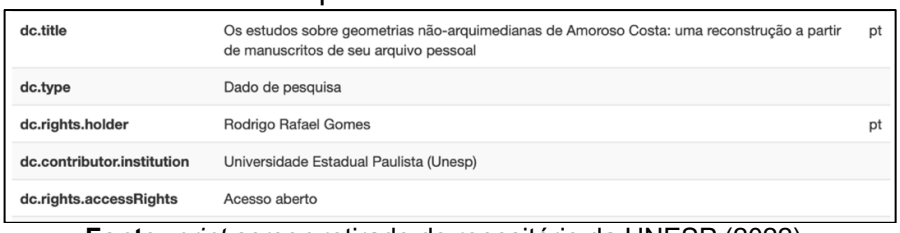
Fazendo uma comparação, ao analisar um conjunto de dados do *Harvard Dataverse* (https://doi.org/10.7910/DVN/WR4S9I) que pontuou como avançado nesse aspecto, é possível encontrar nos metadados exportados do *dataset* campos como

“*name*”, “*type*”, “*format*” e “*size*” para os arquivos do conjunto de dados. O nome, tipo e tamanho do arquivo são disponibilizados em um registro de metadados para cada arquivo que compõe o *dataset* (que também possui um registro de metadados). Nesse caso o *dataset* possui três arquivos publicados, e cada um deles tem seus próprios metadados, especificando tipos e tamanhos diferentes. Assim, a ferramenta foi capaz de localizar essas informações. A equipe gestora pode avaliar a viabilidade de implementar uma estrutura similar de descrição, mas é preciso verificar se o *software* escolhido permite essa funcionalidade.

O segundo aspecto (FsF-R1.1-01M) diz respeito aos metadados incluírem ou não informações de licença sob as quais os dados podem ser reutilizados. Hodson *et al.* (2018) citam em seu trabalho o conceito de interoperabilidade legal, que acontece quando: 1) as condições legais de uso são claras e prontamente determináveis para cada um dos conjuntos de dados, geralmente por meios automatizados; 2) as condições legais de uso impostas a cada conjunto de dados permitem a criação e uso de produtos combinados ou derivados e 3) os usuários podem acessar e usar legalmente cada conjunto de dados sem obter autorização dos detentores de direitos de dados caso a caso, assumindo que as condições de uso acumuladas para todos e cada um dos conjuntos de dados sejam atendidas. Sendo assim, a licença é uma informação extremamente importante para que pesquisadores possam reutilizar os dados de pesquisa, mas a ferramenta F-UJI não conseguiu recuperá-la em um elemento de metadado apropriado para os *datasets* da UNESP.

Conforme mostra a Figura 16 abaixo, é possível identificar o elemento “*dc.rights.accessRights*” com o valor “Acesso Aberto”, que foi validado no aspecto FsF-A1-01M (nível de acesso/condições de acesso). Mas, dessa vez, a ferramenta F-UJI não estava procurando informações sobre restrições ou direitos de acesso nos metadados; mas sim informações da licença em um metadado apropriado, o que não foi recuperado em nenhum *dataset* da UNESP.

Figura 16 – Metadado “*dc.rights.accessRights”* de um *dataset* depositado no repositório da UNESP



dc.title	Os estudos sobre geometrias não-arquimedianas de Amoroso Costa: uma reconstrução a partir de manuscritos de seu arquivo pessoal	pt
dc.type	Dado de pesquisa	
dc.rights.holder	Rodrigo Rafael Gomes	pt
dc.contributor.institution	Universidade Estadual Paulista (Unesp)	
dc.rights.accessRights	Acesso aberto	

Fonte: *print screen* retirado do repositório da UNESP (2022).

Fazendo novamente uma comparação com o conjunto de dados (https://doi.org/10.7910/DVN/WR4S9I) depositado no *Harvard Dataverse*, que foi pontuado como “avançado” no aspecto FsF-R1.1-01M, é possível observar diferenças. Ao exportar os metadados em JSON percebe-se a indicação da licença da seguinte forma:

license":{"name":"CC01.0","uri":"http://creativecommons.org/publicdomain/zero/1.0"}.

A indicação da licença CC0, da forma como aparece para o usuário do repositório em *License/Data Use Agreement,* é seguida pelo *link* da respectiva licença, permitindo que o usuário a acesse e que a máquina faça a interpretação. O mesmo acontece para os dois *datasets* do repositório da UFSCar que pontuaram como “avançado”. Logo, essa é outra questão que a UNESP pode avaliar a pertinência para melhorias, lembrando que para maximizar a reutilização é preciso atribuir uma licença bem definida e reconhecida internacionalmente, de modo que as condições de reutilização sejam claramente comunicadas (HODSON *et al.*, 2018), tanto para os usuários humanos como para os agentes computacionais.

O próximo aspecto (FsF-R1.2-01M) avaliou informações de proveniência nos metadados, e todos os 44 *datasets* obtiveram um nível moderado de descrição conforme mostra o Quadro 13. Alguns dos elementos que foram extraídos pela ferramenta foram: "*creator", "publication_date"* e *"contributor*". Entretanto, a ferramenta não conseguiu validar a presença, nos metadados, de informações de proveniência usando ontologias formais, o que impossibilitou a pontuação dos *datasets* como avançado, mesmo caso da UFSCar, revelando outro ponto fraco em comum. Mas, mesmo *datasets* com altos níveis de *FAIRness*, como o “https://doi.org/10.7910/DVN/WR4S9I” (*Harvard Dataverse),* também não foram validados nesse teste. Supõe-se, portanto, que não é uma funcionalidade de fácil implementação.

Quanto ao quarto aspecto (FsF-R1.3-01M), a ferramenta verificou se um padrão de metadados específico da comunidade é detectado usando *namespaces* e se está listado no registro Re3data. O Re3data, como já visto, visa contribuir para o estabelecimento de um ecossistema de repositórios de dados mais coerente e integrado. O nível inicial obtido pelos 44 *datasets* diz respeito ao uso de um padrão

de metadados multidisciplinar, endossado pela comunidade (*RDA Metadata Standards Catalog*), que é detectado através de *namespaces.*

A ferramenta F-UJI detectou o padrão *Dublin Core*, assim como no caso da UFSCar. Esse resultado já era esperado, uma vez que o DC é o esquema de metadados descritivos pré-definido do *DSpace*. Ele é um padrão que se adequa a vários domínios (como é o caso em repositórios multidisciplinares) e seu uso é um bom indicativo quando se pensa na interoperabilidade entre repositórios digitais.

Por fim, quanto ao último aspecto (FsF-R1.3-02D), a ferramenta verificou se o formato do arquivo de dados é aberto aberto/de longo prazo/científico. Todos os conjuntos de dados foram avaliados como incompletos, ou seja, o formato do arquivo não era fornecido nos metadados/não estava contemplado na lista de formatos de arquivos de longo prazo, formatos de arquivos abertos ou formatos de arquivos científicos.

Após o *download* dos *datasets* foi possível observar que a maioria se tratava de ‘PDF’, ‘mp4’ ou ‘xlsx’, formato proprietário. Assim, seu acesso fica à mercê de ferramentas pagas, que podem ser descontinuadas. Mas, no manual de autoarquivamento da UNESP, tem-se a seguinte mensagem: “formato submetido deve ser PDF para arquivos textuais; para outros tipos de arquivos recomenda-se, se possível, a disponibilização em formatos abertos (não proprietários)”, demonstrando a preocupação com o acesso e reuso dos dados. Logo, assim como no caso da UFSCar, supõe-se que os usuários só optariam pelos formatos abertos se houvesse uma política mandatória da instituição, o que pode ser estudado pela equipe para aumentar a reutilização dos *datasets*.

Seguindo a tendência da primeira avaliação, percebe-se um ponto fraco quanto à reutilização, em que a maior nota obtida foi 1 de 10 (Gráfico 2 e 3). Isso também foi visto na interoperabilidade, em que a maior nota obtida foi 1 de 4. Como já mencionado, para que dados de pesquisa sejam realmente interoperáveis e reutilizáveis, “[...] não devem apenas ser devidamente licenciados, mas os métodos para acessá-los e/ou baixá-los também devem ser bem descritos e, de preferência, totalmente automatizados e usando protocolos bem estabelecidos” (KALINAUSKAITĖ, 2017, p. 22). Não basta pensar os repositórios para os usuários humanos, que conseguem mais facilmente interpretar os dados, é preciso pensá-los também para as máquinas.

Sendo assim, os resultados obtidos por meio da ferramenta F-UJI demonstram uma falta de aderência às quatro facetas FAIR no geral, com foco na interoperabilidade, em que todas as notas variaram entre 0 e 1 (sendo o máximo 4), e reutilização, em que todas as notas foram 1 (sendo o máximo 10). Com relação ao acesso, todos os *datasets* da amostra foram pontuados com a mesma nota: 1,5 de 3, exatamente igual ao repositório da UFSCar. Quanto à encontrabilidade, todos os conjuntos de dados receberam a nota 3,5 (sendo o máximo 7), um resultado melhor do que o encontrado na UFSCar, onde 80% dos *datasets* obteve nota igual a 3.

Resumindo, todos os conjuntos de dados da UNESP foram validados quanto ao uso de identificadores persistentes. Isso é um excelente indicativo porque, segundo Hodson *et al*. (2018), metadados, identificadores persistentes e acesso por protocolos de comunicação padronizados são elementos básicos em uma escala de 5 estrelas em direção ao FAIR, conforme mostra a Figura 17. Para os autores, os dados podem ser vistos dentro de um espectro de aderência e a proposta abaixo coloca os 15 princípios existentes em uma escala. Isso ajuda a visualizar quais pontos podem ser colocados como prioridade pelas equipes para aumentar o nível de *FAIRness* dos dados de pesquisa avaliados.

Figura 17 – Níveis de FAIR: uma escala cinco estrelas

*	**The basic core**: metadata, PID & access	F2. data are described with rich metadata F1. (meta)data are assigned a globally unique and persistent identifier A1. (meta)data are retrievable by their identifier using a standardized communications protocol
**	**Enhanced access**: catalogues for discovery, standard (controlled) access & licences	F4. (meta)data are registered or indexed in a searchable resource A1.1. the protocol is free, open and universally implementable A1.2. the protocol allows for an authentication and authorization procedure, where necessary R1.1. (meta)data are released with a clear and accessible data usage license
***	**Use of standards:** for metadata and data	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation R1.3. (meta)data meet domain relevant community standards F3. metadata clearly and explicitly include the identifier of the data it describes
****	**Rich, FAIR metadata**	R1. (meta)data are richly described with a plurality of accurate and relevant attributes I2. (meta)data uses vocabularies that follow FAIR principles
*****	**Provenance and additional context**	R1.2 (meta)data are associated with data provenance I3. (meta)data include qualified references to other (meta)data A2. metadata are accessible, even when the data are no longer available

Fonte: HODSON *et al*. (2018).

O uso de metadados para descrições ricas dos dados (também no nível de elementos básicos) é um ponto a ser avaliado e aperfeiçoado pela equipe, buscando

aumentar o nível semântico para as máquinas e a descoberta dos *datasets*. Mas, devido ao autoarquivamento, talvez seja necessário reformular o próprio manual disponível no repositório para instruir os usuários quanto ao preenchimento de mais elementos de metadados. É importante frisar que, diferente dos resultados de pesquisa como artigos, livros e dissertações, os dados de pesquisa exigem uma descrição/contextualização muito maior para serem interpretados. Metadados ricos são, portanto, imprescindíveis.

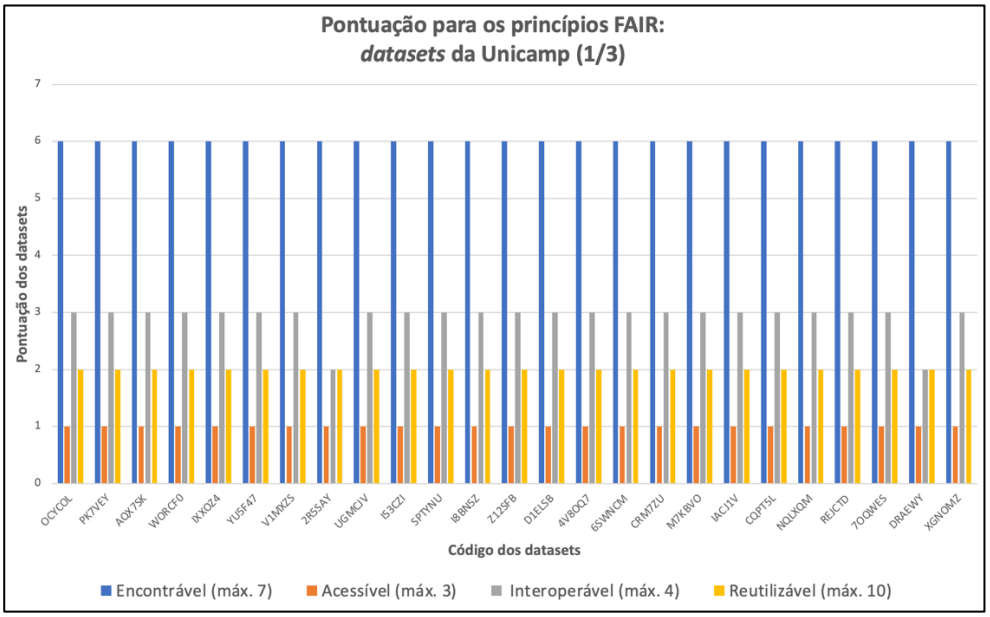
O recomendado é que a equipe gestora do repositório da UNESP avalie os pontos destacados nesta seção e decida quais são pertinentes para a instituição, uma vez que “Alguns dos princípios serão triviais a certos domínios de pesquisa e problemáticos a outros; portanto, cada campo de pesquisa precisa definir o que significa ser FAIR e decidir as medidas apropriadas para avaliar isso” (SILVA; RODRIGUES, 2021, p. 129). Seria vantajoso realizar essa avaliação enquanto a coleção ainda é recente e pequena, podendo ser mais facilmente aperfeiçoada, evitando assim problemas futuros principalmente de falta de padronização e interoperabilidade. O aperfeiçoamento quanto aos princípios FAIR é um passo inicial para que o repositório possa almejar certificações internacionais de qualidade, trazendo maior visibilidade para a instituição e seus pesquisadores.

5.1.3 Repositório institucional da Universidade Estadual de Campinas (Unicamp)

O próximo repositório avaliado foi o da Unicamp, que continha um total de 68 conjuntos de dados depositados no momento da coleta. Como já visto, cada faceta do FAIR possui uma nota máxima, conforme o número indicado na legenda do gráfico: encontrável (máximo sete), acessível (máximo três), interoperável (máximo quatro) e reutilizável (máximo dez). É possível observar pelos Gráficos 4, 5 e 6 que nenhum dos *datasets* do repositório da Unicamp conseguiu atingir nota máxima, o que se aplica a todos os princípios. A maior nota obtida em encontrável foi 6 de 7, em acessível 1 de 3, em interoperável 3 de 4 e em reutilizável foi 2 de 10.

Entretanto, comparado aos resultados da UFSCar e da UNESP, os da Unicamp apresentam uma melhora significativa, principalmente quanto à encontrabilidade e à interoperabilidade, em que as pontuações alcançadas pelos *datasets* chegam muito próximo da nota máximo possível, um ótimo indicativo.

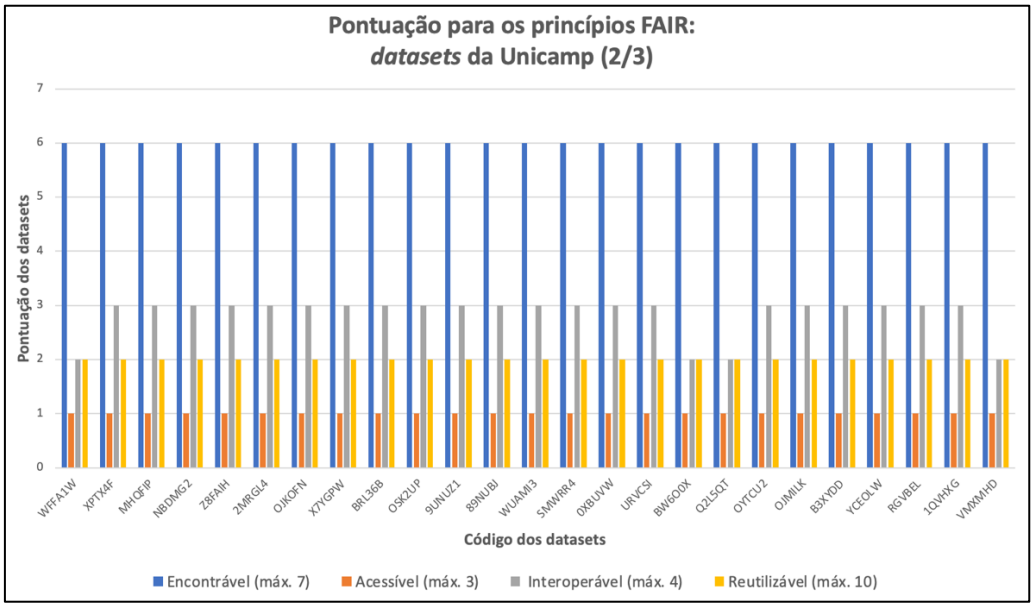
Gráfico 4 - Pontuação para os princípios FAIR: *datasets* da Unicamp (1/3)



Fonte: Elaborado pela autora (2022).

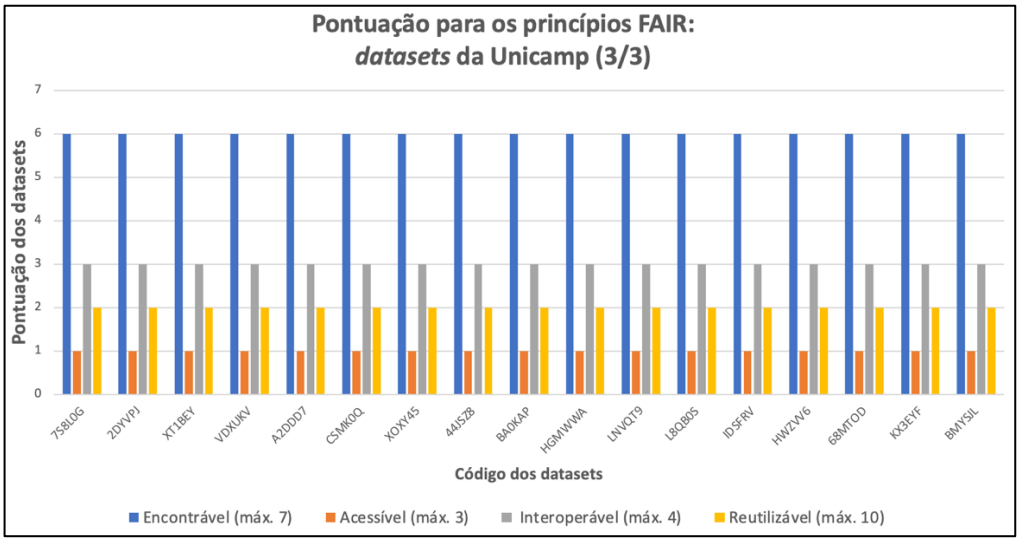
Analisando os gráficos é possível perceber que a única faceta em que o repositório da Unicamp obteve pontuações mais baixas foi em “acessível”. Todos os seus *datasets* receberam a nota 1, enquanto a UFSCar e UNESP pontuaram igualmente: 1,5. Logo, a instituição pode escolher dar atenção especial para a acessibilidade de seus conjuntos de dados.

Gráfico 5 - Pontuação para os princípios FAIR: *datasets* da Unicamp (2/3)



Fonte: Elaborado pela autora (2022).

Gráfico 6 - Pontuação para os princípios FAIR: *datasets* da Unicamp (3/3)



Fonte: Elaborado pela autora (2022).

Os 68 *datasets* da Unicamp foram também avaliados em uma escala de 0 a 100% quanto à aderência geral aos princípios FAIR, sendo a maior pontuação igual a 50%. No total, 62 conjuntos de dados apresentaram uma aderência de 50% ao FAIR, enquanto os outros seis obtiveram um nível de 45% de *FAIRness,* que também é um bom resultado quando comparado às demais instituições do estudo. Conforme mostram os gráficos acima, 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.

De todos os seis repositórios da amostra, o da Unicamp obteve a maior aderência ao FAIR. De cara, uma diferença que pode ser destacada é a escolha do uso do *software Dataverse* em vez do *DSpace.* Rocha *et al*. (2018) explicam que como esse *software* foi criado pensando em repositórios de dados, "[...] a representação e a gestão automatizada dos conjuntos de dados são estruturadas por meio do conceito *dataset*, que inclui dados, metadados de citação, metadados específicos, documentação adicional, citação, gerenciamento de versões etc.".

O *Dataverse*, inclusive, permite tanto a citação para conjuntos de dados inteiros como também para os arquivos de dados, com um identificador para cada arquivo, o que auxilia na descoberta dos *datasets*. Silva e Rodrigues (2021, p. 127) defendem que "[...] uma plataforma de repositório como o *Dataverse* pode facilitar muito a criação

de dados científicos FAIR". Essa é uma semelhança que o repositório da Unicamp guarda com o *Harvard Dataverse,* citado no documento oficial do FAIR. Rodrigues Dias e Lourenço (2022, p. 315) corroboram essa teoria ao alegar que:

> [...] os repositórios implantados a partir do *software* DSpace possuem baixo nível FAIR. Ademais, elas podem apontar o contrário quanto aos repositórios hospedados no Dataverse, onde esses estão próximos de satisfazer a maioria dos critérios baseados nos Princípios FAIR estabelecidos pela ferramenta.

Indo para uma análise mais minuciosa, a ferramenta apresenta o nível (inicial, moderado, avançado ou incompleto) de conformidade quanto a cada aspecto dos princípios FAIR. Dessa forma é possível verificar em quais implementações o repositório pode investir e o que pode ser feito, na prática, para melhorar essas pontuações. Abaixo tem-se os resultados para a **encontrabilidade** dos dados.

Quadro 14 - Aderência quanto à Encontrabilidade (Unicamp)

Código	FsF-F1-01D	FsF-F1-02D	FsF-F2-01M	FsF-F3-01M	FsF-F4-01M
EP4NGO	avançado	avançado	avançado	incompleto	avançado
OCYCOL	avançado	avançado	avançado	incompleto	avançado
PK7VEY	avançado	avançado	avançado	incompleto	avançado
AQX7SK	avançado	avançado	avançado	incompleto	avançado
WORCF0	avançado	avançado	avançado	incompleto	avançado
IXXOZ4	avançado	avançado	avançado	incompleto	avançado
YU5F47	avançado	avançado	avançado	incompleto	avançado
V1MXZS	avançado	avançado	avançado	incompleto	avançado
2R5SAY	avançado	avançado	avançado	incompleto	avançado
UGMCJV	avançado	avançado	avançado	incompleto	avançado
IS3CZI	avançado	avançado	avançado	incompleto	avançado
SPTYNU	avançado	avançado	avançado	incompleto	avançado
I8BN5Z	avançado	avançado	avançado	incompleto	avançado
Z12SFB	avançado	avançado	avançado	incompleto	avançado
D1ELSB	avançado	avançado	avançado	incompleto	avançado
4V8OQ7	avançado	avançado	avançado	incompleto	avançado
6SWNCM	avançado	avançado	avançado	incompleto	avançado
CRM7ZU	avançado	avançado	avançado	incompleto	avançado
M7KBVO	avançado	avançado	avançado	incompleto	avançado
IACJ1V	avançado	avançado	avançado	incompleto	avançado

CQPT5L	avançado	avançado	avançado	incompleto	avançado
NQLXQM	avançado	avançado	avançado	incompleto	avançado
REJCTD	avançado	avançado	avançado	incompleto	avançado
7OQWES	avançado	avançado	avançado	incompleto	avançado
DRAEWY	avançado	avançado	avançado	incompleto	avançado
XGNOMZ	avançado	avançado	avançado	incompleto	avançado
WFFA1W	avançado	avançado	avançado	incompleto	avançado
XPTX4F	avançado	avançado	avançado	incompleto	avançado
MHQFIP	avançado	avançado	avançado	incompleto	avançado
NBDMG2	avançado	avançado	avançado	incompleto	avançado
Z8FAIH	avançado	avançado	avançado	incompleto	avançado
2MRGL4	avançado	avançado	avançado	incompleto	avançado
OJKOFN	avançado	avançado	avançado	incompleto	avançado
X7YGPW	avançado	avançado	avançado	incompleto	avançado
BRL36B	avançado	avançado	avançado	incompleto	avançado
OSK2UP	avançado	avançado	avançado	incompleto	avançado
9UNUZ1	avançado	avançado	avançado	incompleto	avançado
89NUBJ	avançado	avançado	avançado	incompleto	avançado
WUAMI3	avançado	avançado	avançado	incompleto	avançado
SMWRR4	avançado	avançado	avançado	incompleto	avançado
0XBUVW	avançado	avançado	avançado	incompleto	avançado
URVCSI	avançado	avançado	avançado	incompleto	avançado
BW6O0X	avançado	avançado	avançado	incompleto	avançado
Q2L5QT	avançado	avançado	avançado	incompleto	avançado
OYTCU2	avançado	avançado	avançado	incompleto	avançado
OJMILK	avançado	avançado	avançado	incompleto	avançado
B3XYDD	avançado	avançado	avançado	incompleto	avançado
YCEOLW	avançado	avançado	avançado	incompleto	avançado
RGVBEL	avançado	avançado	avançado	incompleto	avançado
1QVHXG	avançado	avançado	avançado	incompleto	avançado
VMXMHD	avançado	avançado	avançado	incompleto	avançado
7S8L0G	avançado	avançado	avançado	incompleto	avançado
2DYVPJ	avançado	avançado	avançado	incompleto	avançado
XT1BEY	avançado	avançado	avançado	incompleto	avançado
VDXUKV	avançado	avançado	avançado	incompleto	avançado
A2DDD7	avançado	avançado	avançado	incompleto	avançado
CSMK0Q	avançado	avançado	avançado	incompleto	avançado
XOXY45	avançado	avançado	avançado	incompleto	avançado

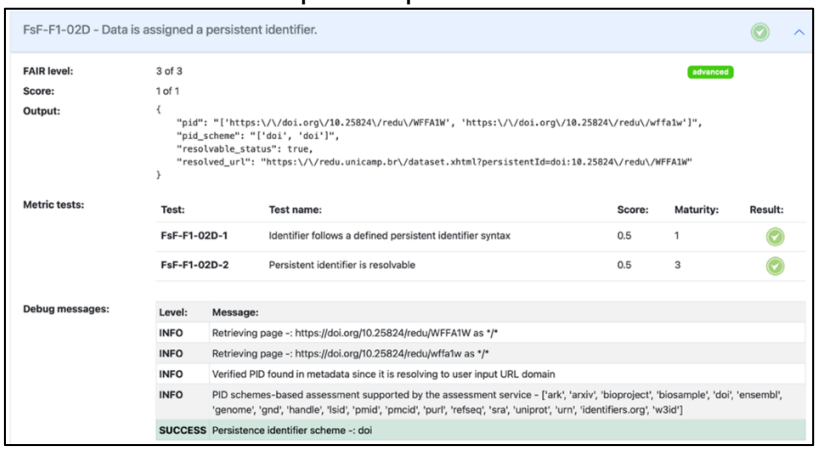
44JSZ8	avançado	avançado	avançado	incompleto	avançado
BA0KAP	avançado	avançado	avançado	incompleto	avançado
HGMWWA	avançado	avançado	avançado	incompleto	avançado
LNVQT9	avançado	avançado	avançado	incompleto	avançado
L8QB0S	avançado	avançado	avançado	incompleto	avançado
IDSFRV	avançado	avançado	avançado	incompleto	avançado
HWZVV6	avançado	avançado	avançado	incompleto	avançado
68MTOD	avançado	avançado	avançado	incompleto	avançado
KX3EYF	avançado	avançado	avançado	incompleto	avançado
BMYSJL	avançado	avançado	avançado	incompleto	avançado

Fonte: elaborado pela autora (2022).

Percebe-se que todos os *datasets* foram bem avaliados pela ferramenta (alcançando o nível avançado em quatro de cinco aspectos avaliados), e só não foram validados em um aspecto, que será visto em mais detalhes. O primeiro (FsF-F1-01D), que diz respeito a: "são atribuídos identificadores globalmente exclusivos aos dados", foi avaliado como "avançado" para todos os conjuntos de dados da Unicamp, seguindo a mesma tendência dos dois repositórios anteriores. A ferramenta auxiliar verificou inicialmente se o identificador atribuído segue uma sintaxe de identificador único definido e, de fato, todos os *datasets* cumprem o requisito.

Já o segundo aspecto (FsF-F1-02D) diz respeito a: "são atribuídos identificadores persistentes aos dados", e novamente todos os *datasets* foram avaliados como “avançado”. A ferramenta foi capaz de localizar um identificador que segue uma sintaxe de identificador persistente, conforme mostra a Figura 18.

Figura 18 – *Output* exibido pela ferramenta F-UJI ao avaliar um conjunto de dados da Unicamp no aspecto FsF-F1-02D



Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

No caso, foi recuperado um DOI para todos os conjuntos de dados do repositório, um ótimo indicativo. Sayão (2007, p.73) explica que:

> Além de identificar, o DOI associa aos objetos digitais dados estruturados – informações bibliográficas e comerciais atualizáveis. Para tanto, o DOI estabelece uma infra-estrutura ampla, cuja perspectiva é ligar os usuários aos fornecedores de conteúdo, dentro de um escopo que considera sempre a facilitação das práticas de comércio eletrônico de conteúdos e a viabilidade da gestão automática de *copyright.*

Além disso, o DOI "[...] foi projetado tendo como perspectiva a máxima conformidade com as normas e padrões correntes voltados para a interoperabilidade" (SAYÃO, 2007, p. 76), o que ainda trás benefícios mais benefícios pensando no FAIR. Sendo assim, dentre os três repositórios já mencionados, apenas a UFSCar não obteve a validação neste aspecto para todos os conjuntos de dados.

O terceiro aspecto avaliado (FsF-F2-01M) se refere a: "metadados incluem elementos essenciais (*'creator', 'title', 'publisher', 'publication_date', 'summary', 'keywords', 'object_identifier', 'object_type'*) para dar suporte à encontrabilidade dos objetos digitais". Como já visto na Figura 17, metadados ricos são elementos básicos quando se pensa na implementação gradual do FAIR. Os *datasets* da UNESP obtiveram nível inicial, os da UFSCar moderado e os da Unicamp avançado, outro ponto forte da instituição.

A ferramenta foi capaz de extrair os metadados "*creator", "keywords", object_identifier*", "*object_type*", "*publication_date", "publisher", "summary"* e *"title"*. Além dos metadados essenciais para a descrição e citação do *dataset*, a ferramenta também validou os seguintes pontos: os metadados foram disponibilizados por meio de métodos comuns da *web*; eles são incorporados no código XHTML/HTML da página de destino e são acessíveis por meio de negociação de conteúdo. De acordo com a ferramenta F-UJI foi possível encontrar metadados *schema.org* JSON-LD na página html; metadados *Dublin Core*; metadados *schema.org* por meio de negociação de conteúdo; e metadados *Datacite*.

O fato dos *datasets* serem descritos de forma que podem ser facilmente encontrados e citados é extremamente positivo, uma vez que incentiva o seu reuso. Inclusive, assim como no caso do repositório da UFSCar e da UNESP, o da Unicamp

também oferece a citação dos conjuntos de dados já pronta, permitindo que o pesquisador apenas copie, dando os devidos créditos ao autor dos dados.

O próximo aspecto (FsF-F3-01M) diz respeito a: "os metadados incluem o identificador do dado que descrevem". A ferramenta avaliou duas características: 1) os metadados contêm informações relacionadas ao conteúdo dos dados (nome do arquivo, tamanho, tipo) e 2) os metadados contêm um PID ou URL que indica a localização do conteúdo dos dados para *download*. Mas, assim como no caso da UFSCar e da UNESP, o identificador dos dados foi tido como ausente e, por isso, todos os *datasets* foram avaliados como "incompleto". Esse é, portanto, um ponto fraco comum entre as três instituições avaliadas até o momento. Recomenda-se novamente o estudo do caso do *dataset* depositado no *Harvard Dataverse*, que conseguiu aderir em nível “avançado” a este aspecto, e foi citado no tópico da UFSCar.

Por fim, no aspecto FsF-F4-01M, os *datasets* foram testados quanto aos metadados fornecidos de uma maneira que os principais mecanismos de pesquisa pudessem inseri-los em seus catálogos (exemplos: JSON-LD, *Dublin Core*, RDFa), e todos foram avaliados como “avançado”, mesmo resultado encontrado nos conjuntos de dados da UFSCar e da UNESP. É um ponto forte comum entre as três instituições.

Inclusive, assim como no caso da UNESP, o repositório da Unicamp também se encontra indexado no Re3data, aumentando a encontrabilidade e a visibilidade dos conjuntos de dados da instituição. É recomendado, portanto, que todos os repositórios busquem estar mapeados no diretório.

Sendo assim, o único aspecto de encontrabilidade que os *datasets* da Unicamp não aderiram foi o FsF-F3-01M, que diz respeito aos metadados incluírem o identificador dos dados que descrevem. É um ponto que a instituição pode trabalhar para aumentar a descoberta de seus dados. Apesar disso, em todos os outros quatro aspectos (FsF-F1-01D, FsF-F1-02D, FsF-F2-01M e FsF-F4-01M), os *datasets* da Unicamp aderiram em nível avançado ao FAIR, um excelente resultado.

Quadro 15 - Aderência quanto à Acessibilidade (Unicamp)

Código	FsF-A1-01M	FsF-A1-03D	FsF-A1-02M
EP4NGO	incompleto	incompleto	avançado
OCYCOL	incompleto	incompleto	avançado
PK7VEY	incompleto	incompleto	avançado

AQX7SK	incompleto	incompleto	avançado
WORCF0	incompleto	incompleto	avançado
IXXOZ4	incompleto	incompleto	avançado
YU5F47	incompleto	incompleto	avançado
V1MXZS	incompleto	incompleto	avançado
2R5SAY	incompleto	incompleto	avançado
UGMCJV	incompleto	incompleto	avançado
IS3CZI	incompleto	incompleto	avançado
SPTYNU	incompleto	incompleto	avançado
I8BN5Z	incompleto	incompleto	avançado
Z12SFB	incompleto	incompleto	avançado
D1ELSB	incompleto	incompleto	avançado
4V8OQ7	incompleto	incompleto	avançado
6SWNCM	incompleto	incompleto	avançado
CRM7ZU	incompleto	incompleto	avançado
M7KBVO	incompleto	incompleto	avançado
IACJ1V	incompleto	incompleto	avançado
CQPT5L	incompleto	incompleto	avançado
NQLXQM	incompleto	incompleto	avançado
REJCTD	incompleto	incompleto	avançado
7OQWES	incompleto	incompleto	avançado
DRAEWY	incompleto	incompleto	avançado
XGNOMZ	incompleto	incompleto	avançado
WFFA1W	incompleto	incompleto	avançado
XPTX4F	incompleto	incompleto	avançado
MHQFIP	incompleto	incompleto	avançado
NBDMG2	incompleto	incompleto	avançado
Z8FAIH	incompleto	incompleto	avançado
2MRGL4	incompleto	incompleto	avançado
OJKOFN	incompleto	incompleto	avançado
X7YGPW	incompleto	incompleto	avançado
BRL36B	incompleto	incompleto	avançado
OSK2UP	incompleto	incompleto	avançado
9UNUZ1	incompleto	incompleto	avançado
89NUBJ	incompleto	incompleto	avançado
WUAMI3	incompleto	incompleto	avançado
SMWRR4	incompleto	incompleto	avançado
0XBUVW	incompleto	incompleto	avançado

URVCSI	incompleto	incompleto	avançado
BW6O0X	incompleto	incompleto	avançado
Q2L5QT	incompleto	incompleto	avançado
OYTCU2	incompleto	incompleto	avançado
OJMILK	incompleto	incompleto	avançado
B3XYDD	incompleto	incompleto	avançado
YCEOLW	incompleto	incompleto	avançado
RGVBEL	incompleto	incompleto	avançado
1QVHXG	incompleto	incompleto	avançado
VMXMHD	incompleto	incompleto	avançado
7S8L0G	incompleto	incompleto	avançado
2DYVPJ	incompleto	incompleto	avançado
XT1BEY	incompleto	incompleto	avançado
VDXUKV	incompleto	incompleto	avançado
A2DDD7	incompleto	incompleto	avançado
CSMK0Q	incompleto	incompleto	avançado
XOXY45	incompleto	incompleto	avançado
44JSZ8	incompleto	incompleto	avançado
BA0KAP	incompleto	incompleto	avançado
HGMWWA	incompleto	incompleto	avançado
LNVQT9	incompleto	incompleto	avançado
L8QB0S	incompleto	incompleto	avançado
IDSFRV	incompleto	incompleto	avançado
HWZVV6	incompleto	incompleto	avançado
68MTOD	incompleto	incompleto	avançado
KX3EYF	incompleto	incompleto	avançado
BMYSJL	incompleto	incompleto	avançado

Fonte: elaborado pela autora (2022).

Já quanto à **acessibilidade,** os conjuntos de dados foram avaliados em três aspectos: 1) Os metadados contêm o nível de acesso e as condições de acesso aos dados (FsF-A1-01M); 2) Os dados são acessíveis por meio de um protocolo de comunicação padronizado (FsF-A1-03D) e 3) Os metadados são acessíveis por meio de um protocolo de comunicação padronizado (FsF-A1-02M).

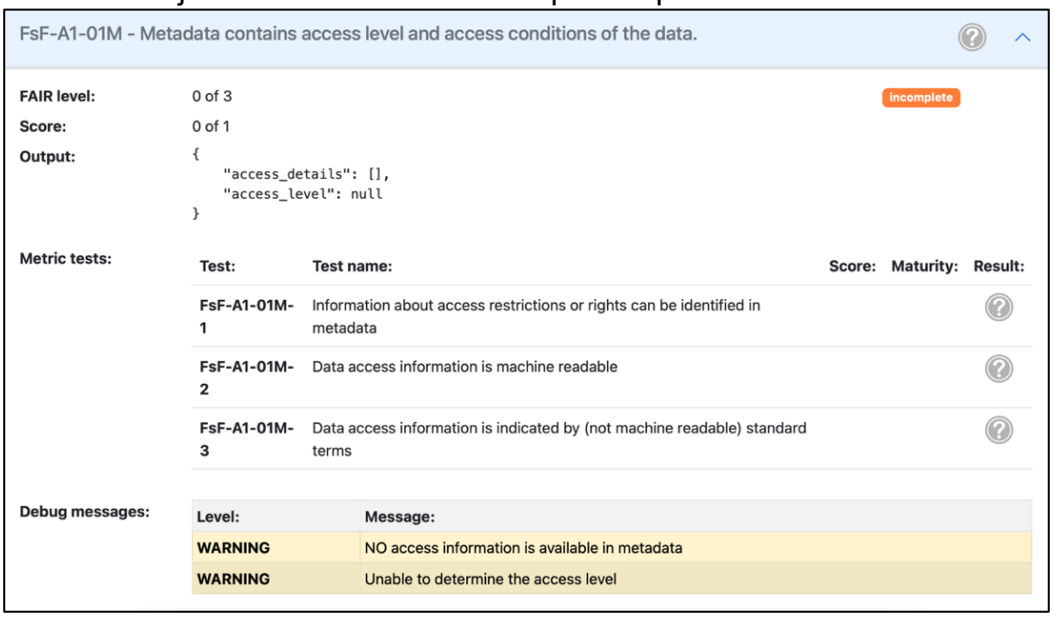
Nota-se que todos os conjuntos de dados da Unicamp só aderiram a um dos três aspectos avaliados pela ferramenta F-UJI. A UFSCar aderiu a dois (FsF-A1-01M em nível inicial e FsF-A1-02M em nível avançado) e a UNESP também aderiu a dois

(FsF-A1-01M em nível inicial e FsF-A1-02M em nível avançado). Ou seja, a acessibilidade, comparando as três instituições, representa um ponto fraco a ser trabalhado pela Unicamp.

Para o aspecto FsF-A1-01M, a ferramenta auxiliar verificou se os *datasets* possuíam informações sobre restrições ou direitos de acesso nos metadados, mas não foi capaz de encontrar essas informações nos metadados, conforme mostra a Figura 19.

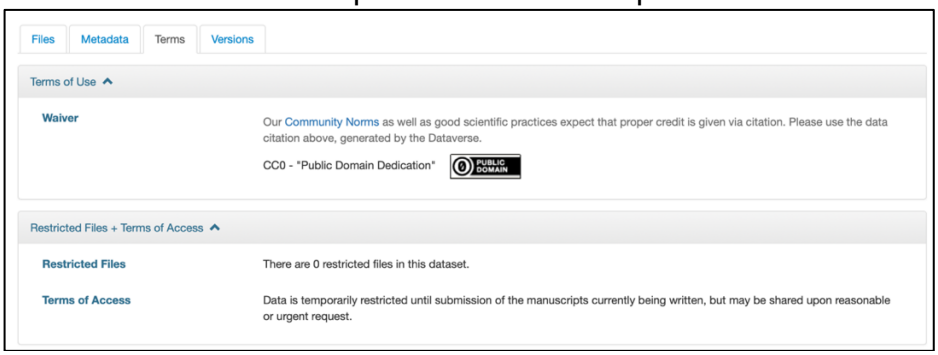
Apesar desse resultado, em alguns conjuntos de dados da Unicamp é possível encontrar os termos de acesso e restrições na aba de "*terms*", conforme mostra a Figura 20. Recomenda-se que essa informação (termos de acesso) esteja disponível para todos os conjuntos de dados do repositório. Isso norteia os usuários que desejam acessar os dados de pesquisa, fornecendo informações sobre restrições e seus motivos, dando a oportunidade de o pesquisador entrar em contato com o criador do *dataset*. Mas nem para esses *datasets* a ferramenta F-UJI foi capaz de recuperar as condições de acesso, o que pode apontar para a falta da indicação num elemento de metadado adequado para o processamento por máquina.

Figura 19 – Métricas e *debug messages* exibidas pela ferramenta F-UJI para um conjunto de dados da Unicamp no aspecto FsF-A1-01M



Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

Figura 20 – Termos de acesso disponibilizado num conjunto de dados depositado no repositório da Unicamp



Fonte: *print screen* retirado do repositório da Unicamp (2022).

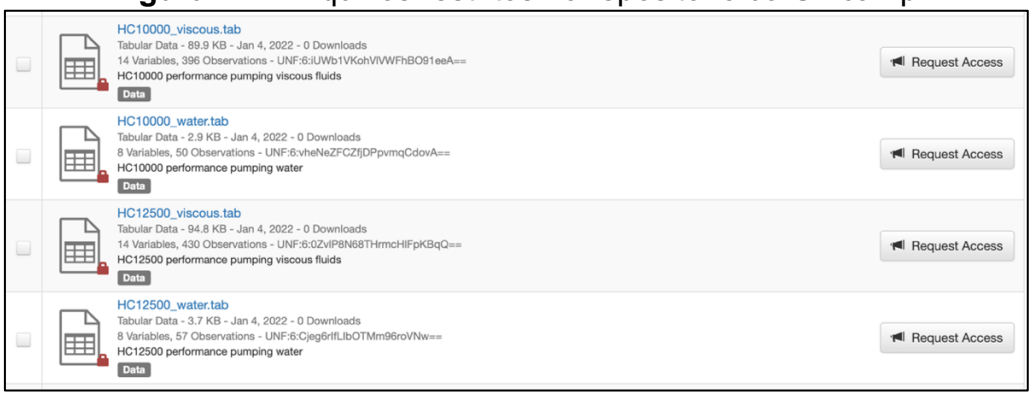
Esse resultado, inclusive, também foi encontrado numa avaliação adicional feita com *datasets* do *Harvard Dataverse* e do *SciELO Data*, que utilizam o *software Dataverse*, assim como a Unicamp. Após testes rápidos com alguns conjuntos de dados desses dois repositórios adicionais, foi possível notar uma tendência de não aderência ao aspecto FsF-A1-01M, mesmo tendo alcançado níveis de *FAIRness* altos. Cabe à equipe gestora do repositório da Unicamp determinar se é preciso aplicar melhorias neste aspecto, uma vez que na publicação original dos princípios FAIR Wilkinson *et al*. (2016) explicam que é preciso fornecer as condições exatas de acesso aos dados e, idealmente, a acessibilidade é especificada de forma que uma máquina possa entender automaticamente os requisitos e, em seguida, executá-los automaticamente ou alertar o usuário sobre os requisitos.

Na página de perguntas frequentes (FAQ) do repositório encontra-se uma seção dedicada ao “Acesso aos Dados do REDU via internet”, onde a instituição afirma que todos os dados são públicos e não há como estabelecer critérios para o acesso e para o uso dos dados depositados. Todos os dados publicados via REDU se tornam automaticamente acessíveis por qualquer pesquisador do mundo inteiro. Entretanto, é possível habilitar períodos de carência, mas somente mediante solicitação por e-mail por parte do pesquisador. Não há como disponibilizar os dados apenas para usuários selecionados, de acordo com o FAQ.

Na Figura 21 é possível observar o caso de *datasets* restritos. Assim como no repositório da UNESP, os metadados continuam disponíveis, mesmo que os arquivos não estejam, uma boa prática importante para o FAIR. Mas, diferente da UNESP, no repositório da Unicamp existe um botão para solicitar diretamente acesso aos arquivos, e somente após o *login* é possível fazer a solicitação. De acordo com Dias, Anjos e Rodrigues (2019), é essencial que essa prática seja adotada, fornecendo

informações nos metadados que possibilitem identificar os indivíduos ou as instituições detentoras dos dados.

Figura 21 – Arquivos restritos no repositório da Unicamp



Fonte: *print screen* retirado do repositório da Unicamp (2022).

Ainda nas perguntas frequentes há uma seção destinada às questões legais e éticas ligadas ao depósito, e a instituição afirma que a publicação de dados pessoais de indivíduos de algum projeto de pesquisa está sujeita a regras éticas de cada disciplina e à Lei Geral de Proteção de Dados. Tais dados poderão ser publicados desde que devidamente pseudonimizados de forma a impedir a identificação dos indivíduos. Além disso, a publicação precisa ser autorizada pelos comitês de ética do projeto. Ou seja, há a preocupação em alertar quanto à importância de não expor dados sensíveis no repositório, cabendo ao pesquisador anonimizar esses dados.

Já em relação ao segundo aspecto (FsF-A1-03D), que diz respeito aos dados serem acessíveis por meio de um protocolo de comunicação padronizado, a ferramenta examinou se os metadados incluíam um *link* para dados com base em protocolos de comunicação da *web* padronizados (*standard_data_protocol*), o que não foi encontrado, mesmo resultado da UFSCar e da UNESP. Esse é outro ponto em comum entre as três instituições já citadas neste trabalho.

Entretanto, todos os *datasets* da Unicamp são identificados por *links* que começam com https://, ou seja, são recuperados utilizando um protocolo de comunicação padrão, o HTTPS. Essa questão também foi, inclusive, testada para os metadados (FsF-A1-02M), e a ferramenta F-UJI conseguiu localizar o protocolo padrão para acesso aos metadado: HTTPS. Logo, quanto aos aspectos FsF-A1-03D e FsF-A1-02M, todos os três repositórios e seus *datasets* pontuaram igualmente, indicando uma consistência (e tendência) nos resultados encontrados em “acessível”.

Cabe à equipe gestora definir o nível de prioridade (e de viabilidade) para as melhorias necessárias.

Os *datasets* também foram avaliados quanto à **interoperabilidade** e, apesar de ser um elemento desafiador, o repositório da Unicamp obteve resultados melhores que os dois anteriores. Foram avaliados três aspectos: 1) Os metadados são representados usando uma linguagem de representação de conhecimento formal (FsF-I1-01M); 2) Metadados usam recursos semânticos (FsF-I1-02M) e 3) Os metadados incluem *links* entre os dados e suas entidades relacionadas (FsF-I3-01M).

É importante citar que, como defendem Hodson *et al.* (2018), o desenvolvimento de componentes tecnológicos compatíveis com FAIR precisa envolver comunidades científicas, especialistas técnicos e outras partes interessadas, e essas equipes multidisciplinares são essenciais quando se pensa em interoperabilidade. Além da implementação, os componentes FAIR precisarão de ciclos de iteração regulares e processos de avaliação para garantir que sejam adequados ao propósito e atendam às necessidades da comunidade.

Quadro 16 - Aderência quanto à Interoperabilidade (Unicamp)

Código	FsF-I1-01M	FsF-I1-02M	FsF-I3-01M
EP4NGO	avançado	inicial	avançado
OCYCOL	avançado	incompleto	avançado
PK7VEY	avançado	inicial	avançado
AQX7SK	avançado	inicial	avançado
W0RCF0	avançado	inicial	avançado
IXXOZ4	avançado	inicial	avançado
YU5F47	avançado	incompleto	avançado
V1MXZS	avançado	incompleto	avançado
2R5SAY	moderado	incompleto	avançado
UGMCJV	avançado	inicial	avançado
IS3CZI	avançado	inicial	avançado
SPTYNU	avançado	inicial	avançado
I8BN5Z	avançado	inicial	avançado
Z12SFB	avançado	inicial	avançado
D1ELSB	avançado	inicial	avançado
4V8OQ7	avançado	inicial	avançado
6SWNCM	avançado	inicial	avançado
CRM7ZU	avançado	inicial	avançado

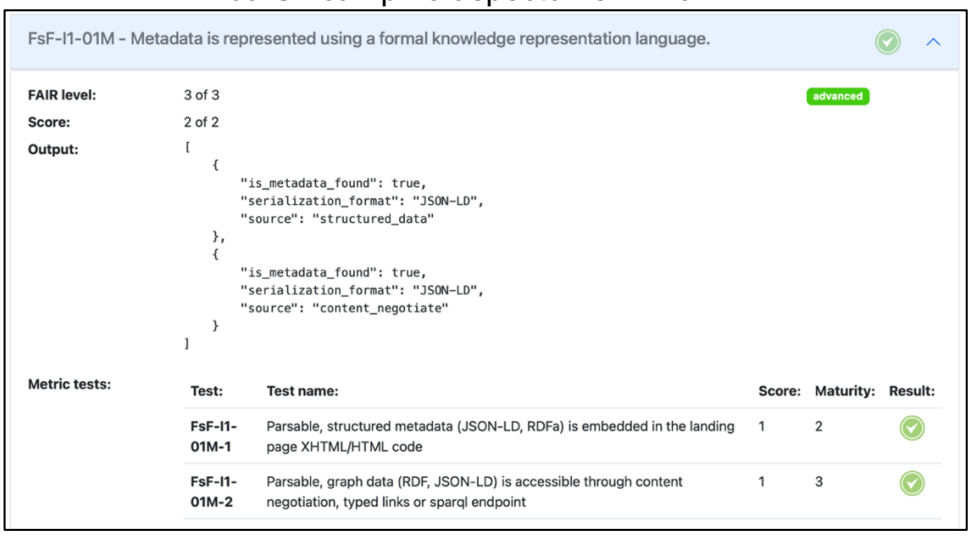
M7KBVO	avançado	inicial	avançado
IACJ1V	avançado	inicial	avançado
CQPT5L	avançado	inicial	avançado
NQLXQM	avançado	inicial	avançado
REJCTD	avançado	inicial	avançado
7OQWES	avançado	incompleto	avançado
DRAEWY	moderado	inicial	avançado
XGNOMZ	avançado	inicial	avançado
WFFA1W	moderado	inicial	avançado
XPTX4F	avançado	inicial	avançado
MHQFIP	avançado	inicial	avançado
NBDMG2	avançado	inicial	avançado
Z8FAIH	avançado	inicial	avançado
2MRGL4	avançado	incompleto	avançado
OJKOFN	avançado	inicial	avançado
X7YGPW	avançado	inicial	avançado
BRL36B	avançado	inicial	avançado
OSK2UP	avançado	inicial	avançado
9UNUZ1	avançado	inicial	avançado
89NUBJ	avançado	inicial	avançado
WUAMI3	avançado	inicial	avançado
SMWRR4	avançado	inicial	avançado
0XBUVW	avançado	inicial	avançado
URVCSI	avançado	inicial	avançado
BW6O0X	moderado	incompleto	avançado
Q2L5QT	moderado	incompleto	avançado
OYTCU2	avançado	inicial	avançado
OJMILK	avançado	inicial	avançado
B3XYDD	avançado	incompleto	avançado
YCEOLW	avançado	incompleto	avançado
RGVBEL	avançado	incompleto	avançado
1QVHXG	avançado	incompleto	avançado
VMXMHD	moderado	incompleto	avançado
7S8L0G	avançado	incompleto	avançado
2DYVPJ	avançado	inicial	avançado
XT1BEY	avançado	inicial	avançado
VDXUKV	avançado	incompleto	avançado
A2DDD7	avançado	incompleto	avançado

CSMK0Q	avançado	incompleto	avançado
XOXY45	avançado	incompleto	avançado
44JSZ8	avançado	incompleto	avançado
BA0KAP	avançado	inicial	avançado
HGMWWA	avançado	inicial	avançado
LNVQT9	avançado	inicial	avançado
L8QB0S	avançado	inicial	avançado
IDSFRV	avançado	inicial	avançado
HWZVV6	avançado	inicial	avançado
68MTOD	avançado	inicial	avançado
KX3EYF	avançado	inicial	avançado
BMYSJL	avançado	inicial	avançado

Fonte: elaborado pela autora (2022).

Quanto ao primeiro aspecto, que diz respeito a uma linguagem de representação de conhecimento formal, a maioria dos conjuntos de dados obteve nível avançado, enquanto seis obtiveram nível moderado, diferente dos *datasets* da UFSCar e da UNESP que foram avaliados como “incompleto”. Os conjuntos de dados foram testados quanto à existência de metadados estruturados e analisáveis (JSON-LD, RDFa) incorporados ao código XHTML/HTML da página de destino, o que foi encontrado (JSON-LD), conforme mostra a Figura 22. Além disso, os dados eram acessíveis por negociação de conteúdo, “*typed links”* ou “*sparql endpoint*”.

Figura 22 – *Output* exibido pela ferramenta F-UJI ao avaliar um conjunto de dados da Unicamp no aspecto FsF-I1-01M

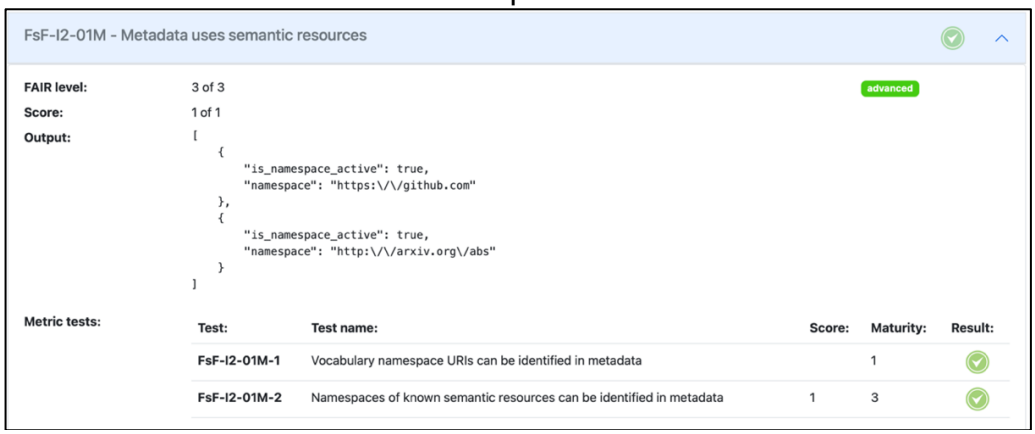


Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

Já no aspecto FsF-I1-02M a ferramenta buscou por uso de *namespaces* de vocabulário e de recursos semânticos nos metadados, e a maioria dos *datasets* pontuaram como inicial, conforme mostra o Quadro 16. Como já visto, o teste buscou extrair *namespaces* dos metadados baseados em RDF, que garantem que determinado conjunto de objetos tenha nomes exclusivos para que possa ser facilmente identificado.

Fazendo uma comparação, ao analisar o conjunto de dados (https://doi.org/10.5281/zenodo.7454759) depositado no *Zenodo*, a ferramenta conseguiu detectar (em nível avançado) os recursos semânticos nos metadados, conforme mostra a Figura 23. Os *namespaces* encontrados foram: 'https://github.com' e 'http://arxiv.org/abs'. Isso fornece contexto para os *datasets* e ainda permite a desambiguação de nomes, determinando o significado preciso dos conceitos e qualidades que os dados representam. Assim os dados são "acionáveis pela máquina", de modo que os valores de um conjunto de atributos possam ser examinados em uma vasta gama de conjuntos de dados com o conhecimento sólido de que os atributos sendo medidos ou representados são de fato os mesmos (HODSON *et al*., 2018). Logo, a equipe gestora do repositório da Unicamp pode estudar a implementação desse aspecto, levando em conta o autoarquivamento, para aumentar o *FAIRness* dos *datasets*.

Figura 23 – *Output* exibido pela ferramenta F-UJI ao avaliar um conjunto de dados do Zenodo no aspecto FsF-I2-01M



Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

Quanto ao último aspecto da interoperabilidade, que diz respeito às ligações entre os dados e suas entidades relacionadas, todos os conjuntos de dados conseguiram pontuar como “avançado”, apresentando assim recursos relacionados

nos metadados indicados por *links* legíveis por máquina. De acordo com a ferramenta foi possível extrair esses recursos do *Schema.org* JSON-LD e *Datacite Search.* Alguns dos tipos de relações encontradas foram: "*isPartOf*" e "*HasPart"*. Isso é um ótimo indicativo para a interoperabilidade dos *datasets*, permitindo que os usuários (tanto humanos quanto computacionais) do repositório naveguem em relações existentes entre vários recursos e entendam melhor o contexto no qual aqueles dados se inserem.

Logo, os conjuntos de dados depositados no repositório da Unicamp apresentam um bom nível de interoperabilidade, com a maioria obtendo a pontuação três (nota máxima sendo quatro). Isso é um excelente resultado quando se pensa na dificuldade de aderência a esta faceta do FAIR. Os *datasets* da UFSCar e da UNESP, por exemplo, variaram entre zero e um, o que explicita o ponto forte da Unicamp quanto à interoperabilidade, podendo até servir como base de estudos para os outros repositórios. Cabe então à equipe da Unicamp decidir sobre a implementação de outras melhorias, conforme foi apontado no aspecto FsF-I2-01M (recursos semânticos nos metadados).

Por fim, todos os conjuntos de dados foram avaliados quanto à **reutilização**. Como já visto, ela está ligada a metadados e documentação ricos que atendam aos padrões da comunidade e forneçam informações sobre a proveniência. A capacidade de humanos e máquinas avaliarem e selecionarem dados com base em critérios relacionados às informações de proveniência é essencial para a reutilização. Para uma boa aderência, os *datasets* deveriam ser descritos em um nível de granularidade significativo para que detalhes e atributos fossem representados, e a descrição deveria ser precisa, clara e relevante, tanto para a comunidade quanto para o domínio do repositório de dados, com o uso de licenças explícitas e claras.

Quadro 17 - Aderência quanto à Reutilização (Unicamp)

Código	FsF-R1-01MD	FsF-R1.1-01M	FsF-R1.2-01M	FsF-R1.3-01M	FsF-R1.3-02D
EP4NGO	inicial	incompleto	moderado	inicial	incompleto
OCYCOL	inicial	incompleto	moderado	inicial	incompleto
PK7VEY	inicial	incompleto	moderado	inicial	incompleto
AQX7SK	inicial	incompleto	moderado	inicial	incompleto
WORCF0	inicial	incompleto	moderado	inicial	incompleto
IXXOZ4	inicial	incompleto	moderado	inicial	incompleto

YU5F47	inicial	incompleto	moderado	inicial	incompleto
V1MXZS	inicial	incompleto	moderado	inicial	incompleto
2R5SAY	inicial	incompleto	moderado	inicial	incompleto
UGMCJV	inicial	incompleto	moderado	inicial	incompleto
IS3CZI	inicial	incompleto	moderado	inicial	incompleto
SPTYNU	inicial	incompleto	moderado	inicial	incompleto
I8BN5Z	inicial	incompleto	moderado	inicial	incompleto
Z12SFB	inicial	incompleto	moderado	inicial	incompleto
D1ELSB	inicial	incompleto	moderado	inicial	incompleto
4V8OQ7	inicial	incompleto	moderado	inicial	incompleto
6SWNCM	inicial	incompleto	moderado	inicial	incompleto
CRM7ZU	inicial	incompleto	moderado	inicial	incompleto
M7KBVO	inicial	incompleto	moderado	inicial	incompleto
IACJ1V	inicial	incompleto	moderado	inicial	incompleto
CQPT5L	inicial	incompleto	moderado	inicial	incompleto
NQLXQM	inicial	incompleto	moderado	inicial	incompleto
REJCTD	inicial	incompleto	moderado	inicial	incompleto
7OQWES	inicial	incompleto	moderado	inicial	incompleto
DRAEWY	inicial	incompleto	moderado	inicial	incompleto
XGNOMZ	inicial	incompleto	moderado	inicial	incompleto
WFFA1W	inicial	incompleto	moderado	inicial	incompleto
XPTX4F	inicial	incompleto	moderado	inicial	incompleto
MHQFIP	inicial	incompleto	moderado	inicial	incompleto
NBDMG2	inicial	incompleto	moderado	inicial	incompleto
Z8FAIH	inicial	incompleto	moderado	inicial	incompleto
2MRGL4	inicial	incompleto	moderado	inicial	incompleto
OJKOFN	inicial	incompleto	moderado	inicial	incompleto
X7YGPW	inicial	incompleto	moderado	inicial	incompleto
BRL36B	inicial	incompleto	moderado	inicial	incompleto
OSK2UP	inicial	incompleto	moderado	inicial	incompleto
9UNUZ1	inicial	incompleto	moderado	inicial	incompleto
89NUBJ	inicial	incompleto	moderado	inicial	incompleto
WUAMI3	inicial	incompleto	moderado	inicial	incompleto
SMWRR4	inicial	incompleto	moderado	inicial	incompleto
0XBUVW	inicial	incompleto	moderado	inicial	incompleto
URVCSI	inicial	incompleto	moderado	inicial	incompleto
BW6O0X	inicial	incompleto	moderado	inicial	incompleto
Q2L5QT	inicial	incompleto	moderado	inicial	incompleto

OYTCU2	inicial	incompleto	moderado	inicial	incompleto
OJMILK	inicial	incompleto	moderado	inicial	incompleto
B3XYDD	inicial	incompleto	moderado	inicial	incompleto
YCEOLW	inicial	incompleto	moderado	inicial	incompleto
RGVBEL	inicial	incompleto	moderado	inicial	incompleto
1QVHXG	inicial	incompleto	moderado	inicial	incompleto
VMXMHD	inicial	incompleto	moderado	inicial	incompleto
7S8L0G	inicial	incompleto	moderado	inicial	incompleto
2DYVPJ	inicial	incompleto	moderado	inicial	incompleto
XT1BEY	inicial	incompleto	moderado	inicial	incompleto
VDXUKV	inicial	incompleto	moderado	inicial	incompleto
A2DDD7	inicial	incompleto	moderado	inicial	incompleto
CSMK0Q	inicial	incompleto	moderado	inicial	incompleto
XOXY45	inicial	incompleto	moderado	inicial	incompleto
44JSZ8	inicial	incompleto	moderado	inicial	incompleto
BA0KAP	inicial	incompleto	moderado	inicial	incompleto
HGMWWA	inicial	incompleto	moderado	inicial	incompleto
LNVQT9	inicial	incompleto	moderado	inicial	incompleto
L8QB0S	inicial	incompleto	moderado	inicial	incompleto
IDSFRV	inicial	incompleto	moderado	inicial	incompleto
HWZVV6	inicial	incompleto	moderado	inicial	incompleto
68MTOD	inicial	incompleto	moderado	inicial	incompleto
KX3EYF	inicial	incompleto	moderado	inicial	incompleto
BMYSJL	inicial	incompleto	moderado	inicial	incompleto

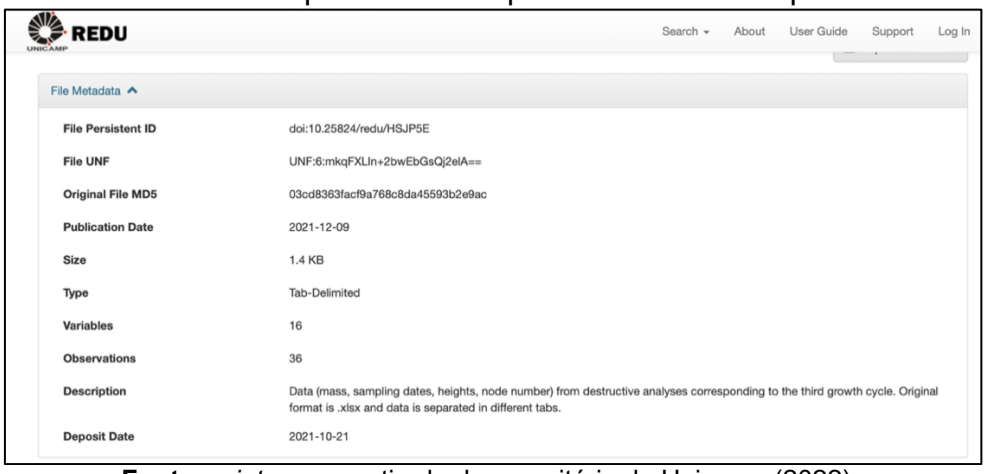
Fonte: elaborado pela autora (2022).

Os dados de pesquisa foram testados em cinco aspectos pela ferramenta auxiliar: 1) Metadados especificam o conteúdo dos dados (FsF-R1-01MD); 2) Os metadados incluem informações de licença sob as quais os dados podem ser reutilizados (FsF-R1.1-01M); 3) Os metadados incluem informações de proveniência sobre a criação ou geração dos dados (FsF-R1.2-01M); 4) Os metadados seguem um padrão recomendado pela comunidade de pesquisa-alvo dos dados (FsF-R1.3-01M) e 5) Os dados estão disponíveis em um formato de arquivo recomendado pela comunidade de pesquisa-alvo (FsF-R1.3-02D). O Quadro 17 mostra a consistência entre os resultados obtidos por todos os 68 *datasets* avaliados, que receberam as mesmas pontuações em todos os aspectos de “reutilizável”. Isso se mostrou uma tendência entre o repositório da UNESP e da Unicamp.

No primeiro aspecto (FsF-R1-01MD) todos os *datasets* pontuaram como “inicial”. Ou seja, a ferramenta encontrou um nível mínimo de informações sobre o conteúdo dos dados nos metadados, mais precisamente o tipo do recurso, que é justamente “*dataset*”. Assim como no caso da UFSCar e da UNESP, o “tipo” de recurso é “dado de pesquisa”. Mas, ao contrário das outras duas instituições citadas, a Unicamp possui um repositório destinado exclusivamente aos dados de pesquisa, em vez de adotar um modelo híbrido. Ou seja, é facilmente dedutível que o tipo de recurso é justamente um dado de pesquisa. Isso abre margem para descrições mais ricas por parte da instituição, contextualizando o tipo de dado depositado. Entretanto, o uso de “*dataset*” no campo “*type*” pode auxiliar na troca de informações entre sistemas diferentes, como os outros repositórios da amostra.

A ferramenta, assim como no caso da UFSCar e da UNESP, não conseguiu extrair o tamanho do arquivo. Mas, ao entrar nos arquivos dos *datasets*, é possível encontrar informações como tamanho, tipo, descrição, data de depósito, data de publicação, etc. (Figura 24). Ou seja, as informações estão declaradas, mas a ferramenta não foi capaz de extrair.

Figura 24 – Visualização do nome, tamanho e formato do arquivo de um conjunto de dados depositado no repositório da Unicamp



Fonte: *print screen* retirado do repositório da Unicamp (2022).

Isso pode estar ligado ao fato delas não estarem disponibilizadas nos metadados (do item) que podem ser extraídos em *Schema.org* JSON-LD, diferente do metadado “*type*”, que estava presente. Deduz isso por causa do conjunto de dados do *Harvard Dataverse* (https://doi.org/10.7910/DVN/WR4S9I) citado anteriormente e que pontuou como avançado nesse aspecto. Nos metadados exportados do *dataset*

em si é possível encontrar campos como “*name*”, “*type*”, “*format*” e “*size*” dos arquivos do conjunto de dados.

A equipe gestora do repositório pode avaliar essa aplicação de acordo com as necessidades da comunidade. Isso porque, apesar de um usuário humano conseguir encontrar facilmente essas informações, um computador pode ter dificuldades de processar automaticamente, o que é preconizado pelo FAIR. Ou seja, não basta fornecer as informações sobre o *dataset*, elas precisam ser legíveis por máquina para serem FAIR.

O segundo aspecto (FsF-R1.1-01M) diz respeito aos metadados incluírem ou não informações de licença sob as quais os dados podem ser reutilizados. Essa é uma informação extremamente importante para que pesquisadores possam reutilizar os dados de pesquisa, mas a ferramenta F-UJI não conseguiu recuperá-la em um elemento de metadado apropriado para nenhum dos *datasets* da Unicamp. A ferramenta retornou a seguinte mensagem: “Parece que a representação schema.org da licença está incorreta, pulando o teste”.

Entretanto, na aba “*terms*” do item, é possível encontrar os termos de uso com uma licença CC atribuída: CC0 - "*Public Domain Dedication*". Também foi possível encontrar, nos termos de uso de alguns itens, outra informação: “Esta licença permite que os reutilizadores distribuam, remixem, adaptem e desenvolvam o material em qualquer meio ou formato, desde que a atribuição seja dada ao criador. A licença permite o uso comercial”, em vez da licença CC em si. Já para alguns outros *datasets* há apenas a mensagem: “Nenhuma isenção foi selecionada para este conjunto de dados.”. Ou seja, é importante que a equipe gestora busque instruir a padronização dessa informação, e em metadados legíveis por máquina, conforme preconiza o FAIR.

O próximo aspecto (FsF-R1.2-01M) avaliou informações de proveniência nos metadados, e todos os 68 *datasets* obtiveram um nível moderado de descrição, mesmo resultado encontrado para a UFSCar e UNESP, indicando uma tendência. Alguns dos elementos que foram extraídos pela ferramenta foram: "*creator", "publication_date", "modified_date", "publisher"* e *"contributor*". Essas informações de proveniência relacionadas à criação dos dados de pesquisa foram encontradas, mas a ferramenta não conseguiu validar a presença, nos metadados, de informações de proveniência usando ontologias formais, o que impossibilitou a pontuação dos *datasets* como avançado, assim como no caso da UFSCar e da UNESP.

O aspecto FsF-R1.2-01M é, portanto, um ponto em comum entre os três repositórios avaliados até o momento. Esse nível moderado de informações de proveniência é um bom indicativo, uma vez que sem esse contexto os usuários dos dados podem se sentir receosos de reutilizá-los.

Quanto ao quarto aspecto (FsF-R1.3-01M), a ferramenta verificou se um padrão de metadados específico da comunidade é detectado usando *namespaces* e se está listado no registro Re3data. O nível "inicial" obtido pelos 68 *datasets* diz respeito ao uso de um padrão de metadados multidisciplinar, endossado pela comunidade (*RDA Metadata Standards Catalog*), que é detectado através de *namespaces.*

A ferramenta F-UJI detectou o padrão *Dublin Core* e o *DataCite Metadata Schema,* mas indicou que os padrões de metadados não estão listados no registro Re3data do repositório responsável. A equipe gestora pode avaliar a pertinência dessa melhoria, mas o uso de um padrão validado pela comunidade como o DC, por exemplo, já é um ótimo indicativo, ainda mais quando se pensa na interoperabilidade entre sistemas. Todos os repositórios avaliados até então utilizam justamente o DC, e essa adoção padronizada é o maior foco para este estudo.

Por fim, quanto ao último aspecto (FsF-R1.3-02D), a ferramenta verificou se o formato do arquivo de dados é aberto/de longo prazo/científico. Seguindo a tendência dos outros repositórios da amostra, todos os conjuntos de dados foram avaliados como incompletos, ou seja, o formato do arquivo não era fornecido nos metadados/não estava contemplado na lista de formatos de arquivos de longo prazo, formatos de arquivos abertos ou formatos de arquivos científicos.

Foi feita, então, uma análise manual de alguns conjuntos de dados, encontrado valores como: *Tab-Delimited*, MS Excel (XLSX), MS Word (docx), Adobe PDF, TIFF Image, *Plain Text, Comma Separated Values*, dentre outros. Ou seja, assim como no caso da UFSCar e da UNESP há o depósito de arquivos em formatos proprietários, elaborados em *softwares* pagos e, portanto, não estão disponíveis sem barreiras de acesso e custo como é defendido pelo movimento da Ciência Aberta. Isso pode acabar afetando tanto o acesso quanto a reutilização desses dados ao longo do tempo. Outro ponto que pode ser citado é a diferença de sintaxe apresentada entre os repositórios já avaliados quanto à indicação de formato. Essa falta de padronização pode levar a dificuldades de processamento automático por máquinas, demandando a limpeza desses dados.

Logo, é possível observar algumas diferenças entre os resultados obtidos pela UFSCar/UNESP e os obtidos pela Unicamp. Ao contrário das duas primeiras, a Unicamp apresenta um ponto forte quanto à interoperabilidade, com a maioria de seus conjuntos de dados recebendo a pontuação 3 (sendo o máximo 4). Apesar disso, quanto à acessibilidade, a Unicamp obteve pontuações mais baixas. Enquanto os conjuntos de dados da UFSCar e a UNESP receberam nota 1,5 de 3, os da Unicamp receberam nota 1, indicando um ponto que pode ser trabalhado com prioridade.

Outro ponto a ser observado pela Unicamp, mas que foi uma tendência entre as três universidades avaliadas até então, é a reutilização. Todos os *datasets* da Unicamp receberam nota 2 (sendo o máximo 10) da ferramenta F-UJI, resultado semelhante ao da UFSCar, em que apenas 2 dos 15 conjuntos de dados receberam nota 4 e todos os demais receberam nota 2 de 10. Mas, como já visto, a reutilização é um desafio comum para os dados de pesquisa. Quanto à encontrabilidade, todos os *datasets* da Unicamp receberam nota 6 (sendo o máximo 7), o que também representa um ponto forte do repositório da instituição.

Os pontos que podem ser trabalhados pela equipe da Unicamp estão ligados, em grande parte, aos metadados, o que aponta novamente para a importância que eles assumem no contexto FAIR. Os metadados devem incluir o identificador dos dados que descrevem (nome, tamanho e tipo de arquivo, em formato legível por máquina); o nível e as restrições de acesso devem ser explicitamente declarados nos metadados; é preciso investir em metadados que utilizam recursos semânticos; e os metadados devem incluir explicitamente a licença de uso para os dados de pesquisa (em formato legível por máquina).

Esses são alguns aspectos a serem avaliados pela equipe gestora, que pode definir a pertinência de implementação e investimento, lembrando que o FAIR existe dentro de um espectro. O núcleo básico do FAIR proposto por Hodson *et al.* (2018), por exemplo, diz respeito a metadados de descoberta, identificadores persistentes e acesso aos dados ou, no mínimo, metadados. Escalas como essa ajudam na priorização de investimentos.

Por fim, é interessante citar que a Unicamp declara possuir uma Comissão de Gestão de Dados de Pesquisa (CGDP), com a incumbência de promover a política institucional de dados de pesquisa, propondo ações segundo as boas práticas nacionais e internacionais. A CGDP é ainda responsável pela gestão do Repositório

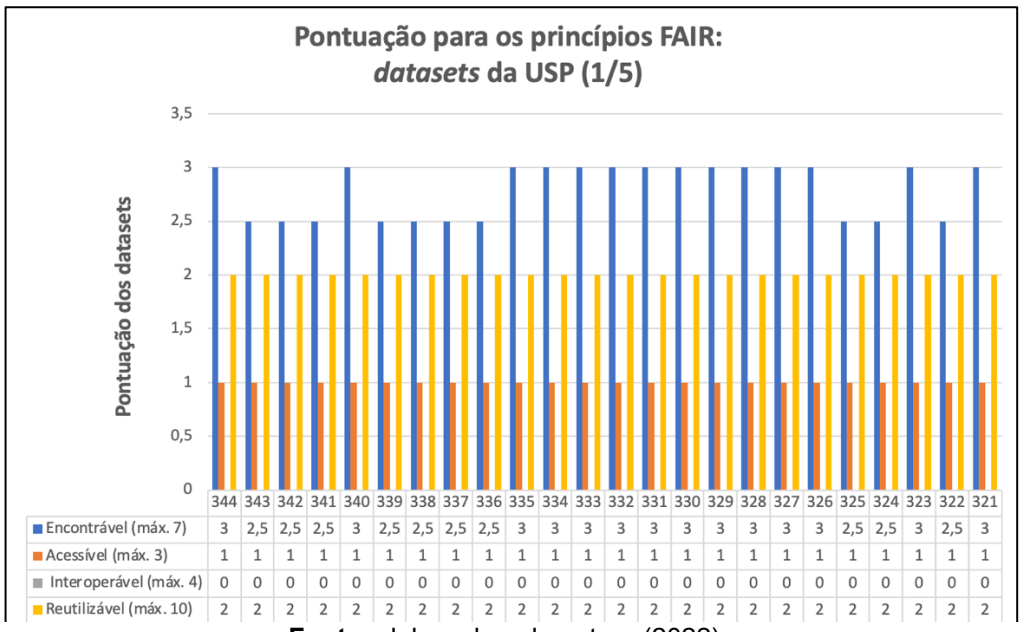
de Dados de Pesquisa (REDU), criado pela Deliberação CCP 006/2020. A instauração da Comissão pode estar relacionada com os melhores resultados encontrados.

5.1.4 Repositório institucional da Universidade de São Paulo (USP)

A quarta avaliação diz respeito aos 118 conjuntos de dados depositados no repositório institucional da USP. A legenda dos gráficos exibe a pontuação máxima possível: encontrável (máximo sete), acessível (máximo três), interoperável (máximo quatro) e reutilizável (máximo dez). As notas foram entregues automaticamente pela ferramenta F-UJI com base nas suas métricas e os resultados podem ser vistos nos gráficos abaixo.

É possível observar pelos Gráficos 7, 8, 9, 10 e 11 que nenhum dos *datasets* do repositório da USP conseguiu atingir nota máxima, o que se aplica a todos os princípios. A maior nota obtida em encontrável foi 3 de 7, em acessível 1 de 3, em interoperável 0 de 4 e em reutilizável foi 2 de 10.

Gráfico 7 - Pontuação para os princípios FAIR: *datasets* da USP (1/5)

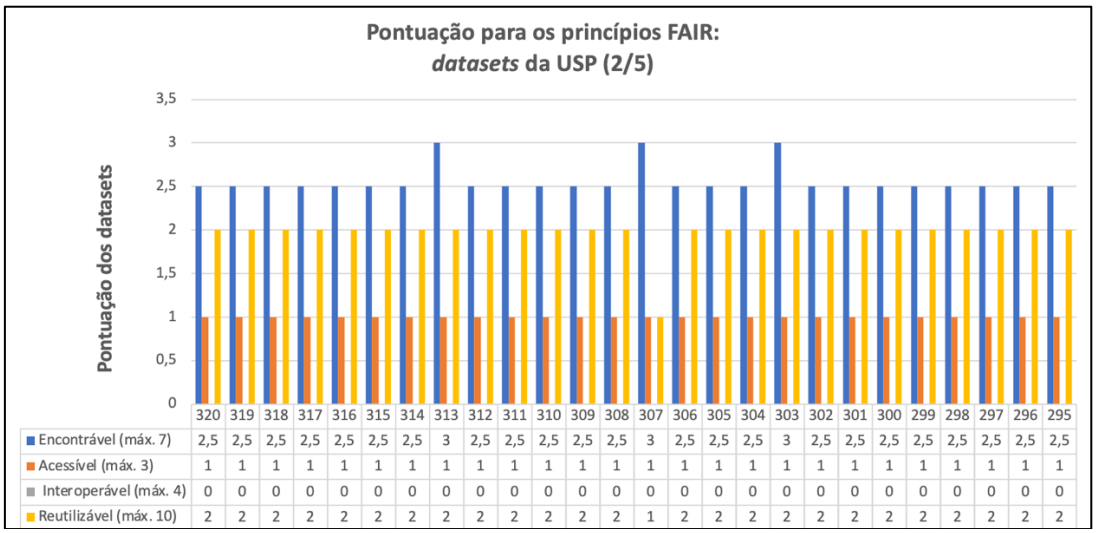


	344	343	342	341	340	339	338	337	336	335	334	333	332	331	330	329	328	327	326	325	324	323	322	321
Encontrável (máx. 7)	3	2,5	2,5	2,5	3	2,5	2,5	2,5	2,5	3	3	3	3	3	3	3	3	3	3	2,5	2,5	3	2,5	3
Acessível (máx. 3)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Interoperável (máx. 4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Reutilizável (máx. 10)	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Fonte: elaborado pela autora (2022).

Os *datasets* obtiveram baixas pontuações no geral, mas principalmente quanto à interoperabilidade e reutilização, revelando áreas que podem ser priorizadas pela instituição. Cabe relembrar que Dunning, Smaele e Böhmero (2017) corroboram esse resultado.

Gráfico 8 - Pontuação para os princípios FAIR: *datasets* da USP (2/5)

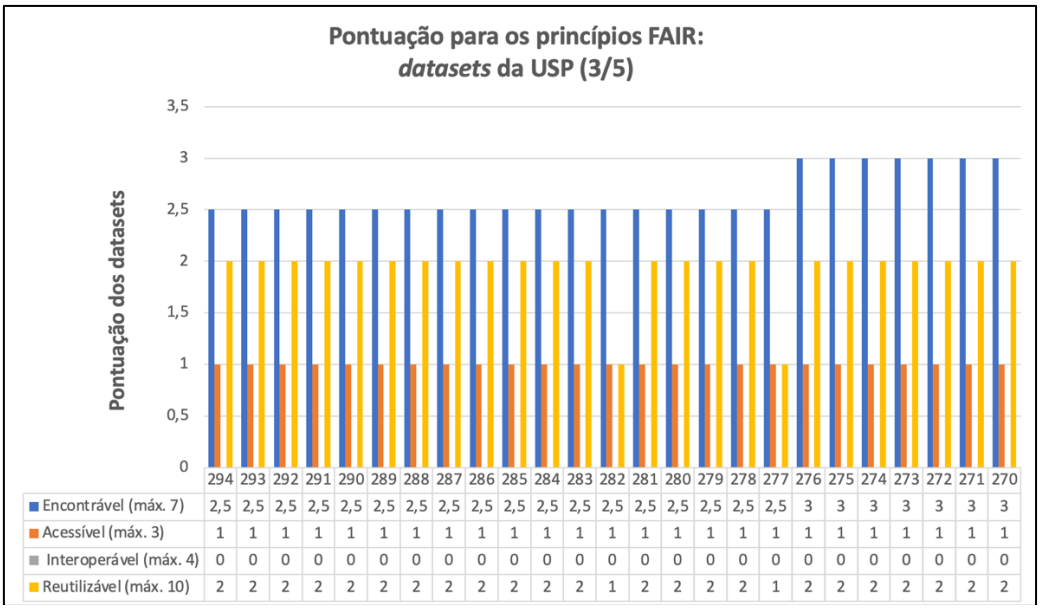


	320	319	318	317	316	315	314	313	312	311	310	309	308	307	306	305	304	303	302	301	300	299	298	297	296	295
Encontrável (máx. 7)	2,5	2,5	2,5	2,5	2,5	2,5	2,5	3	2,5	2,5	2,5	2,5	2,5	3	2,5	2,5	2,5	3	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5
Acessível (máx. 3)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Interoperável (máx. 4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Reutilizável (máx. 10)	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2

Fonte: elaborado pela autora (2022).

Os 118 *datasets* da USP 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.

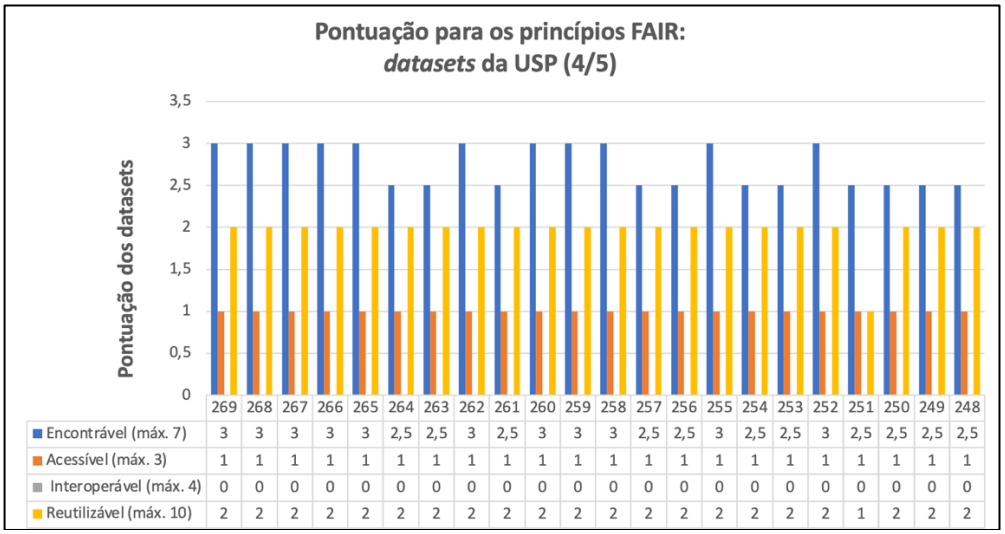
Gráfico 9 - Pontuação para os princípios FAIR: *datasets* da USP (3/5)



	294	293	292	291	290	289	288	287	286	285	284	283	282	281	280	279	278	277	276	275	274	273	272	271	270
Encontrável (máx. 7)	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	3	3	3	3	3	3	3
Acessível (máx. 3)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Interoperável (máx. 4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Reutilizável (máx. 10)	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	1	2	2	2	2	2	2	2

Fonte: elaborado pela autora (2022).

Gráfico 10 - Pontuação para os princípios FAIR: *datasets* da USP (4/5)

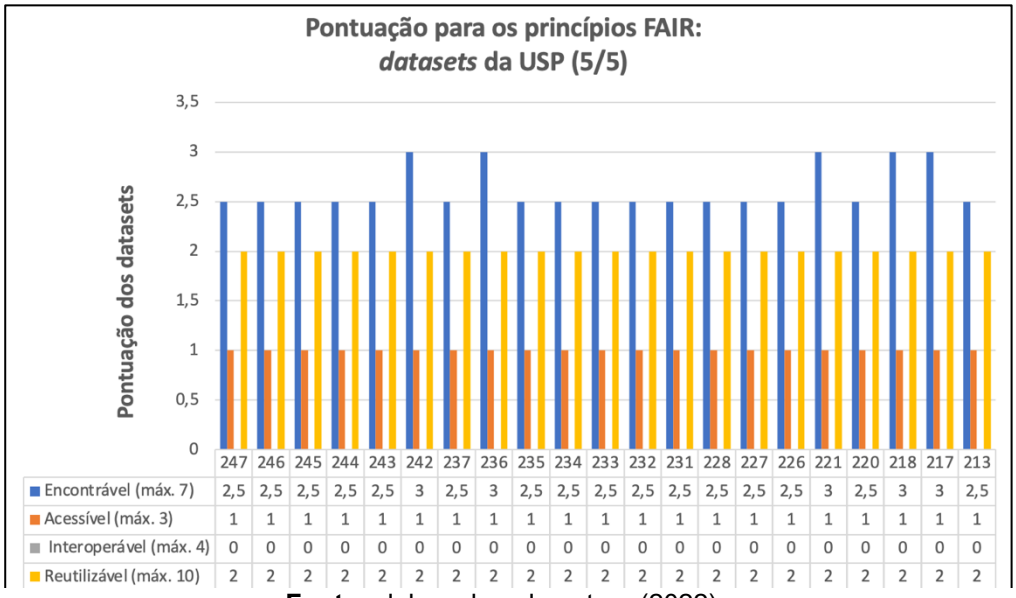


	269	268	267	266	265	264	263	262	261	260	259	258	257	256	255	254	253	252	251	250	249	248
Encontrável (máx. 7)	3	3	3	3	3	2,5	2,5	3	2,5	3	3	3	2,5	2,5	3	2,5	2,5	3	2,5	2,5	2,5	2,5
Acessível (máx. 3)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Interoperável (máx. 4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Reutilizável (máx. 10)	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2

Fonte: elaborado pela autora (2022).

É possível observar novamente o papel de destaque dos metadados para alcançar maiores níveis de conformidade com o FAIR. A própria escala 5 estrelas de níveis de FAIR (HODSON *et al.*, 2018) coloca os metadados num nível básico, como visto anteriormente. Ou seja, já é possível apontar uma questão que pode ser avaliada pela equipe gestora do repositório da USP. Isso será mais bem apresentado nos quadros abaixo.

Gráfico 11 - Pontuação para os princípios FAIR: *datasets* da USP (5/5)



	247	246	245	244	243	242	237	236	235	234	233	232	231	228	227	226	221	220	218	217	213
Encontrável (máx. 7)	2,5	2,5	2,5	2,5	2,5	3	2,5	3	2,5	2,5	2,5	2,5	2,5	2,5	2,5	2,5	3	2,5	3	3	2,5
Acessível (máx. 3)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Interoperável (máx. 4)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Reutilizável (máx. 10)	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Fonte: elaborado pela autora (2022).

Em âmbito geral, percebe-se que o percentual alcançado quanto ao nível de *FAIRness* pelos dados de pesquisa da instituição é baixo, indo de acordo com o que é defendido por Henning *et al*. (2019a). Comparado aos repositórios anteriormente avaliados, o da USP foi o que obteve as menores pontuações pela ferramenta F-UJI.

Indo para uma análise mais minuciosa, a ferramenta apresenta o nível (inicial, moderado, avançado ou incompleto) de conformidade quanto a cada aspecto dos princípios FAIR. Dessa forma é possível verificar em quais implementações o repositório pode investir e o que pode ser feito, na prática, para melhorar essas pontuações. No Quadro 18 tem-se os resultados a **encontrabilidade** dos dados.

Quadro 18 - Aderência quanto à Encontrabilidade (USP)

Código	FsF-F1-01D	FsF-F1-02D	FsF-F2-01M	FsF-F3-01M	FsF-F4-01M
344	avançado	incompleto	moderado	incompleto	avançado
343	avançado	incompleto	inicial	incompleto	avançado
342	avançado	incompleto	inicial	incompleto	avançado
341	avançado	incompleto	inicial	incompleto	avançado
340	avançado	incompleto	moderado	incompleto	avançado
339	avançado	incompleto	inicial	incompleto	avançado
338	avançado	incompleto	inicial	incompleto	avançado
337	avançado	incompleto	inicial	incompleto	avançado
336	avançado	incompleto	inicial	incompleto	avançado
335	avançado	incompleto	moderado	incompleto	avançado
334	avançado	incompleto	moderado	incompleto	avançado
333	avançado	incompleto	moderado	incompleto	avançado
332	avançado	incompleto	moderado	incompleto	avançado
331	avançado	incompleto	moderado	incompleto	avançado
330	avançado	incompleto	moderado	incompleto	avançado
329	avançado	incompleto	moderado	incompleto	avançado
328	avançado	incompleto	moderado	incompleto	avançado
327	avançado	incompleto	moderado	incompleto	avançado
326	avançado	incompleto	moderado	incompleto	avançado
325	avançado	incompleto	inicial	incompleto	avançado
324	avançado	incompleto	inicial	incompleto	avançado
323	avançado	incompleto	moderado	incompleto	avançado
322	avançado	incompleto	inicial	incompleto	avançado
321	avançado	incompleto	moderado	incompleto	avançado
320	avançado	incompleto	inicial	incompleto	avançado

319	avançado	incompleto	inicial	incompleto	avançado
318	avançado	incompleto	inicial	incompleto	avançado
317	avançado	incompleto	inicial	incompleto	avançado
316	avançado	incompleto	inicial	incompleto	avançado
315	avançado	incompleto	inicial	incompleto	avançado
314	avançado	incompleto	inicial	incompleto	avançado
313	avançado	incompleto	moderado	incompleto	avançado
312	avançado	incompleto	inicial	incompleto	avançado
311	avançado	incompleto	inicial	incompleto	avançado
310	avançado	incompleto	inicial	incompleto	avançado
309	avançado	incompleto	inicial	incompleto	avançado
308	avançado	incompleto	inicial	incompleto	avançado
307	avançado	incompleto	moderado	incompleto	avançado
306	avançado	incompleto	inicial	incompleto	avançado
305	avançado	incompleto	inicial	incompleto	avançado
304	avançado	incompleto	inicial	incompleto	avançado
303	avançado	incompleto	moderado	incompleto	avançado
302	avançado	incompleto	inicial	incompleto	avançado
301	avançado	incompleto	inicial	incompleto	avançado
300	avançado	incompleto	inicial	incompleto	avançado
299	avançado	incompleto	inicial	incompleto	avançado
298	avançado	incompleto	inicial	incompleto	avançado
297	avançado	incompleto	inicial	incompleto	avançado
296	avançado	incompleto	inicial	incompleto	avançado
295	avançado	incompleto	inicial	incompleto	avançado
294	avançado	incompleto	inicial	incompleto	avançado
293	avançado	incompleto	inicial	incompleto	avançado
292	avançado	incompleto	inicial	incompleto	avançado
291	avançado	incompleto	inicial	incompleto	avançado
290	avançado	incompleto	inicial	incompleto	avançado
289	avançado	incompleto	inicial	incompleto	avançado
288	avançado	incompleto	inicial	incompleto	avançado
287	avançado	incompleto	inicial	incompleto	avançado
286	avançado	incompleto	inicial	incompleto	avançado
285	avançado	incompleto	inicial	incompleto	avançado
284	avançado	incompleto	inicial	incompleto	avançado
283	avançado	incompleto	inicial	incompleto	avançado
282	avançado	incompleto	inicial	incompleto	avançado
281	avançado	incompleto	inicial	incompleto	avançado

280	avançado	incompleto	inicial	incompleto	avançado
279	avançado	incompleto	inicial	incompleto	avançado
278	avançado	incompleto	inicial	incompleto	avançado
277	avançado	incompleto	inicial	incompleto	avançado
276	avançado	incompleto	moderado	incompleto	avançado
275	avançado	incompleto	moderado	incompleto	avançado
274	avançado	incompleto	moderado	incompleto	avançado
273	avançado	incompleto	moderado	incompleto	avançado
272	avançado	incompleto	moderado	incompleto	avançado
271	avançado	incompleto	moderado	incompleto	avançado
270	avançado	incompleto	moderado	incompleto	avançado
269	avançado	incompleto	moderado	incompleto	avançado
268	avançado	incompleto	moderado	incompleto	avançado
267	avançado	incompleto	moderado	incompleto	avançado
266	avançado	incompleto	moderado	incompleto	avançado
265	avançado	incompleto	moderado	incompleto	avançado
264	avançado	incompleto	inicial	incompleto	avançado
263	avançado	incompleto	inicial	incompleto	avançado
262	avançado	incompleto	moderado	incompleto	avançado
261	avançado	incompleto	inicial	incompleto	avançado
260	avançado	incompleto	moderado	incompleto	avançado
259	avançado	incompleto	moderado	incompleto	avançado
258	avançado	incompleto	moderado	incompleto	avançado
257	avançado	incompleto	inicial	incompleto	avançado
256	avançado	incompleto	inicial	incompleto	avançado
255	avançado	incompleto	moderado	incompleto	avançado
254	avançado	incompleto	inicial	incompleto	avançado
253	avançado	incompleto	inicial	incompleto	avançado
252	avançado	incompleto	moderado	incompleto	avançado
251	avançado	incompleto	inicial	incompleto	avançado
250	avançado	incompleto	inicial	incompleto	avançado
249	avançado	incompleto	inicial	incompleto	avançado
248	avançado	incompleto	inicial	incompleto	avançado
247	avançado	incompleto	inicial	incompleto	avançado
246	avançado	incompleto	inicial	incompleto	avançado
245	avançado	incompleto	inicial	incompleto	avançado
244	avançado	incompleto	inicial	incompleto	avançado
243	avançado	incompleto	inicial	incompleto	avançado
242	avançado	incompleto	moderado	incompleto	avançado

237	avançado	incompleto	inicial	incompleto	avançado
236	avançado	incompleto	moderado	incompleto	avançado
235	avançado	incompleto	inicial	incompleto	avançado
234	avançado	incompleto	inicial	incompleto	avançado
233	avançado	incompleto	inicial	incompleto	avançado
232	avançado	incompleto	inicial	incompleto	avançado
231	avançado	incompleto	inicial	incompleto	avançado
228	avançado	incompleto	inicial	incompleto	avançado
227	avançado	incompleto	inicial	incompleto	avançado
226	avançado	incompleto	inicial	incompleto	avançado
221	avançado	incompleto	moderado	incompleto	avançado
220	avançado	incompleto	inicial	incompleto	avançado
218	avançado	incompleto	moderado	incompleto	avançado
217	avançado	incompleto	moderado	incompleto	avançado
213	avançado	incompleto	inicial	incompleto	avançado

Fonte: elaborado pela autora (2022).

O primeiro aspecto (FsF-F1-01D), que diz respeito a: "são atribuídos identificadores globalmente exclusivos aos dados", foi avaliado como "avançado" para todos os conjuntos de dados da USP. A ferramenta auxiliar verificou inicialmente se o identificador atribuído segue uma sintaxe de identificador único definido (IRI, URL) e, de fato, todos os *datasets* cumprem o requisito. Como visto até o momento, todos os conjuntos de dados dos repositórios da amostra conseguiram pontuar como avançado nesse quesito, sendo, portanto, algo esperado.

No segundo aspecto (FsF-F1-02D), que diz respeito a: "são atribuídos identificadores persistentes aos dados", todos os *datasets* foram avaliados como "incompleto" pela ferramenta F-UJI, que buscou por esquemas baseados em *process ID* (PID) como o DOI. Era preciso que o identificador do *dataset* seguisse uma sintaxe de identificador persistente definida, o que a ferramenta teve dificuldade de encontrar nos conjuntos de dados da USP.

O resultado aqui encontrado é similar ao caso da UFSCar. Logo, recomenda-se que o repositório também busque investir, com certa prioridade, no uso de identificadores como o DOI, em que os metadados anexos estão disponíveis sob uma licença CC0, aberta a todos. Isso garante que os conjuntos de dados sejam encontrados, interpretados e devidamente citados (o que aumenta também sua reusabilidade).

O terceiro aspecto avaliado (FsF-F2-01M) se refere a: "metadados incluem elementos essenciais (*'creator', 'title', 'publisher', 'publication_date', 'summary', 'keywords', 'object_identifier', 'object_type'*) para dar suporte à encontrabilidade dos objetos digitais". Os princípios FAIR dão atenção especial aos metadados, e os *datasets* da USP apresentaram níveis diferentes de conformidade quanto a este critério. Dos 118 conjuntos de dados, 40 foram avaliados como “moderado”, e os demais 78 como “inicial”.

De acordo com a ferramenta F-UJI, um *dataset* avaliado como “moderado” apresentava os metadados: “*creator”, “keywords”, “title”, “summary*”, “*'object_identifier'*”, “*publisher*”, *“publication_date*” e “*object_type*”, exatamente os metadados elencados no aspecto FsF-F2-01M. Apesar disso, a ferramenta não encontrou metadados acessíveis por meio de "*typed links*" ou "*signposting links*". Esses termos, como já visto, dizem respeito às sinalizações e ligações que tornam a *web* acadêmica mais amigável para as máquinas. Já para um *dataset* avaliado como “inicial”, a ferramenta não conseguiu localizar o metadado “*publisher*”.

Em caráter adicional, os conjuntos de dados pontuados como “moderado” apresentavam alguns outros metadados como “*dc.description.sponsorship”, “dc.coverage.spatial”, “dc.coverage.temporal”* e *“dc.format*”. Descrições ricas são essenciais para a descoberta e futuro reuso desses dados de pesquisa e, por isso, recomenda-se que a instituição busque incentivar o uso de metadados ricos, mas, principalmente, o uso de um conjunto mínimo de metadados para os dados de pesquisa.

Nesse sentido, pode-se citar novamente a iniciativa *FAIR Data Point*, voltada para a publicação de metadados de forma FAIR. O FDP está sob responsabilidade da iniciativa GO-FAIR e é um *software* que trabalha os metadados. A equipe gestora pode, então, verificar essa iniciativa para melhor contextualização das descrições de seus conjuntos de dados. Mas, de forma inicial, o repositório pode focar naquele conjunto de metadados elencado pela ferramenta F-UJI no aspecto FsF-F2-01M, que alguns conjuntos de dados já conseguiram apresentar. Esses *datasets* avaliados como “moderado”, inclusive, dispunham dos metadados essenciais para sua citação, outro ponto importante no contexto FAIR.

Isso permite que os *datasets* sejam reutilizados e devidamente citados por outros pesquisadores, tornando o conteúdo científico mais visível. A possibilidade de citá-los é de suma importância na popularização do seu reuso, incentivando os

pesquisadores a adotarem a prática. Mas, diferente dos repositórios da UFSCar, UNESP e Unicamp, o da USP não apresenta a citação já pronta do *dataset* para todos os conjuntos de dados. Oferecer um “como citar” com a devida citação na página do item é uma funcionalidade que a equipe gestora pode avaliar para implementar em todos os *datasets* depositados, tornando assim um padrão de boas práticas. O *dataset* 221, por exemplo, oferece essa informação no elemento ‘*dc.identifier.citation’*.

O próximo aspecto (FsF-F3-01M) diz respeito a: "os metadados incluem o identificador do dado que descrevem". A ferramenta avaliou duas características: 1) os metadados contêm informações relacionadas ao conteúdo dos dados (nome do arquivo, tamanho, tipo) e 2) os metadados contêm um PID ou URL que indica a localização do conteúdo dos dados para *download*. Mas o identificador dos dados foi tido como ausente e, por isso, todos os *datasets* foram avaliados como "incompleto".

Esse foi um ponto em comum entre todos os repositórios avaliados até o momento, apontando para uma tendência que demanda atenção das quatro instituições. Pode citar-se novamente, então, o conjunto de dados (https://doi.org/10.7910/DVN/WR4S9I) depositado no *Harvard Dataverse*, que obteve nível avançado neste aspecto e pode ser uma base de estudo para melhorias nos repositórios da amostra.

Por fim, no aspecto "FsF-F4-01M", os *datasets* foram testados quanto aos metadados fornecidos de uma maneira que os principais mecanismos de pesquisa pudessem inseri-los em seus catálogos (exemplos: JSON-LD, *Dublin Core*, RDFa), e todos foram avaliados como “avançado”. Isso é um bom indicativo para sua encontrabilidade e para a visibilidade da instituição, e é o mesmo resultado encontrado para os outros três repositórios avaliados até o momento, indicando outra tendência.

A ferramenta rodou vários testes para o aspecto "FsF-F4-01M" e alegou que os metadados são fornecidos de forma que os principais mecanismos de pesquisa podem consumí-los em seus catálogos (JSON-LD, Dublin Core, RDFa), um bom indicativo. Apesar disso, os metadados não são registrados nos principais registros de dados de pesquisa (*DataCite*).

O *DataCite Search,* como já visto, oferece uma interface de busca integrada, onde é possível pesquisar, filtrar e extrair todos os detalhes de uma coleção de registros. É um portal para encontrar, acessar e reutilizar dados. Esse é, então, um ponto que garante maior visibilidade aos *datasets* do repositório*,* que poderão ser recuperados por outros mecanismos de busca na *web*.

Já quanto à **acessibilidade**, os conjuntos de dados foram avaliados em três aspectos: 1) Os metadados contêm o nível de acesso e as condições de acesso aos dados (FsF-A1-01M); 2) Os dados são acessíveis por meio de um protocolo de comunicação padronizado (FsF-A1-03D) e 3) Os metadados são acessíveis por meio de um protocolo de comunicação padronizado (FsF-A1-02M). Os resultados de cada conjunto de dados depositado no repositório da USP podem ser vistos no Quadro 19.

Quadro 19 - Aderência quanto à Acessibilidade (USP)

Código	FsF-A1-01M	FsF-A1-03D	FsF-A1-02M
344	incompleto	incompleto	avançado
343	incompleto	incompleto	avançado
342	incompleto	incompleto	avançado
341	incompleto	incompleto	avançado
340	incompleto	incompleto	avançado
339	incompleto	incompleto	avançado
338	incompleto	incompleto	avançado
337	incompleto	incompleto	avançado
336	incompleto	incompleto	avançado
335	incompleto	incompleto	avançado
334	incompleto	incompleto	avançado
333	incompleto	incompleto	avançado
332	incompleto	incompleto	avançado
331	incompleto	incompleto	avançado
330	incompleto	incompleto	avançado
329	incompleto	incompleto	avançado
328	incompleto	incompleto	avançado
327	incompleto	incompleto	avançado
326	incompleto	incompleto	avançado
325	incompleto	incompleto	avançado
324	incompleto	incompleto	avançado
323	incompleto	incompleto	avançado
322	incompleto	incompleto	avançado
321	incompleto	incompleto	avançado
320	incompleto	incompleto	avançado
319	incompleto	incompleto	avançado
318	incompleto	incompleto	avançado
317	incompleto	incompleto	avançado
316	incompleto	incompleto	avançado

315	incompleto	incompleto	avançado
314	incompleto	incompleto	avançado
313	incompleto	incompleto	avançado
312	incompleto	incompleto	avançado
311	incompleto	incompleto	avançado
310	incompleto	incompleto	avançado
309	incompleto	incompleto	avançado
308	incompleto	incompleto	avançado
307	incompleto	incompleto	avançado
306	incompleto	incompleto	avançado
305	incompleto	incompleto	avançado
304	incompleto	incompleto	avançado
303	incompleto	incompleto	avançado
302	incompleto	incompleto	avançado
301	incompleto	incompleto	avançado
300	incompleto	incompleto	avançado
299	incompleto	incompleto	avançado
298	incompleto	incompleto	avançado
297	incompleto	incompleto	avançado
296	incompleto	incompleto	avançado
295	incompleto	incompleto	avançado
294	incompleto	incompleto	avançado
293	incompleto	incompleto	avançado
292	incompleto	incompleto	avançado
291	incompleto	incompleto	avançado
290	incompleto	incompleto	avançado
289	incompleto	incompleto	avançado
288	incompleto	incompleto	avançado
287	incompleto	incompleto	avançado
286	incompleto	incompleto	avançado
285	incompleto	incompleto	avançado
284	incompleto	incompleto	avançado
283	incompleto	incompleto	avançado
282	incompleto	incompleto	avançado
281	incompleto	incompleto	avançado
280	incompleto	incompleto	avançado
279	incompleto	incompleto	avançado
278	incompleto	incompleto	avançado
277	incompleto	incompleto	avançado

276	incompleto	incompleto	avançado
275	incompleto	incompleto	avançado
274	incompleto	incompleto	avançado
273	incompleto	incompleto	avançado
272	incompleto	incompleto	avançado
271	incompleto	incompleto	avançado
270	incompleto	incompleto	avançado
269	incompleto	incompleto	avançado
268	incompleto	incompleto	avançado
267	incompleto	incompleto	avançado
266	incompleto	incompleto	avançado
265	incompleto	incompleto	avançado
264	incompleto	incompleto	avançado
263	incompleto	incompleto	avançado
262	incompleto	incompleto	avançado
261	incompleto	incompleto	avançado
260	incompleto	incompleto	avançado
259	incompleto	incompleto	avançado
258	incompleto	incompleto	avançado
257	incompleto	incompleto	avançado
256	incompleto	incompleto	avançado
255	incompleto	incompleto	avançado
254	incompleto	incompleto	avançado
253	incompleto	incompleto	avançado
252	incompleto	incompleto	avançado
251	incompleto	incompleto	avançado
250	incompleto	incompleto	avançado
249	incompleto	incompleto	avançado
248	incompleto	incompleto	avançado
247	incompleto	incompleto	avançado
246	incompleto	incompleto	avançado
245	incompleto	incompleto	avançado
244	incompleto	incompleto	avançado
243	incompleto	incompleto	avançado
242	incompleto	incompleto	avançado
237	incompleto	incompleto	avançado
236	incompleto	incompleto	avançado
235	incompleto	incompleto	avançado
234	incompleto	incompleto	avançado

233	incompleto	incompleto	avançado
232	incompleto	incompleto	avançado
231	incompleto	incompleto	avançado
228	incompleto	incompleto	avançado
227	incompleto	incompleto	avançado
226	incompleto	incompleto	avançado
221	incompleto	incompleto	avançado
220	incompleto	incompleto	avançado
218	incompleto	incompleto	avançado
217	incompleto	incompleto	avançado
213	incompleto	incompleto	avançado

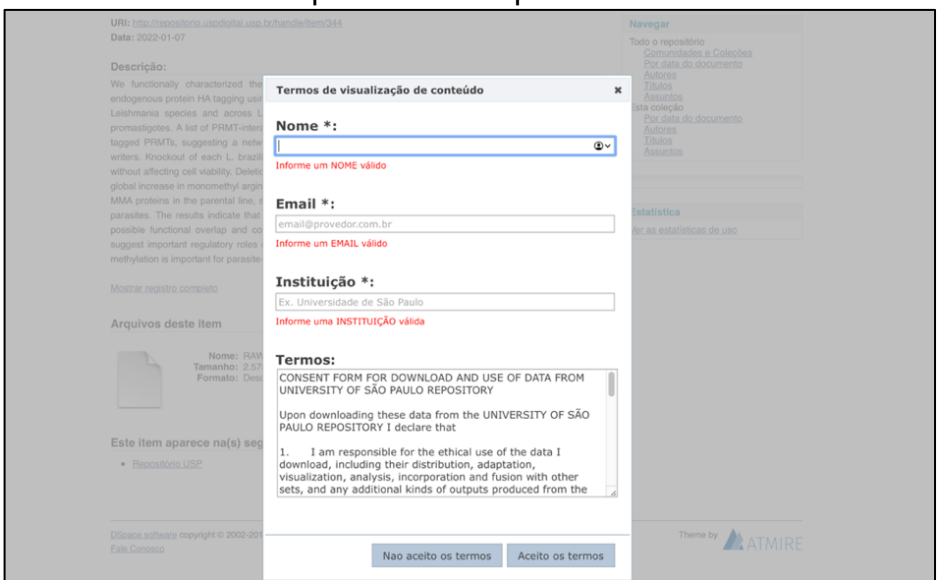
Fonte: elaborado pela autora (2022).

Para o aspecto FsF-A1-01M, a ferramenta auxiliar verificou se os *datasets* possuíam informações sobre restrições ou direitos de acesso nos metadados, mas não conseguiu recuperá-las, o que levou à pontuação como “incompleto” de todos os 118 conjuntos de dados. O resultado foi semelhante ao da Unicamp, indicando um ponto fraco em comum entre os repositórios.

Quando o usuário entra em um conjunto de dados da USP e clica em cima de um arquivo, o sistema exibe os “termos de visualização de conteúdo”, conforme mostra a Figura 25. O repositório solicita então o nome, e-mail e a instituição do usuário que deseja acessar aquele arquivo, junto com os termos de consentimento para baixar e usar os dados. Os termos são exibidos tanto em inglês quanto em português.

Portanto, existe a indicação de termos de acesso, mas é importante relembrar que os princípios FAIR preconizam o processamento automático por máquinas, de tal forma que um computador possa interpretar e determinar facilmente as condições exatas de acesso aos dados, o que não foi validado pela F-UJI. A equipe do repositório pode, então, avaliar a pertinência de adotar uma solução que permita esse processamento automático. De qualquer forma, ao preencher os campos solicitados e aceitar os termos, o usuário humano consegue baixar o arquivo e está ciente das restrições.

Figura 25 – Termos de visualização de conteúdo disponibilizados num conjunto de dados depositado no repositório da USP



Fonte: *print screen* retirado do repositório da USP (2022).

A funcionalidade exibida na Figura 25 foi a única encontrada no sentido de autenticação ou autorização de acesso na interface do repositório. Nos demais repositórios avaliados há uma opção de *login*, geralmente ligada à matrícula institucional do usuário. Isso permite autenticar o proprietário (ou colaborador) de cada conjunto de dados e potencialmente definir direitos específicos do usuário.

Já no segundo aspecto (FsF-A1-03D), que diz respeito aos dados serem acessíveis por meio de um protocolo de comunicação padronizado, a ferramenta examinou se os metadados incluíam um *link* para dados com base em protocolos de comunicação da *web* padronizados (*standard_data_protocol*), o que não foi encontrado. Entretanto, exatamente como nos casos dos outros repositórios avaliados, todos os *datasets* são identificados por *links* que começam com https://, ou seja, são recuperados usando um protocolo de comunicação padrão, o *Hyper Text Transfer Protocol Secure* (HTTPS). Essa questão também foi, inclusive, testada para os metadados (FsF-A1-02M), e a ferramenta F-UJI conseguiu localizar o protocolo padrão para acesso aos metadado: HTTPS.

Sendo assim, todos os conjuntos de dados de todos os quatro repositórios institucionais foram avaliados igualmente nos dois últimos aspectos de “acessível” (FsF-A1-03D e FsF-A1-02M). No segundo princípio do F(A)IR há uma certa consistência de resultados entre os repositórios mapeados no metabuscador da FAPESP, indicando pontos fortes e fracos em comum, que podem ser inclusive

trabalhados em conjunto, adotando soluções padronizadas. Os *datasets* do Zenodo e do *Harvard Dataverse* podem servir como base de estudo para melhorias.

A ferramenta F-UJI também avaliou os *datasets* da USP quanto à **interoperabilidade**, que segundo Henning *et al.* (2019b, p. 401), "[...] é um problema de longo prazo e não trivial, que exigirá um esforço mais criativo na criação dos dados FAIR". E, de fato, se mostrou como um desafio do repositório. Foram avaliados três aspectos: 1) Os metadados são representados usando uma linguagem de representação de conhecimento formal (FsF-I1-01M); 2) Metadados usam recursos semânticos (FsF-I1-02M) e 3) Os metadados incluem *links* entre os dados e suas entidades relacionadas (FsF-I3-01M). Como visto nos gráficos 7 a 11, todos os conjuntos de dados obtiveram nota zero (sendo o máximo 4) em interoperável. O Quadro 20 apresenta os resultados para cada um dos três aspectos citados.

Quadro 20 - Aderência quanto à Interoperabilidade (USP)

Código	FsF-I1-01M	FsF-I1-02M	FsF-I3-01M
344	incompleto	inicial	incompleto
343	incompleto	inicial	incompleto
342	incompleto	inicial	incompleto
341	incompleto	inicial	incompleto
340	incompleto	inicial	incompleto
339	incompleto	inicial	incompleto
338	incompleto	inicial	incompleto
337	incompleto	inicial	incompleto
336	incompleto	inicial	incompleto
335	incompleto	inicial	incompleto
334	incompleto	inicial	incompleto
333	incompleto	inicial	incompleto
332	incompleto	inicial	incompleto
331	incompleto	inicial	incompleto
330	incompleto	inicial	incompleto
329	incompleto	inicial	incompleto
328	incompleto	inicial	incompleto
327	incompleto	inicial	incompleto
326	incompleto	inicial	incompleto
325	incompleto	inicial	incompleto
324	incompleto	inicial	incompleto
323	incompleto	inicial	incompleto

322	incompleto	inicial	incompleto
321	incompleto	inicial	incompleto
320	incompleto	inicial	incompleto
319	incompleto	inicial	incompleto
318	incompleto	inicial	incompleto
317	incompleto	inicial	incompleto
316	incompleto	inicial	incompleto
315	incompleto	inicial	incompleto
314	incompleto	inicial	incompleto
313	incompleto	inicial	incompleto
312	incompleto	inicial	incompleto
311	incompleto	inicial	incompleto
310	incompleto	inicial	incompleto
309	incompleto	inicial	incompleto
308	incompleto	inicial	incompleto
307	incompleto	inicial	incompleto
306	incompleto	inicial	incompleto
305	incompleto	inicial	incompleto
304	incompleto	inicial	incompleto
303	incompleto	inicial	incompleto
302	incompleto	inicial	incompleto
301	incompleto	inicial	incompleto
300	incompleto	inicial	incompleto
299	incompleto	inicial	incompleto
298	incompleto	inicial	incompleto
297	incompleto	inicial	incompleto
296	incompleto	inicial	incompleto
295	incompleto	inicial	incompleto
294	incompleto	inicial	incompleto
293	incompleto	inicial	incompleto
292	incompleto	inicial	incompleto
291	incompleto	inicial	incompleto
290	incompleto	inicial	incompleto
289	incompleto	inicial	incompleto
288	incompleto	inicial	incompleto
287	incompleto	inicial	incompleto
286	incompleto	inicial	incompleto
285	incompleto	inicial	incompleto
284	incompleto	inicial	incompleto

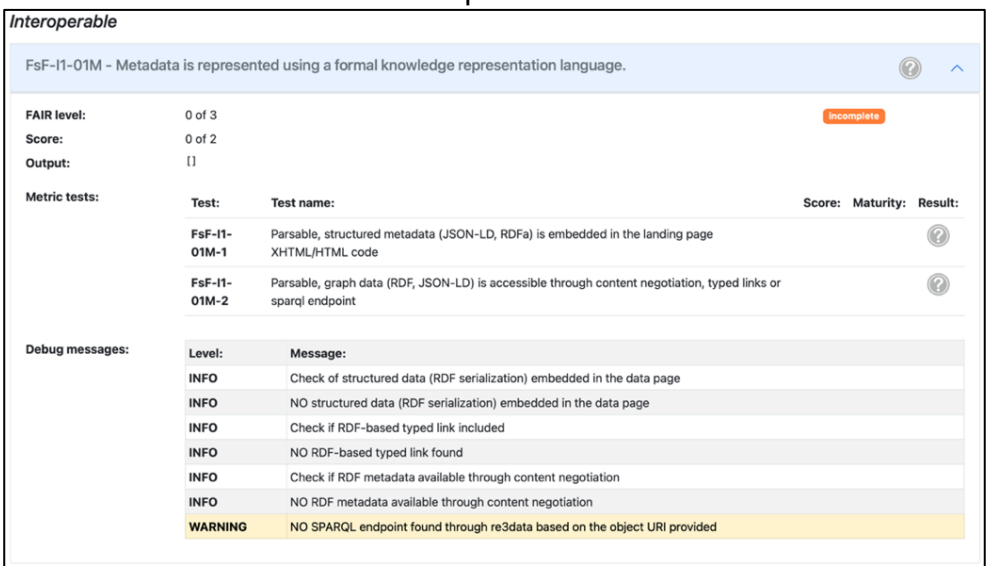
283	incompleto	inicial	incompleto
282	incompleto	inicial	incompleto
281	incompleto	inicial	incompleto
280	incompleto	inicial	incompleto
279	incompleto	inicial	incompleto
278	incompleto	inicial	incompleto
277	incompleto	inicial	incompleto
276	incompleto	inicial	incompleto
275	incompleto	inicial	incompleto
274	incompleto	inicial	incompleto
273	incompleto	inicial	incompleto
272	incompleto	inicial	incompleto
271	incompleto	inicial	incompleto
270	incompleto	inicial	incompleto
269	incompleto	inicial	incompleto
268	incompleto	inicial	incompleto
267	incompleto	inicial	incompleto
266	incompleto	inicial	incompleto
265	incompleto	inicial	incompleto
264	incompleto	inicial	incompleto
263	incompleto	inicial	incompleto
262	incompleto	inicial	incompleto
261	incompleto	inicial	incompleto
260	incompleto	inicial	incompleto
259	incompleto	inicial	incompleto
258	incompleto	inicial	incompleto
257	incompleto	inicial	incompleto
256	incompleto	inicial	incompleto
255	incompleto	inicial	incompleto
254	incompleto	inicial	incompleto
253	incompleto	inicial	incompleto
252	incompleto	inicial	incompleto
251	incompleto	inicial	incompleto
250	incompleto	inicial	incompleto
249	incompleto	inicial	incompleto
248	incompleto	inicial	incompleto
247	incompleto	inicial	incompleto
246	incompleto	inicial	incompleto
245	incompleto	inicial	incompleto

244	incompleto	inicial	incompleto
243	incompleto	inicial	incompleto
242	incompleto	inicial	incompleto
237	incompleto	inicial	incompleto
236	incompleto	inicial	incompleto
235	incompleto	inicial	incompleto
234	incompleto	inicial	incompleto
233	incompleto	inicial	incompleto
232	incompleto	inicial	incompleto
231	incompleto	inicial	incompleto
228	incompleto	inicial	incompleto
227	incompleto	inicial	incompleto
226	incompleto	inicial	incompleto
221	incompleto	inicial	incompleto
220	incompleto	inicial	incompleto
218	incompleto	inicial	incompleto
217	incompleto	inicial	incompleto
213	incompleto	inicial	incompleto

Fonte: elaborado pela autora (2022).

Quanto ao primeiro aspecto, que diz respeito a uma linguagem de representação de conhecimento formal, todos os conjuntos de dados da USP foram avaliados como “incompleto”, mesmo resultado encontrado nos casos da UFSCar e da UNESP. Os conjuntos de dados foram testados quanto à existência de metadados estruturados e analisáveis (JSON-LD, RDFa) incorporados ao código XHTML/HTML da página de destino, o que não foi encontrado, conforme mostra a Figura 26. Além disso, os dados não eram acessíveis por negociação de conteúdo, “*typed links”* ou “*sparql endpoint*”.

Figura 26 – *Output* exibido pela ferramenta F-UJI ao avaliar um conjunto de dados da USP no aspecto FsF-I1-01M



Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

Já no aspecto FsF-I1-02M a ferramenta buscou por uso de *namespaces* de vocabulário e de recursos semânticos nos metadados, e todos os *datasets* pontuaram como inicial, conforme mostra o Quadro 20. Como já visto, o teste buscou extrair *namespaces* dos metadados baseados em RDF*,* que garantem que determinado conjunto de objetos tenha nomes exclusivos para que possa ser facilmente identificado. Isso fornece contexto para os *datasets* e ainda permite a desambiguação de nomes, determinando o significado preciso dos conceitos e qualidades que os dados representam. Todos os repositórios avaliados até o momento obtiveram resultados semelhantes neste aspecto da interoperabilidade, indicando um ponto fraco em comum que pode ser trabalhado em conjunto pelas instituições, garantindo soluções padronizadas.

É importante lembrar que, conforme defendem Hodson *et al*. (2018), para alcançar a interoperabilidade os dados precisam ser descritos usando padrões e vocabulários normativos e reconhecidos pela comunidade que determinam o significado preciso dos conceitos e qualidades que os dados representam. É preciso usar uma linguagem formal e compartilhada. Os autores, ao abordarem o FAIR de modo interdisciplinar, afirmam que as estruturas de interoperabilidade devem ser articuladas de maneira comum e adotar padrões globais sempre que possível para permitir a pesquisa interdisciplinar. Padrões comuns, cruzamentos inteligentes, mecanismos de intermediação e aprendizado de máquina devem ser explorados.

Quanto ao último aspecto da interoperabilidade, que diz respeito às ligações entre os dados e suas entidades relacionadas, todos os conjuntos de dados pontuaram como “incompleto”. Era preciso indicar, de forma explícita, nos metadados do item, os recursos relacionados (e preferencialmente por meio de *links* ou identificadores legíveis por máquina). Uma referência qualificada é uma referência cruzada que explica sua intenção. O objetivo é criar tantos *links* significativos quanto possível entre recursos de (meta)dados para enriquecer o conhecimento contextual sobre os dados.

De forma mais concreta é possível, por exemplo, especificar se um conjunto de dados se baseia em outro conjunto de dados, se são necessários conjuntos de dados adicionais para completar os dados ou se informações complementares são armazenadas em um conjunto de dados diferente. Essas ligações foram encontradas em registros de outros repositórios da amostra como Unicamp (‘*HasPart’, ‘isPartOf’*) e são importantes para, como visto, fornecer maior contexto aos dados de pesquisa depositados.

Logo, os conjuntos de dados do repositório da USP apresentam um baixo nível de interoperabilidade, com todos obtendo a pontuação zero (nota máxima sendo quatro). Os *datasets* da UFSCar e da UNESP variaram entre as notas zero e um, resultado semelhante, e apenas a Unicamp obteve pontuações mais altas (entre 2 e 3). Sendo assim, é possível observar que, de fato, a interoperabilidade se mostra de difícil aderência para os repositórios da amostra. Seria recomendado, então, sua priorização para investimentos, permitindo que os conjuntos de dados interajam com aplicativos ou fluxos de trabalho para análise, armazenamento e processamento.

Por fim, todos os 118 *datasets* foram avaliados quanto à **reutiliação**. O objetivo final do FAIR é otimizar a reutilização dos dados. Para conseguir isso, metadados e dados devem ser bem descritos para que possam ser replicados e/ou combinados em diferentes configurações.

Os dados foram então testados em cinco aspectos pela ferramenta auxiliar: 1) Metadados especificam o conteúdo dos dados (FsF-R1-01MD); 2) Os metadados incluem informações de licença sob as quais os dados podem ser reutilizados (FsF-R1.1-01M); 3) Os metadados incluem informações de proveniência sobre a criação ou geração dos dados (FsF-R1.2-01M); 4) Os metadados seguem um padrão recomendado pela comunidade de pesquisa-alvo dos dados (FsF-R1.3-01M) e 5) Os

dados estão disponíveis em um formato de arquivo recomendado pela comunidade de pesquisa-alvo (FsF-R1.3-02D). O Quadro 21 apresenta os resultados encontrados.

Quadro 21 - Aderência quanto à Reutilização (USP)

Código	FsF-R1-01MD	FsF-R1.1-01M	FsF-R1.2-01M	FsF-R1.3-01M	FsF-R1.3-02D
344	inicial	incompleto	moderado	inicial	incompleto
343	inicial	incompleto	moderado	inicial	incompleto
342	inicial	incompleto	moderado	inicial	incompleto
341	inicial	incompleto	moderado	inicial	incompleto
340	inicial	incompleto	moderado	inicial	incompleto
339	inicial	incompleto	moderado	inicial	incompleto
338	inicial	incompleto	moderado	inicial	incompleto
337	inicial	incompleto	moderado	inicial	incompleto
336	inicial	incompleto	moderado	inicial	incompleto
335	inicial	incompleto	moderado	inicial	incompleto
334	inicial	incompleto	moderado	inicial	incompleto
333	inicial	incompleto	moderado	inicial	incompleto
332	inicial	incompleto	moderado	inicial	incompleto
331	inicial	incompleto	moderado	inicial	incompleto
330	inicial	incompleto	moderado	inicial	incompleto
329	inicial	incompleto	moderado	inicial	incompleto
328	inicial	incompleto	moderado	inicial	incompleto
327	inicial	incompleto	moderado	inicial	incompleto
326	inicial	incompleto	moderado	inicial	incompleto
325	inicial	incompleto	moderado	inicial	incompleto
324	inicial	incompleto	moderado	inicial	incompleto
323	inicial	incompleto	moderado	inicial	incompleto
322	inicial	incompleto	moderado	inicial	incompleto
321	inicial	incompleto	moderado	inicial	incompleto
320	inicial	incompleto	moderado	inicial	incompleto
319	inicial	incompleto	moderado	inicial	incompleto
318	inicial	incompleto	moderado	inicial	incompleto
317	inicial	incompleto	moderado	inicial	incompleto
316	inicial	incompleto	moderado	inicial	incompleto
315	inicial	incompleto	moderado	inicial	incompleto
314	inicial	incompleto	moderado	inicial	incompleto
313	inicial	incompleto	moderado	inicial	incompleto
312	inicial	incompleto	moderado	inicial	incompleto

311	inicial	incompleto	moderado	inicial	incompleto
310	inicial	incompleto	moderado	inicial	incompleto
309	inicial	incompleto	moderado	inicial	incompleto
308	inicial	incompleto	moderado	inicial	incompleto
307	incompleto	incompleto	moderado	inicial	incompleto
306	inicial	incompleto	moderado	inicial	incompleto
305	inicial	incompleto	moderado	inicial	incompleto
304	inicial	incompleto	moderado	inicial	incompleto
303	inicial	incompleto	moderado	inicial	incompleto
302	inicial	incompleto	moderado	inicial	incompleto
301	inicial	incompleto	moderado	inicial	incompleto
300	inicial	incompleto	moderado	inicial	incompleto
299	inicial	incompleto	moderado	inicial	incompleto
298	inicial	incompleto	moderado	inicial	incompleto
297	inicial	incompleto	moderado	inicial	incompleto
296	inicial	incompleto	moderado	inicial	incompleto
295	inicial	incompleto	moderado	inicial	incompleto
294	inicial	incompleto	moderado	inicial	incompleto
293	inicial	incompleto	moderado	inicial	incompleto
292	inicial	incompleto	moderado	inicial	incompleto
291	inicial	incompleto	moderado	inicial	incompleto
290	inicial	incompleto	moderado	inicial	incompleto
289	inicial	incompleto	moderado	inicial	incompleto
288	inicial	incompleto	moderado	inicial	incompleto
287	inicial	incompleto	moderado	inicial	incompleto
286	inicial	incompleto	moderado	inicial	incompleto
285	inicial	incompleto	moderado	inicial	incompleto
284	inicial	incompleto	moderado	inicial	incompleto
283	inicial	incompleto	moderado	inicial	incompleto
282	incompleto	incompleto	moderado	inicial	incompleto
281	inicial	incompleto	moderado	inicial	incompleto
280	inicial	incompleto	moderado	inicial	incompleto
279	inicial	incompleto	moderado	inicial	incompleto
278	inicial	incompleto	moderado	inicial	incompleto
277	inicial	incompleto	moderado	inicial	incompleto
276	inicial	incompleto	moderado	inicial	incompleto
275	inicial	incompleto	moderado	inicial	incompleto
274	inicial	incompleto	moderado	inicial	incompleto
273	inicial	incompleto	moderado	inicial	incompleto

272	inicial	incompleto	moderado	inicial	incompleto
271	inicial	incompleto	moderado	inicial	incompleto
270	inicial	incompleto	moderado	inicial	incompleto
269	inicial	incompleto	moderado	inicial	incompleto
268	inicial	incompleto	moderado	inicial	incompleto
267	inicial	incompleto	moderado	inicial	incompleto
266	inicial	incompleto	moderado	inicial	incompleto
265	inicial	incompleto	moderado	inicial	incompleto
264	inicial	incompleto	moderado	inicial	incompleto
263	inicial	incompleto	moderado	inicial	incompleto
262	inicial	incompleto	moderado	inicial	incompleto
261	inicial	incompleto	moderado	inicial	incompleto
260	inicial	incompleto	moderado	inicial	incompleto
259	inicial	incompleto	moderado	inicial	incompleto
258	inicial	incompleto	moderado	inicial	incompleto
257	inicial	incompleto	moderado	inicial	incompleto
256	inicial	incompleto	moderado	inicial	incompleto
255	inicial	incompleto	moderado	inicial	incompleto
254	inicial	incompleto	moderado	inicial	incompleto
253	inicial	incompleto	moderado	inicial	incompleto
252	inicial	incompleto	moderado	inicial	incompleto
251	incompleto	incompleto	moderado	inicial	incompleto
250	inicial	incompleto	moderado	inicial	incompleto
249	inicial	incompleto	moderado	inicial	incompleto
248	inicial	incompleto	moderado	inicial	incompleto
247	inicial	incompleto	moderado	inicial	incompleto
246	inicial	incompleto	moderado	inicial	incompleto
245	inicial	incompleto	moderado	inicial	incompleto
244	inicial	incompleto	moderado	inicial	incompleto
243	inicial	incompleto	moderado	inicial	incompleto
242	inicial	incompleto	moderado	inicial	incompleto
237	inicial	incompleto	moderado	inicial	incompleto
236	inicial	incompleto	moderado	inicial	incompleto
235	inicial	incompleto	moderado	inicial	incompleto
234	inicial	incompleto	moderado	inicial	incompleto
233	inicial	incompleto	moderado	inicial	incompleto
232	inicial	incompleto	moderado	inicial	incompleto
231	inicial	incompleto	moderado	inicial	incompleto
228	inicial	incompleto	moderado	inicial	incompleto

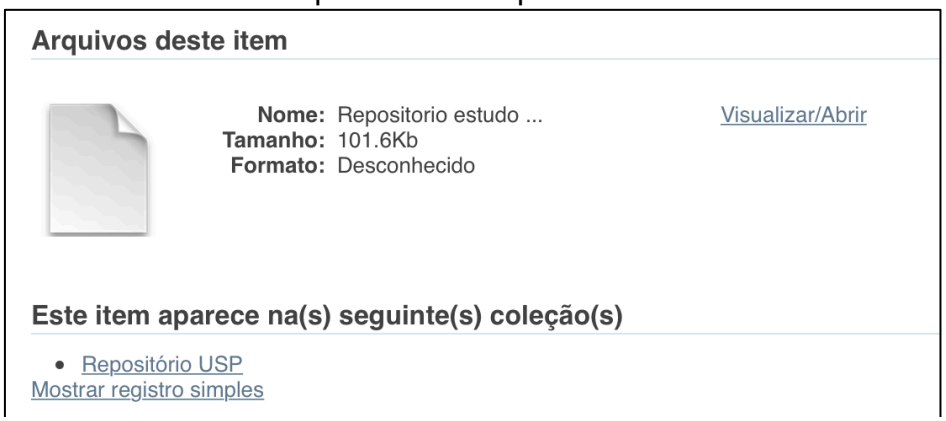
227	inicial	incompleto	moderado	inicial	incompleto
226	inicial	incompleto	moderado	inicial	incompleto
221	inicial	incompleto	moderado	inicial	incompleto
220	inicial	incompleto	moderado	inicial	incompleto
218	inicial	incompleto	moderado	inicial	incompleto
217	inicial	incompleto	moderado	inicial	incompleto
213	inicial	incompleto	moderado	inicial	incompleto

Fonte: elaborado pela autora (2022).

O primeiro aspecto de reutilização (FsF-R1-01MD) avaliou pontos como a existência de informações mínimas fornecidas nos metadados sobre o conteúdo dos dados disponíveis. A maioria dos *datasets* pontuaram como “inicial”, enquanto três pontuaram como “incompleto”. A ferramenta F-UJI, ao analisar os metadados, foi capaz de encontrar o tipo de recurso, retornando: “*dataset”*. Entretanto, não foi capaz de recuperar as outras informações como tamanho. É o mesmo caso das demais instituições da amostra.

Ao observar a Figura 27 é possível notar que algumas informações sobre o arquivo são disponibilizadas no registro: nome, tamanho e formato, mas não de uma forma que a máquina consegue recuperar automaticamente. O mesmo caso é visto no repositório da UNESP (Figura 13), onde todos os conjuntos de dados foram avaliados como incompletos no aspecto FsF-R1-01MD.

Figura 27 – Visualização do nome, tamanho e formato do arquivo de um conjunto de dados depositado no repositório da USP



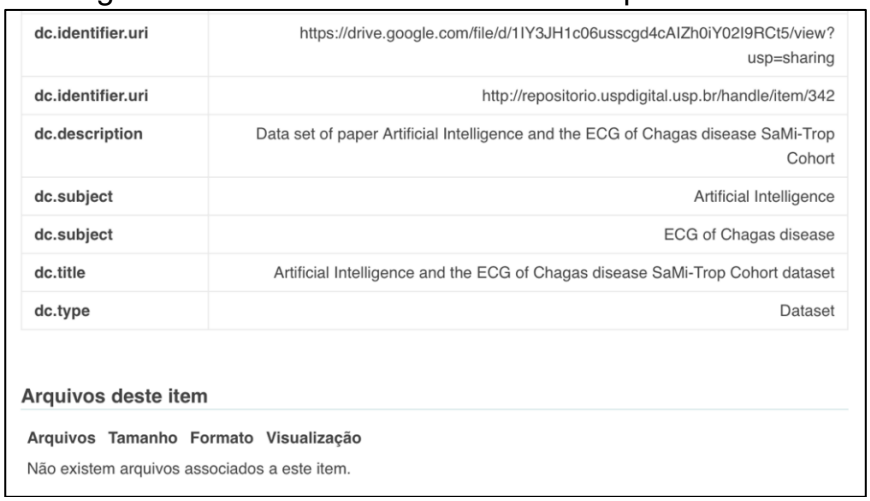
Fonte: *print screen* retirado do repositório da USP (2022).

Uma diferença entre essas duas instituições é que no caso da UNESP, ao clicar em cima do arquivo, o *download* começa imediatamente, sem a necessidade de

aceitar termos. Já na USP, como visto na Figura 25, é preciso aceitar os termos do formulário de consentimento. O tipo do recurso foi, então, a única informação recuperada sobre o conjunto de dados, e ela está, justamente, indicada no registro de metadados do *dataset* (*dc.type*).

Além disso, vale destacar que alguns registros não possuem nenhum arquivo associado, ou seja, há apenas a disponibilização dos metadados. Em um *dataset*, por exemplo, é fornecido (no elemento "*dc.identifier.uri*) um *link* do *Google Drive* para acessar os dados. O problema dessa forma de disponibilização é que a máquina tem dificuldade de processar e interpretar as informações sobre o arquivo. Na Figura 28 é possível observar um exemplo desse caso. Mas isso pode estar ligado ao tamanho do arquivo que o usuário desejava depositar, que talvez ultrapasse o limite definido pelo repositório. Seria preciso uma maior investigação para confirmar essa hipótese.

Figura 28 – Registro de um *dataset* da USP sem arquivos associados ao item



dc.identifier.uri	https://drive.google.com/file/d/1IY3JH1c06usscgd4cAIZh0iY02I9RCt5/view?usp=sharing
dc.identifier.uri	http://repositorio.uspdigital.usp.br/handle/item/342
dc.description	Data set of paper Artificial Intelligence and the ECG of Chagas disease SaMi-Trop Cohort
dc.subject	Artificial Intelligence
dc.subject	ECG of Chagas disease
dc.title	Artificial Intelligence and the ECG of Chagas disease SaMi-Trop Cohort dataset
dc.type	Dataset

Arquivos deste item

Arquivos Tamanho Formato Visualização

Não existem arquivos associados a este item.

Fonte: *print screen* retirado do repositório da USP (2022).

O segundo aspecto (FsF-R1.1-01M) diz respeito aos metadados incluírem ou não informações de licença sob as quais os dados podem ser reutilizados. É preciso informar, de maneira clara e explícita (para humanos e máquinas), quais direitos de uso são atribuídos aos dados. A ambiguidade pode limitar severamente a sua reutilização. Conforme visto no Quadro 21, todos os *datasets* da USP foram avaliados como "incompletos" neste quesito.

A ferramenta F-UJI alegou que não foi encontrada uma licença num elemento de metadados adequado e no registro completo/simples do item não é possível identificar essa declaração. Elementos como '*dc.rights'* ou *'dc.rights.uri'*, presentes,

por exemplo, no repositório da UFSCar, não foram identificados no repositório da USP, afetando seu potencial de reuso.

Entretanto, pode-se citar novamente o estudo de Dunning, Smaele e Böhmero (2017) na Holanda, em que a reutilização pareceu ser a faceta do FAIR mais difícil de ser aderir, com apenas 41% dos repositórios atribuindo uma licença clara aos dados. O estudo demonstra que há essa necessidade geral de melhorias para adesão aos princípios FAIR, e a atribuição de licenças é um dos pontos essenciais que podem mudar drasticamente a capacidade de reutilização. Tempo e suporte suficientes devem ser dados para permitir que os repositórios implementem as políticas necessárias.

O próximo aspecto (FsF-R1.2-01M) avaliou informações de proveniência nos metadados, e todos os 118 *datasets* obtiveram um nível moderado de descrição, seguindo o mesmo padrão dos demais repositórios da amostra analisados. Alguns dos elementos que foram extraídos pela ferramenta foram: "*creator", "publication_date"* e *"publisher"*. Essas informações de proveniência relacionadas à criação dos dados de pesquisa foram encontradas, mas a ferramenta não conseguiu validar a presença, nos metadados, de informações de proveniência usando ontologias formais, o que impossibilitou a pontuação dos *datasets* como avançado. Esse foi o mesmo caso dos demais repositórios de São Paulo, indicando outra tendência que pode ser estudada para melhorias posteriores.

Quanto ao quarto aspecto (FsF-R1.3-01M), a ferramenta verificou se um padrão de metadados específico da comunidade é detectado usando *namespaces* e se está listado no registro Re3data. Assim como nos demais casos, a ferramenta recuperou o DC, um padrão de metadados multidisciplinar, mas endossado pela comunidade (*RDA Metadata Standards Catalog*).

Como já visto, é mais fácil reutilizar conjuntos de dados se eles forem semelhantes: mesmo tipo de dados, dados organizados de maneira padronizada, formatos de arquivo bem estabelecidos e sustentáveis, documentação (metadados) seguindo um modelo comum e usando vocabulário comum. Ou seja, é um ponto positivo que os *datasets* da USP sejam representados com o uso do DC, um padrão que se adequa a vários domínios (como é o caso em repositórios multidisciplinares) e auxilia na interoperabilidade entre repositórios digitais. Mas é importante relembrar que o *DSpace* oferece o *Dublin Core* como esquema de metadados descritivos pré-definido, o que é o caso de alguns repositórios da amostra.

Para alcançar níveis mais avançados neste aspecto, seria preciso investir em dois outros pontos que a ferramenta F-UJI não foi capaz de encontrar: 1) o padrão de metadados da comunidade é detectado usando *namespaces* ou esquemas encontrados em metadados fornecidos ou saídas de serviços de metadados e 2) o padrão de metadados da comunidade está listado no registro Re3data do repositório responsável. Entretanto, o uso de um padrão endossado pela comunidade já é um bom indicativo, e a instituição pode dar prioridade para outros aspectos mais urgentes quando o assunto é conformidade com os princípios FAIR.

Por fim, quanto ao último aspecto (FsF-R1.3-02D), a ferramenta verificou se o formato do arquivo de dados é aberto/de longo prazo/científico. Seguindo a tendência dos outros repositórios da amostra, todos os conjuntos de dados foram avaliados como incompletos, ou seja, o formato do arquivo não era fornecido nos metadados/não estava contemplado na lista de formatos de arquivos de longo prazo, formatos de arquivos abertos ou formatos de arquivos científicos.

Analisando alguns arquivos associados aos itens do repositório encontrou-se o seguinte resultado: para a maioria dos *datasets*, o formato do arquivo é declarado apenas como “desconhecido”. Logo, para saber de que formato se trata, é preciso baixar o arquivo, aceitando os termos de uso e fornecendo seus dados como e-mail e filiação institucional. Algumas extensões encontradas dentro do repositório foram: ‘.pdf’, ‘.xlsx’, ‘.dat’ e ‘.txt’. Ou seja, assim como no caso da UFSCar, da Unicamp e da UNESP, há o depósito de arquivos em formatos proprietários. Isso pode acabar afetando tanto o acesso quanto a reutilização desses dados ao longo do tempo. Mas cabe à equipe gestora definir se deseja fazer o controle dos formatos passíveis de depósito no repositório.

Logo, é possível observar diferenças e semelhanças entre os quatro repositórios avaliados até o momento. Assim como no caso da UFSCar e da Unicamp, os *datasets* da USP receberam a nota 2 (sendo o máximo 10) em reutilizável. Fica nítido que esse é um ponto fraco comum entre as instituições da amostra, revelando que, de fato, o R de FAIR é de difícil aderência (DUNNING; SMAELE; BÖHMERO, 2017) e exige maiores investimentos. A USP obteve resultados iguais aos da Unicamp em acessibilidade, com todos os conjuntos de dados obtendo nota 1 de no máximo 3. Os *datasets* da UFSCar e a UNESP receberam nota 1,5.

A interoperabilidade também necessita de maiores investimentos dos repositórios. Tirando a Unicamp (onde a maioria dos conjuntos de dados receberam

nota 3 de 4), todas as demais instituições tiveram seus conjuntos de dados avaliados entre 0 e 1 de 4. Recomenda-se que a USP dê prioridade a este aspecto, investindo em vocabulários e recursos semânticos, além de referencias qualificadas entre os conjuntos de dados e outros recursos relacionados, como já foi apontado.

Outros pontos que podem ser trabalhados pela equipe da USP são os identificadores persistentes; metadados com o identificador dos dados que descrevem (ponto que nenhum *dataset* até o momento conseguiu aderir); declaração das restrições de acesso em metadados apropriados e declaração das licenças de uso em metadados apropriados. Isso porque, como já mencionado, é fundamental que todas as informações sejam legíveis por máquina para se falar em dados FAIR.

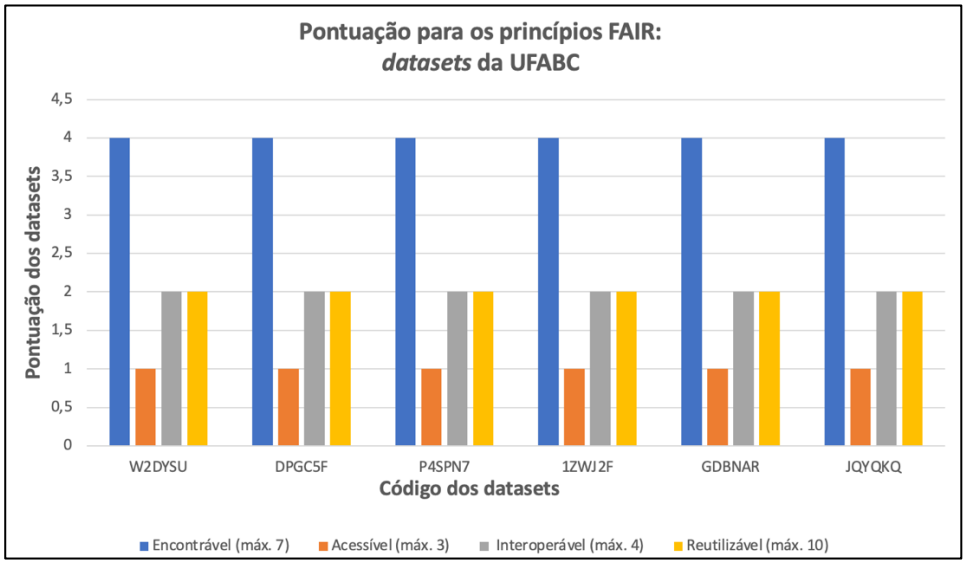
Esses são alguns aspectos a serem avaliados pela equipe, que pode definir a pertinência de implementação e investimento, lembrando que o FAIR existe dentro de um espectro. A questão dos metadados e dos identificadores persistentes estão dentro do núcleo básico proposto por Hodson *et al.* (2018) em sua escala 5 estrelas do FAIR. Ela pode ajudar a definir prioridades para investimentos.

5.1.5 Repositório institucional da Universidade Federal do ABC (UFABC)

O repositório da UFABC continha um total de seis conjuntos de dados (*datasets*) depositados no momento da coleta, e todos foram avaliados pela ferramenta tendo como base os princípios FAIR. Cada faceta possui uma nota máxima, conforme é visto na legenda do gráfico: encontrável (máximo sete), acessível (máximo três), interoperável (máximo quatro) e reutilizável (máximo dez). As notas, como citado anteriormente, foram estipuladas de forma automática pela ferramenta F-UJI com base nas suas métricas.

É possível observar pelo Gráfico 12 que nenhum dos *datasets* do repositório da UFABC conseguiu atingir nota máxima em encontrável, acessível, interoperável ou reutilizável. Mas isso foi uma tendência para todos os repositórios da amostra até o momento. A maior nota obtida em encontrável 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 de difícil aderência.

Gráfico 12 - Pontuação para os princípios FAIR: *datasets* da UFABC



Fonte: elaborado pela autora (2022).

Os seis *datasets* da UFABC foram também avaliados em uma escala de 0 a 100% quanto à aderência geral aos princípios FAIR, e todos obtiveram a mesma pontuação: 37%. Ou seja, há um alto grau de consistência entre os pontos fortes e fracos dos conjuntos de dados do repositório que, como visto no Gráfico 12, pontuaram exatamente igual. Isso pode estar ligado ao Grupo de Trabalho criado na instituição para elaborar diretrizes para gestão do repositório; identificar necessidades para seu funcionamento; promover a articulação com outras instituições para troca de experiências; e implantar e disponibilizar versões beta do sistema.

Indo para uma análise mais minuciosa, a ferramenta F-UJI ainda apresenta o nível (inicial, moderado, avançado ou incompleto) de conformidade quanto a cada aspecto dos quatro princípios FAIR. Dessa forma é possível verificar em quais implementações o repositório pode investir e o que pode ser feito, na prática, para melhorar essas pontuações. No Quadro 22 tem-se os resultados para a **encontrabilidade** dos dados de pesquisa da UFABC.

Quadro 22 - Aderência quanto à Encontrabilidade (UFABC)

Código	FsF-F1-01D	FsF-F1-02D	FsF-F2-01M	FsF-F3-01M	FsF-F4-01M
W2DYSU	avançado	incompleto	avançado	incompleto	avançado
DPGC5F	avançado	incompleto	avançado	incompleto	avançado
P4SPN7	avançado	incompleto	avançado	incompleto	avançado

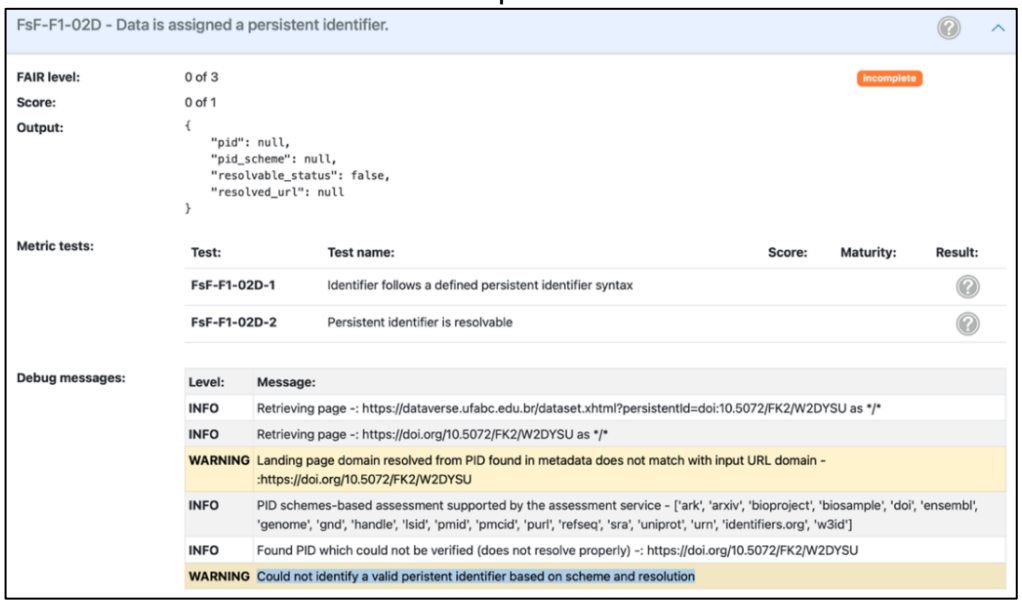
1ZWJ2F	avançado	incompleto	avançado	incompleto	avançado
GDBNAR	avançado	incompleto	avançado	incompleto	avançado
JQYQKQ	avançado	incompleto	avançado	incompleto	avançado

Fonte: elaborado pela autora (2022).

O primeiro aspecto (FsF-F1-01D), que diz respeito a: "são atribuídos identificadores globalmente exclusivos aos dados", foi avaliado como "avançado" para todos os conjuntos de dados da instituição, seguindo a mesma tendência dos demais repositórios da amostra. Não houve, portanto, lacunas na atribuição de identificadores globalmente exclusivos. A ferramenta auxiliar verificou inicialmente se o identificador atribuído segue uma sintaxe de identificador único definido (IRI, URL) e, de fato, todos os *datasets* cumprem o requisito.

Já o segundo aspecto (FsF-F1-02D) diz respeito a: "são atribuídos identificadores persistentes aos dados", e todos os conjuntos de dados foram avaliados como incompletos. É possível ver, na Figura 29, que a ferramenta até chegou a recuperar o DOI, mas alegou que o domínio da página de destino (*resolved from* PID) encontrado nos metadados não corresponde ao domínio do URL de entrada. O PID não pôde ser verificado e a ferramenta não conseguiu encontrar um identificador persistente válido.

Figura 29 – *Output* exibido pela ferramenta F-UJI ao avaliar um conjunto de dados da UFABC no aspecto FsF-F1-02D



Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

Inclusive, ao tentar acessar um conjunto de dados do repositório da UFABC através do DOI, o usuário é levado para uma página de erro com a seguinte mensagem: “A página que você está solicitando não pode ser encontrada. Verifique novamente a URL ou entre em contato conosco em support@datacite.org.”. Isso pode estar ligado a algum erro na sintaxe do DOI que é definida pela norma ANSI/NISO Z39.84-2000: *Syntax For The Digital Object Identifier.* Ela define a ordem e a composição do DOI utilizado para identificar objetos no ambiente digital. A instituição pode, então, verificar essa questão para garantir que os dados de pesquisa depositados sejam identificados de forma persistente.

Somado a isso, é interessante apontar que, inicialmente, as avaliações dos conjuntos de dados foram feitas tendo como base o DOI que é fornecido no repositório da UFABC. Ao inserir o DOI na ferramenta F-UJI o resultado encontrado foi bem abaixo dos 37%. Todos os conjuntos de dados pontuaram com apenas 8% de aderência aos princípios FAIR, indicando um problema para a extração correta das informações desse *link*. Optou-se, portanto, por utilizar a URL dos conjuntos de dados em vez do DOI na avaliação, levando à pontuação de 37% de aderência ao FAIR. Essa é outra informação que pode ser relevante para a equipe gestora do repositório da UFABC.

O terceiro aspecto avaliado (FsF-F2-01M) se refere a: "metadados incluem elementos essenciais (*'creator', 'title', 'publisher', 'publication_date', 'summary', 'keywords', 'object_identifier', 'object_type'*) para dar suporte à encontrabilidade dos objetos digitais". Como já visto, os metadados são citados no núcleo básico proposto por Hodson *et al.* (2018) em sua escala 5 estrelas do FAIR, ou seja, são essenciais e demandam atenção das instituições. Assim como no caso da Unicamp, todos os conjuntos de dados da UFABC foram avaliados como “avançado”, um excelente indicativo.

A ferramenta foi capaz de recuperar elementos como “*creator”, “keywords”, object_identifier*”, “*object_type*”, “*publication_date”, “publisher”, “summary”* e *“title”*. São os mesmos elementos encontrados para os conjuntos de dados da Unicamp, indicando um bom conjunto mínimo de descrição para os dados de pesquisa. Com eles é possível fazer a citação adequada do *dataset,* o que, como visto, é fundamental para a reutilização segura dos dados. Inclusive, assim como no caso do repositório da UFSCar, da UNESP e da Unicamp, o da UFABC também oferece a citação dos

conjuntos de dados já pronta, permitindo que o pesquisador apenas copie, dando os devidos créditos ao autor dos dados (Figura 30).

Figura 30 – Citação pronta disponibilizada no repositório da UFABC

Fonte: *print screen* retirado do site do repositório da UFABC (2022).

O próximo aspecto (FsF-F3-01M) diz respeito a: "os metadados incluem o identificador do dado que descrevem". A ferramenta avaliou duas características: 1) os metadados contêm informações relacionadas ao conteúdo dos dados (nome do arquivo, tamanho, tipo) e 2) os metadados contêm um PID ou URL que indica a localização do conteúdo dos dados para *download*. Seguindo a tendência de todos os demais repositórios, o identificador dos dados foi tido como ausente e, por isso, todos os *datasets* foram avaliados como incompletos. Esse é, portanto, um ponto fraco comum entre as instituições avaliadas até o momento. Recomenda-se novamente estudar o caso do *dataset* depositado no *Harvard Dataverse*, que conseguiu aderir em nível “avançado” a este aspecto, e foi citado em detalhes no tópico da UFSCar.

Por fim, no aspecto FsF-F4-01M, os *datasets* foram testados quanto aos metadados fornecidos de uma maneira que os principais mecanismos de pesquisa pudessem inseri-los em seus catálogos (exemplos: JSON-LD, *Dublin Core*, RDFa), e todos foram avaliados como “avançado”, mesmo resultado encontrado para as outras instituições da amostra. É um ponto forte comum, como já apontado.

Sendo assim, a equipe gestora do repositório da UFABC pode avaliar a necessidade de investimentos quanto ao aspecto FsF-F3-01M, que diz respeito aos metadados incluírem o identificador dos dados que descrevem (um ponto fraco de todos os repositórios). Mas recomenda-se que a equipe dê atenção prioritária à questão dos identificadores persistentes, uma vez que “[...] sem identificadores globais

únicos e persistentes será difícil conseguir outros elementos de dados FAIR" (HENNING *et al*., 2019b, p. 400).

Quadro 23 - Aderência quanto à Acessibilidade (UFABC)

Código	FsF-A1-01M	FsF-A1-03D	FsF-A1-02M
W2DYSU	incompleto	incompleto	avançado
DPGC5F	incompleto	incompleto	avançado
P4SPN7	incompleto	incompleto	avançado
1ZWJ2F	incompleto	incompleto	avançado
GDBNAR	incompleto	incompleto	avançado
JQYQKQ	incompleto	incompleto	avançado

Fonte: elaborado pela autora (2022).

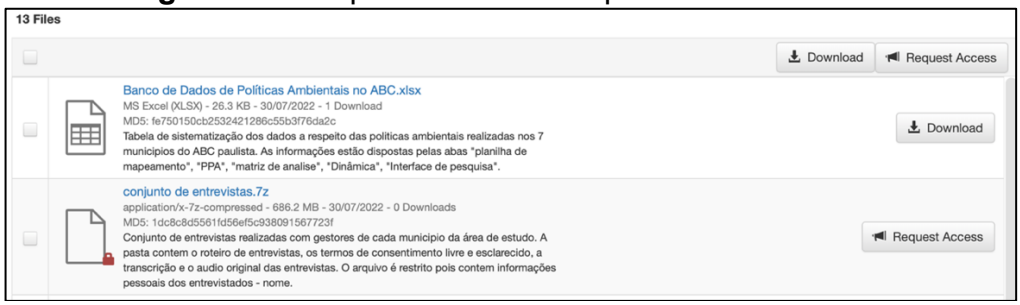
Já quanto à **acessibilidade**, os conjuntos de dados foram avaliados em três aspectos: 1) Os metadados contêm o nível de acesso e as condições de acesso aos dados (FsF-A1-01M); 2) Os dados são acessíveis por meio de um protocolo de comunicação padronizado (FsF-A1-03D) e 3) Os metadados são acessíveis por meio de um protocolo de comunicação padronizado (FsF-A1-02M). De acordo com Dias, Anjos e Rodrigues (2019, p. 183), ser acessível "[...] está relacionado ao fato de que os dados deveriam poder ser acessados por protocolos-padrões. Tecnológicas "esotéricas", fechadas, com poucas implementações e mal documentadas devem ser evitadas".

Nota-se, pelo Quadro 23, que todos os conjuntos de dados da UFABC só aderiram a um dos três aspectos avaliados pela ferramenta F-UJI, exatamente o mesmo caso da Unicamp e da USP. A UFSCar aderiu a dois (FsF-A1-01M em nível inicial e FsF-A1-02M em nível avançado) e a UNESP também aderiu a dois (FsF-A1-01M em nível inicial e FsF-A1-02M em nível avançado). Ou seja, a acessibilidade, comparando as cinco instituições, representa um ponto fraco a ser trabalhado pela UFABC, assim como pela Unicamp e pela USP.

Para o aspecto FsF-A1-01M, a ferramenta auxiliar verificou se os *datasets* possuíam informações sobre restrições ou direitos de acesso nos metadados, mas não foi capaz de encontrar essas informações. Esse resultado, inclusive, também foi encontrado no repositório da Unicamp e numa avaliação adicional feita com *datasets* do *Harvard Dataverse* e do *SciELO Data*, que utilizam o *software Dataverse*, assim

como a UFABC. Após testes rápidos com alguns conjuntos de dados desses dois repositórios adicionais, foi possível notar uma tendência de não aderência ao aspecto FsF-A1-01M, mesmo tendo alcançado níveis de *FAIRness* altos.

Figura 31 – Arquivo restrito no repositório da UFABC



Fonte: *print screen* retirado do site do repositório da UFABC (2022).

Na Figura 31 é possível observar o caso de um *dataset* que possui arquivos restritos. Assim como no repositório da UNESP e da Unicamp, os metadados continuam disponíveis, mesmo que os arquivos não estejam, uma boa prática importante para o FAIR. E, assim como no caso da Unicamp, no repositório da UFABC o usuário tem a possibilidade de contatar o criador dos dados (por meio do botão '*contact*' na parte superior da página) ou solicitar o acesso ao arquivo restrito, conforme mostra a Figura 31. A interface é igual a da Unicamp, tornando intuitiva a navegação em ambos os sistemas que utilizam o *Dataverse*.

Já em relação ao segundo aspecto (FsF-A1-03D), que diz respeito aos dados serem acessíveis por meio de um protocolo de comunicação padronizado, a ferramenta examinou se os metadados incluíam um *link* para dados com base em protocolos de comunicação da *web* padronizados (*standard_data_protocol*), o que não foi encontrado, mesmo resultado dos demais repositórios avaliados. Ou seja, nenhum deles conseguiu aderir ao FsF-A1-03D.

Entretanto, todos os *datasets* da UFABC são identificados por *links* que começam com https://, ou seja, são recuperados utilizando um protocolo de comunicação padrão, o HTTPS. Essa questão também foi, inclusive, testada para os metadados (FsF-A1-02M), e a ferramenta F-UJI conseguiu localizar o protocolo padrão para acesso aos metadado: HTTPS. Logo, quanto aos aspectos FsF-A1-03D e FsF-A1-02M, todos os repositórios avaliados e seus *datasets* pontuaram igualmente, indicando uma consistência (e tendência) nos resultados encontrados em "acessível".

É interessante, portanto, um trabalho conjunto para aplicar melhorias nos repositórios da amostra.

A ferramenta F-UJI também avaliou os *datasets* quanto à **interoperabilidade** e, apesar de ser o elemento mais desafiador, o repositório obteve bons resultados se comparado com as demais instituições: todos os conjuntos de dados da UFABC tiraram a nota 2 de 4. Foram avaliados três aspectos: 1) Os metadados são representados usando uma linguagem de representação de conhecimento formal (FsF-I1-01M); 2) Metadados usam recursos semânticos (FsF-I1-02M) e 3) Os metadados incluem *links* entre os dados e suas entidades relacionadas (FsF-I3-01M). Os resultados podem ser vistos no Quadro 24.

Quadro 24 - Aderência quanto à Interoperabilidade (UFABC)

Código	FsF-I1-01M	FsF-I1-02M	FsF-I3-01M
W2DYSU	moderado	inicial	avançado
DPGC5F	moderado	inicial	avançado
P4SPN7	moderado	inicial	avançado
1ZWJ2F	moderado	inicial	avançado
GDBNAR	moderado	inicial	avançado
JQYQKQ	moderado	inicial	avançado

Fonte: elaborado pela autora (2022).

Quanto ao primeiro aspecto, que diz respeito a uma linguagem de representação de conhecimento formal, todos os conjuntos de dados pontuaram em nível moderado, diferente dos *datasets* da UFSCar, da USP e da UNESP que foram avaliados como incompletos. Os conjuntos de dados foram testados quanto à existência de metadados estruturados e analisáveis (JSON-LD, RDFa) incorporados ao código XHTML/HTML da página de destino, o que foi encontrado. De acordo com a ferramenta F-UJI foi possível encontrar dados estruturados (RDF *serialization*) na página, mas não foi encontrado nenhum *endpoint* SPARQL por meio do Re3data com base no URI fornecido. Além disso, não foram encontrados metadados RDF disponíveis por meio de negociação de conteúdo.

A Unicamp foi a outra instituição que conseguiu aderir ao aspecto FsF-I1-01M. Em comum, ambas utilizam o *Dataverse*, que traz várias vantagens para níveis maiores de conformidade com os princípios FAIR, como a atribuição de identificadores

persistentes, citações geradas automaticamente, controle de versões (versionamento) e *harvesting* (OAI-PMH). O usuário pode, inclusive, exportar os metadados em DC ou JSON. Isso contribui para níveis maiores de interoperabilidade.

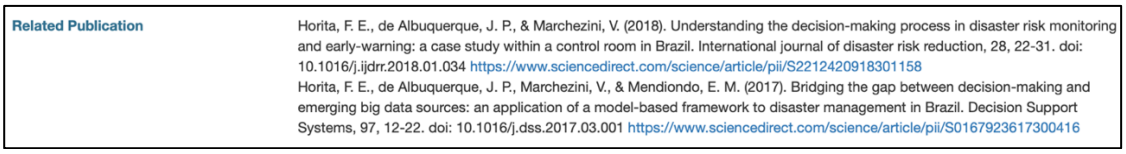
Já no aspecto FsF-I1-02M a ferramenta buscou por uso de *namespaces* de vocabulário e de recursos semânticos nos metadados, e todos os *datasets* pontuaram como inicial. Como já visto, o teste buscou extrair *namespaces* dos metadados baseados em RDF*,* que garantem que determinado conjunto de objetos tenha nomes exclusivos para que possa ser facilmente identificado. O mesmo resultado foi encontrado para os conjuntos de dados da UFSCar, da Unicamp (com algumas exceções), da USP e da UNESP, indicando um ponto que pode ser trabalhado em comum para maiores níveis de *FAIRness*.

Um vocabulário que é testado pela ferramenta F-UJI é o *Schema.org.* É uma iniciativa colaborativa, fundada pelo Google, Microsoft, Yahoo e Yandex, com a missão de criar, manter e promover esquemas para dados estruturados na Internet. O vocabulário *Schema.org* pode ser usado com muitas codificações diferentes, incluindo RDFa e JSON-LD. Esses vocabulários abrangem entidades, relacionamentos entre entidades e ações (OUCHI; SIMIONATO, 2018). Iniciativas como essa permitem aumentar a interoperabilidade dos dados de pesquisa, e podem ser estudadas para implementação futura.

Quanto ao último aspecto da interoperabilidade, que diz respeito às ligações entre os dados e suas entidades relacionadas, todos os conjuntos de dados pontuaram como avançados, um excelente resultado. Era preciso indicar de forma explícita, nos metadados do item, os recursos relacionados (e preferencialmente por meio de *links* ou identificadores legíveis por máquina).

Uma referência qualificada é uma referência cruzada que explica sua intenção. O objetivo é criar tantos *links* significativos quanto possível entre recursos de (meta)dados para enriquecer o conhecimento contextual sobre os dados. Assim como no caso da Unicamp e da UFSCar, a ferramenta recuperou ligações entre recursos. Todos os *datasets* da UFABC apresentavam a ligação: '*isPartOf*'. Na Figura 32 é possível observar, nos metadados de um *dataset*, as publicações relacionadas, criando essa conexão entre recursos e aumentando a interoperabilidade do item.

Figura 32 – Publicações relacionadas indicadas em um registro de um *dataset* da UFABC



Fonte: *print screen* retirado do site do repositório da UFABC (2022).

Logo, os conjuntos de dados depositados no repositório da UFABC apresentam um bom nível de interoperabilidade, com todos obtendo a pontuação dois (nota máxima sendo quatro). O único repositório que obteve pontuações maiores (maioria dos *datasets* obteve nota 3) em interoperabilidade foi o da Unicamp, que também utiliza o *Dataverse*. Para aumentar a interoperabilidade o repositório pode investir no uso de linguagem formal de representação do conhecimento e no uso de recursos semânticos nos metadados (vocabulários/*namespaces*) em maiores níveis de conformidade com o FAIR.

Por fim, todos os *datasets* foram avaliados quanto à **reutilização**. O objetivo final do FAIR é otimizar a reutilização dos dados. Como já visto, a reutilização vêm sendo o ponto fraco dos repositórios da amostra, com a maioria dos conjuntos de dados obtendo nota 2 de 10, mesmo caso da UFABC. O único repositório que, até o momento, obteve uma pontuação mais alta que 2 foi o da UFSCar, mas somente dois de 15 *datasets* tiraram a nota 4, sendo uma exceção no padrão da instituição.

Os dados da UFABC foram testados em cinco aspectos pela ferramenta auxiliar: 1) Metadados especificam o conteúdo dos dados (FsF-R1-01MD); 2) Os metadados incluem informações de licença sob as quais os dados podem ser reutilizados (FsF-R1.1-01M); 3) Os metadados incluem informações de proveniência sobre a criação ou geração dos dados (FsF-R1.2-01M); 4) Os metadados seguem um padrão recomendado pela comunidade de pesquisa-alvo dos dados (FsF-R1.3-01M) e 5) Os dados estão disponíveis em um formato de arquivo recomendado pela comunidade de pesquisa-alvo (FsF-R1.3-02D). O Quadro 25 apresenta os resultados encontrados.

Quadro 25 - Aderência quanto à Reutilização (UFABC)

Código	FsF-R1-01MD	FsF-R1.1-01M	FsF-R1.2-01M	FsF-R1.3-01M	FsF-R1.3-02D
W2DYSU	inicial	incompleto	moderado	inicial	incompleto
DPGC5F	inicial	incompleto	moderado	inicial	incompleto

P4SPN7	inicial	incompleto	moderado	inicial	incompleto
1ZWJ2F	inicial	incompleto	moderado	inicial	incompleto
GDBNAR	inicial	incompleto	moderado	inicial	incompleto
JQYQKQ	inicial	incompleto	moderado	inicial	incompleto

Fonte: elaborado pela autora (2022).

O primeiro aspecto (FsF-R1-01MD) avaliou pontos como a existência de informações mínimas fornecidas nos metadados sobre o conteúdo dos dados disponíveis, o que foi localizado em nível inicial com o "*object_type*", que recebe como valor "*dataset*". Assim como no caso da Unicamp, o repositório da UFABC possui apenas conjuntos de dados (ou seja, não há outros tipos de recursos como teses e dissertações).

Logo, é facilmente dedutível que o tipo de recurso é justamente um dado de pesquisa. Isso abre margem para descrições mais ricas por parte da instituição, contextualizando o tipo de dado depositado. Entretanto, o uso de "*dataset*" no campo "*type*", como já mencionado, pode auxiliar na troca de informações entre sistemas diferentes, como os outros repositórios da amostra. É, portanto, o mesmo caso da Unicamp.

Assim como no caso das demais instituições da amostra, a ferramenta não conseguiu recuperar outras informações do recurso como o tamanho do arquivo. Mas, ao entrar nos arquivos dos *datasets*, é possível encontrar informações como tamanho, tipo, descrição, data de depósito, data de publicação, *etc*. É a mesma interface encontrada na Figura 24. Ou seja, as informações estão declaradas, mas a ferramenta F-UJI não foi capaz de extrair.

Novamente, isso pode estar ligado ao fato delas não estarem disponibilizadas nos metadados (do item) que podem ser extraídos em *Schema.org* JSON-LD, diferente do metadado "*type*", que estava presente. Já foi citado no caso da Unicamp o conjunto de dados do *Harvard Dataverse* (https://doi.org/10.7910/DVN/WR4S9I) que pontuou como avançado nesse aspecto. Nos metadados exportados do *dataset* em sim (não do arquivo) é possível encontrar campos como "*name*", "*type*", "*format*" e "*size*" dos arquivos do conjunto de dados. Ou seja, as informações dos arquivos constam no registro de metadados do recurso.

A equipe gestora do repositório da UFABC pode avaliar essa aplicação de acordo com as necessidades da comunidade. Isso porque, apesar de um usuário humano conseguir encontrar facilmente essas informações, um computador pode ter

dificuldades de processar automaticamente, o que é preconizado pelo FAIR. Ou seja, não basta fornecer as informações sobre o *dataset*, elas precisam ser legíveis por máquina para serem FAIR. Mas a equipe pode optar por dar prioridade a outras questões mais urgentes.

O segundo aspecto (FsF-R1.1-01M) diz respeito aos metadados incluírem ou não informações de licença sob as quais os dados podem ser reutilizados, e todos os conjuntos de dados foram avaliados como incompletos. Essa é uma informação extremamente importante para que pesquisadores possam reutilizar os dados de pesquisa. A ferramenta F-UJI alegou encontrar (*schema.org*) a licença CC0, mas a representação da licença estava incorreta, invalidando o teste. Isso pode dificultar que máquinas interpretem a informação.

Apesar disso, é um bom indicativo que todos os *datasets* possuam uma licença CC associada na página do item. É importante relembrar, porém, que todos os conjuntos de dados adicionados em um repositório *Dataverse* recebem a CC0 *Public Domain Dedication* como padrão. Logo, as questões aqui apontadas dizem respeito ao processamento automático pelas máquinas, uma vez que usuários humanos conseguem encontrar e interpretar a licença facilmente.

O próximo aspecto (FsF-R1.2-01M) avaliou informações de proveniência nos metadados, e todos os *datasets* obtiveram um nível moderado de descrição. Alguns dos elementos que foram extraídos pela ferramenta foram: "*creator", "publication_date", "publisher"* e *"modified_date*". Esse último elemento não foi encontrado em outros repositórios (além da Unicamp, que também utiliza o *Dataverse*) e está ligado ao versionamento que é possibilitado pelo *software*. O versionamento é essencial para que os usuários entendam se estão acessando/reutilizando as versões mais recentes de um conjunto de dados, e quantas modificações já houve. Oferecer esse *background* na interface do repositório é uma boa prática recomendada, inclusive, pela W3C.

Como visto, as informações relacionadas à criação dos dados de pesquisa foram encontradas, mas a ferramenta não conseguiu validar a presença, nos metadados, de informações de proveniência usando ontologias formais, o que impossibilitou a pontuação dos *datasets* como avançado. Esse foi o mesmo caso dos demais repositórios, indicando uma tendência para futuras melhorias conjuntas.

Quanto ao quarto aspecto (FsF-R1.3-01M), a ferramenta verificou se um padrão de metadados específico da comunidade é detectado usando *namespaces* e

se está listado no registro Re3data. O nível “inicial” obtido por todos os *datasets* diz respeito ao uso de um padrão de metadados multidisciplinar, endossado pela comunidade (*RDA Metadata Standards Catalog*), que é detectado através de *namespaces.*

A ferramenta F-UJI detectou o padrão *Dublin Core,* mas indicou que os padrões de metadados não estão listados no registro *Re3data.org* do repositório responsável. A equipe gestora pode avaliar a pertinência dessa melhoria, mas o uso de um padrão validado pela comunidade como o DC, por exemplo, já é um ótimo indicativo, ainda mais quando se pensa na interoperabilidade entre sistemas. Todos os repositórios avaliados até então utilizam justamente o DC, e essa adoção padronizada é o maior foco para este estudo.

Por fim, quanto ao último aspecto (FsF-R1.3-02D), a ferramenta verificou se o formato do arquivo de dados é aberto/de longo prazo/científico. Seguindo a tendência dos outros repositórios da amostra, todos os conjuntos de dados foram avaliados como incompletos, ou seja, o formato do arquivo não era fornecido nos metadados/não estava contemplado na lista de formatos de arquivos de longo prazo, formatos de arquivos abertos ou formatos de arquivos científicos.

Entretanto, por conter apenas seis conjuntos de dados, foi feita a análise manual dos formatos dos arquivos, para entender quais eram usados. No elemento “*type*” do arquivo foi possível identificar os seguintes formatos: Adobe PDF, MS *Excel* (XLSX), *application/x-7z-compressed, Tabular Data*, *Plain Text, Comma Separated Values.* São valores similares aos encontrados no caso da Unicamp, indicando uma consistência entre os dois repositórios que utilizam *Dataverse*.

Assim como no caso da UFSCar, da Unicamp, da USP e da UNESP, há o depósito de arquivos em formatos proprietários, ou seja, elaborados em *softwares* pagos e, portanto, não estão disponíveis sem barreiras de acesso e custo como é defendido pelo movimento da Ciência Aberta. Isso pode acabar afetando tanto o acesso quanto a reutilização desses dados ao longo do tempo. Cabe à equipe gestora definir se deseja fazer um controle dos formatos depositados, incentivando o depósito de formatos não proprietários ou criando políticas mandatórias nesse sentido.

Logo, é possível observar que todos os repositórios da amostra obtiveram resultados semelhantes em reutilização. Tirando a UNESP, em que todos os *datasets* pontuaram como incompleto no aspecto FsF-R1-01MD, os demais repositórios da

amostra receberam as mesmas avaliações da ferramenta F-UJI, indicando um ponto fraco comum que as instituições podem trabalhar em conjunto.

Outro ponto interessante para analisar são as semelhanças entre os resultados da UFABC e da Unicamp. Além de seus conjuntos de dados terem recebido as maiores pontuações gerais de toda a amostra (37% e 50%, respectivamente), elas ainda apresentaram pontos fortes e fracos em comum. Comparando com os demais repositórios estudados, as duas obtiveram resultados mais baixos em acessibilidade e resultados mais altos em encontrabilidade e interoperabilidade. Silva e Rodrigues (2021, p. 127) defendem que "[...] uma plataforma de repositório como o *Dataverse* pode facilitar muito a criação de dados científicos FAIR", e, de fato, é o que os resultados deste trabalho mostram até o momento.

Os pontos que podem ser trabalhados pela equipe da UFABC são os seguintes: verificar a sintaxe dos DOIs atribuídos, uma vez que a ferramenta não conseguiu validá-los; metadados devem incluir o identificador dos dados que descrevem (nome, tamanho e tipo de arquivo, em formato legível por máquina); o nível e as restrições de acesso devem ser explicitamente declarados nos metadados; é preciso investir em metadados que utilizam recursos semânticos; e os metadados devem incluir explicitamente a licença de uso para os dados de pesquisa (em formato legível por máquina).

Esses são alguns aspectos a serem avaliados pela equipe, que pode definir a pertinência de implementação e investimento, lembrando que o FAIR existe dentro de um espectro. O núcleo básico proposto por Hodson *et al.* (2018), por exemplo, diz respeito a metadados de descoberta, identificadores persistentes e acesso aos dados ou, no mínimo, metadados. Escalas como essa ajudam na priorização de investimentos.

5.1.6 Repositório institucional da Universidade Federal de São Paulo (Unifesp)

A última análise diz respeito ao repositório da Unifesp, que continha um total de 15 conjuntos de dados (*datasets*) depositados no momento da coleta. Todos foram avaliados pela ferramenta tendo como base os princípios FAIR. As notas máximas podem ser vistas na legenda do gráfico: encontrável (máximo sete), acessível (máximo três), interoperável (máximo quatro) e reutilizável (máximo dez). As notas,

como citado anteriormente, foram estipuladas de forma automática pela ferramenta F-UJI com base nas suas métricas.

O repositório da Unifesp apresentava um problema semelhante ao da UFABC: ao tentar acessar o conjunto de dados pelo DOI disponibilizado no metabuscador da FAPESP, o usuário é levado para uma página de erro com a seguinte mensagem: "404 *Not Found! The page you are requesting cannot be found. Please check again the URL or contact us at support@datacite.org."*. Ao usar o DOI como identificador dos *datasets* na ferramenta F-UJI, o resultado foi o seguinte: todos receberam a pontuação geral de 8% de aderência, exatamente igual ao repositório da UFABC. Por isso, optou-se por utilizar a URL do conjunto de dados, o que acabou elevando as pontuações atribuídas pela ferramenta.

É possível observar pelo Gráfico 13 que nenhum dos *datasets* do repositório da Unifesp conseguiu atingir nota máxima, o que se aplica à encontrabilidade, acessibilidade, interoperabilidade e reutilização. Mas, como já visto, foi exatamente o mesmo caso de todos os outros repositórios da amostra. Logo, todos eles podem optar por investir em melhorias para aderir aos princípios FAIR, lembrando que ele se trata de um espectro, e dificilmente um conjunto de dados vai obter 100% de aderência.

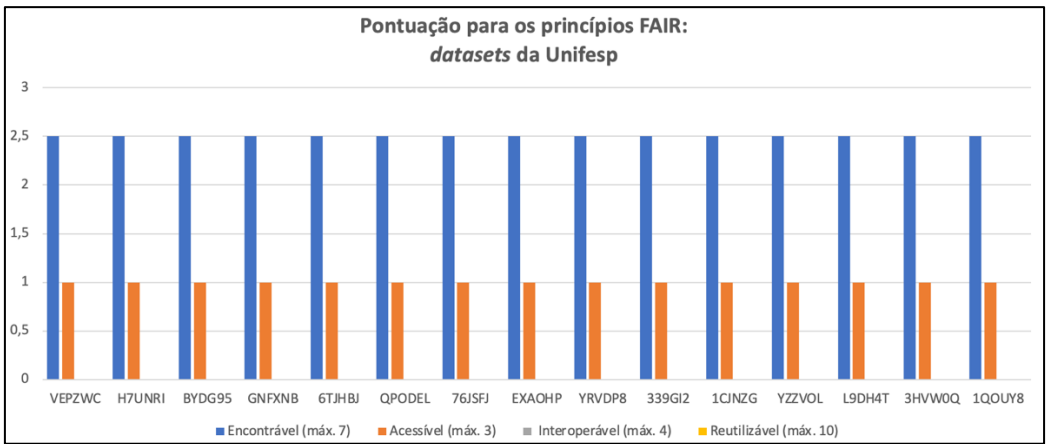
O objetivo é, então, buscar sempre aumentar o nível de *FAIRness*, investindo naqueles aspectos que estão de acordo com as necessidades da instituição mantenedora do repositório. A maior nota obtida em encontrável, no caso da Unifesp, foi 2,5 de 7, em acessível foi 1 de 3, em interoperável foi 0 de 4 e em reutilizável foi 0 de 10. Novamente a reutilização se mostra de difícil aderência, junto com a interoperabilidade. Fica nítido que eles representam desafios para os repositórios de São Paulo e seus dados de pesquisa.

Os 15 *datasets* da Unifesp foram também avaliados em uma escala de 0 a 100% quanto à aderência geral aos princípios FAIR, e todos obtiveram a mesma pontuação: 14%. 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 que, como visto no Gráfico 13, pontuaram exatamente igual.

Isso pode estar ligado ao autoarquivamento guiado por instruções da instituição. Na página de perguntas frequentes (FAQ) do repositório existe um passo a passo de como depositar os dados e os campos que são obrigatórios para preenchimento. Além disso, após o autoarquivamento, é preciso enviar um e-mail para 'repositoriodedados@unifesp.br' para que a equipe possa verificar as informações e

disponibilizar o conjunto de dados para a comunidade. Essa verificação pode garantir maior consistência na representação dos recursos.

Gráfico 13 - Pontuação para os princípios FAIR: *datasets* da Unifesp



Fonte: elaborado pela autora (2022).

Indo para uma análise mais minuciosa, a ferramenta F-UJI ainda apresenta o nível (inicial, moderado, avançado ou incompleto) de conformidade quanto a cada aspecto do FAIR. Dessa forma é possível verificar em quais implementações o repositório pode investir e o que pode ser feito, na prática, para melhorar essas pontuações. No Quadro 26 tem-se os resultados para a **encontrabilidade** dos dados de pesquisa da Unifesp.

Quadro 26 - Aderência quanto à Encontrabilidade (Unifesp)

Código	FsF-F1-01D	FsF-F1-02D	FsF-F2-01M	FsF-F3-01M	FsF-F4-01M
VEPZWC	avançado	incompleto	inicial	incompleto	avançado
H7UNRI	avançado	incompleto	inicial	incompleto	avançado
BYDG95	avançado	incompleto	inicial	incompleto	avançado
GNFXNB	avançado	incompleto	inicial	incompleto	avançado
6TJHBJ	avançado	incompleto	inicial	incompleto	avançado
QPODEL	avançado	incompleto	inicial	incompleto	avançado
76JSFJ	avançado	incompleto	inicial	incompleto	avançado
EXAOHP	avançado	incompleto	inicial	incompleto	avançado
YRVDP8	avançado	incompleto	inicial	incompleto	avançado
339GI2	avançado	incompleto	inicial	incompleto	avançado
1CJNZG	avançado	incompleto	inicial	incompleto	avançado
YZZVOL	avançado	incompleto	inicial	incompleto	avançado
L9DH4T	avançado	incompleto	inicial	incompleto	avançado

3HVW0Q	avançado	incompleto	inicial	incompleto	avançado
1QOUY8	avançado	incompleto	inicial	incompleto	avançado

Fonte: elaborado pela autora (2022).

O primeiro aspecto (FsF-F1-01D), que diz respeito a: "são atribuídos identificadores globalmente exclusivos aos dados", foi avaliado como "avançado" para todos os conjuntos de dados da instituição, seguindo a mesma tendência dos demais repositórios da amostra. Ou seja, todos os repositórios de São Paulo da amostra atribuíram identificadores globalmente exclusivos para seus conjuntos de dados. A ferramenta auxiliar verificou, como nos demais casos, se o identificador atribuído segue uma sintaxe de identificador único definido (IRI, URL) e, de fato, todos os *datasets* cumprem o requisito.

Já o segundo aspecto (FsF-F1-02D) diz respeito a: "são atribuídos identificadores persistentes aos dados", e todos os conjuntos de dados foram avaliados como incompletos. Diferente do caso da UFABC, a ferramenta F-UJI não conseguiu recuperar o DOI atribuído aos conjuntos de dados. Esse foi um ponto em que os resultados entre os repositórios da amostra divergiram um pouco.

No caso da UFSCar, apenas dois *datasets* (de 15) pontuaram como avançado, um pontuou como inicial e todos os outros como incompletos. No caso da Unicamp, que obteve os melhores resultados dentre os repositórios avaliados, todos os 68 conjuntos de dados possuíam um identificador persistente (DOI) que foi validado pela ferramenta F-UJI. Os conjuntos de dados da UNESP também pontuaram como avançado neste aspecto, possuindo identificadores persistentes (*handle*). No caso da USP, todos os *datasets* foram avaliados como incompletos. O mesmo resultado foi encontrado para a UFABC. Ou seja, apenas dois repositórios tiveram todos os seus dados de pesquisa validados no aspecto FsF-F1-02D, indicando um ponto fraco que pode ser visto como prioridade pelas instituições, uma vez que se encontra no núcleo básico proposto por Hodson *et al.* (2018) em sua escala 5 estrelas do FAIR.

Vale refrisar que, ao tentar acessar um conjunto de dados do repositório da Unifesp através do DOI disponibilizado, o resultado foi uma página “erro 404 *not found*”. É o mesmo caso da UFABC. O erro 404 é um código de resposta HTTP que indica que o cliente pôde comunicar-se com o servidor, mas o servidor não pôde encontrar o que foi pedido, ou foi configurado para não cumprir o pedido e não revelar a razão, a página não existe mais ou a URL foi inserida incorretamente. É importante,

portanto, que a equipe gestora do repositório verifique esta questão para solucionar o problema, garantindo que os usuários consigam acessar o conteúdo dos conjuntos de dados pelo DOI fornecido.

O terceiro aspecto avaliado (FsF-F2-01M) se refere a: "metadados incluem elementos essenciais (*'creator', 'title', 'publisher', 'publication_date', 'summary', 'keywords', 'object_identifier', 'object_type'*) para dar suporte à encontrabilidade dos objetos digitais". Como já visto, os metadados são citados no núcleo básico proposto por Hodson *et al.* (2018) em sua escala 5 estrelas do FAIR, ou seja, são essenciais e demandam atenção das instituições. Todos os conjuntos de dados da Unifesp foram avaliados em nível inicial, e a ferramenta alegou que não estão presentes os metadados essenciais para a citação ou os metadados descritivos essenciais do *dataset* (Figura 33).

Figura 33 – *Output* exibido pela ferramenta F-UJI ao avaliar um conjunto de dados da Unifesp no aspecto FsF-F2-01M

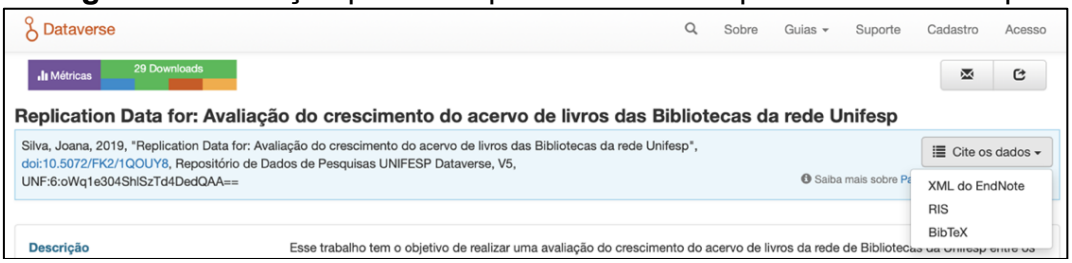
Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

A ferramenta só recuperou um elemento: "*summary*". Entretanto, ao avaliar o registro de metadados, é possível encontrar elementos como *"title"*, "*author*", *"keywords"*, "*publication_date" e "publisher".* Alguns outros metadados são fornecidos como "*language", "subject", "description*", *etc*. Não se sabe o porquê de a ferramenta não ter recuperado os metadados, uma vez que foi capaz de recuperar para outros repositórios da amostra. Mesmo no caso da UNESP, por exemplo, que também teve seus conjuntos de dados avaliados como nível inicial neste aspecto, foi possível

recuperar outros elementos além de “*summary*”. A equipe técnica do repositório pode investigar essa questão caso seja pertinente para a instituição. O próprio caso do repositório da Unicamp, que também utiliza o *software Dataverse,* pode ser usado como exemplo. Todos os conjuntos de dados foram avaliados como avançados e os elementos de metadados essenciais foram recuperados.

Um fator positivo é que, assim como no caso dos repositórios da UFSCar, da UNESP, da Unicamp, e da UFABC, o da Unifesp também oferece a citação dos conjuntos de dados já pronta (uma funcionalidade do *Dataverse*), permitindo que o pesquisador apenas copie, dando os devidos créditos ao autor dos dados (Figura 34). Isso pode aumentar a reutilização dos dados, facilitando o trabalho do usuário que deseja aproveitar aqueles dados já coletados para a sua pesquisa.

Figura 34 – Citação pronta disponibilizada no repositório da Unifesp



Fonte: *print screen* retirado do site do repositório da Unifesp (2022).

O próximo aspecto (FsF-F3-01M) diz respeito a: "os metadados incluem o identificador do dado que descrevem". A ferramenta avaliou duas características: 1) os metadados contêm informações relacionadas ao conteúdo dos dados (nome do arquivo, tamanho, tipo) e 2) os metadados contêm um PID ou URL que indica a localização do conteúdo dos dados para *download*. Assim como no caso dos outros cinco repositórios da amostra, o identificador dos dados foi tido como ausente e, por isso, todos os *datasets* foram avaliados como "incompleto". Ou seja, esse é um aspecto que nenhum repositório conseguiu aderir e pode ser trabalhado em conjunto pelas equipes gestoras. Como já citado, o *dataset* depositado no *Harvard Dataverse*, que conseguiu aderir em nível “avançado” a este aspecto, pode ser usado como base de estudo para futuras melhorias. Ele foi citado em detalhes na seção da UFSCar.

Por fim, no aspecto FsF-F4-01M, os *datasets* foram testados quanto aos metadados fornecidos de uma maneira que os principais mecanismos de pesquisa pudessem inseri-los em seus catálogos (exemplos: JSON-LD, *Dublin Core*, RDFa). Todos foram avaliados como “avançado”, mesmo resultado encontrado para as outras

instituições da amostra. Ou seja, mesmo no caso das instituições que obtiveram níveis mais baixos de *FAIRness*, este não foi um aspecto de difícil aderência. É um ponto positivo a ser exposto, aumentando a encontrabilidade dos dados na *web*.

Sendo assim, a equipe gestora do repositório da Unifesp pode avaliar a necessidade de investimentos quanto ao aspecto FsF-F3-01M, que diz respeito aos metadados incluírem o identificador dos dados que descrevem (e foi um ponto fraco de todos os repositórios). Mas recomenda-se que a equipe dê atenção prioritária à questão dos identificadores persistentes e dos metadados descritivos, que são essenciais para a descoberta/citação dos dados e estão contemplados dentro do núcleo básico proposto por Hodson *et al.* (2018) em sua escala 5 estrelas do FAIR.

Fazendo, então, uma comparação quanto à **encontrabilidade**, é possível notar que as pontuações variaram para cada repositório. Os aspectos nos quais todos os conjuntos de dados da amostra pontuaram igualmente foram o FsF-F1-01D (avançado), o FsF-F3-01M (incompleto) e o FsF-F4-01M (avançado). Ou seja, todos possuíam identificadores globalmente exclusivos e metadados fornecidos de uma maneira que os principais mecanismos de pesquisa podem inseri-los em seus catálogos.

Em compensação, nenhum deles conseguiu pontuar com relação aos metadados incluírem o identificador do dado que descrevem. O identificador persistente (FsF-F1-02D), como visto, variou entre os repositórios, com a maioria não conseguindo a validação pela ferramenta F-UJI. Já quanto aos metadados descritivos com elementos essenciais também houve uma variação entre os resultados: alguns conjuntos de dados pontuaram como inicial, outros como moderado e outros como avançado. Cada repositório, portanto, se encontra em um nível diferente quanto à descrição dos *datasets*.

Quadro 27 - Aderência quanto à Acessibilidade (Unifesp)

Código	FsF-A1-01M	FsF-A1-03D	FsF-A1-02M
VEPZWC	incompleto	incompleto	avançado
H7UNRI	incompleto	incompleto	avançado
BYDG95	incompleto	incompleto	avançado
GNFXNB	incompleto	incompleto	avançado
6TJHBJ	incompleto	incompleto	avançado
QPODEL	incompleto	incompleto	avançado

76JSFJ	incompleto	incompleto	avançado
EXAOHP	incompleto	incompleto	avançado
YRVDP8	incompleto	incompleto	avançado
339GI2	incompleto	incompleto	avançado
1CJNZG	incompleto	incompleto	avançado
YZZVOL	incompleto	incompleto	avançado
L9DH4T	incompleto	incompleto	avançado
3HVW0Q	incompleto	incompleto	avançado
1QOUY8	incompleto	incompleto	avançado

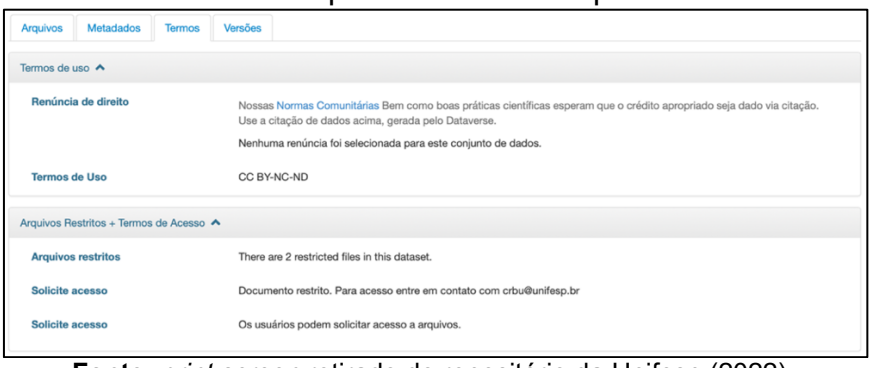
Fonte: elaborado pela autora (2022).

Já quanto à **acessibilidade**, os conjuntos de dados foram avaliados em três aspectos: 1) Os metadados contêm o nível de acesso e as condições de acesso aos dados (FsF-A1-01M); 2) Os dados são acessíveis por meio de um protocolo de comunicação padronizado (FsF-A1-03D) e 3) Os metadados são acessíveis por meio de um protocolo de comunicação padronizado (FsF-A1-02M).

Nota-se, pelo Quadro 27, que todos os conjuntos de dados da Unifesp só aderiram a um dos três aspectos avaliados pela ferramenta F-UJI: o FsF-A1-02M, em que todos foram avaliados como avançados. Para o primeiro aspecto (FsF-A1-01M), a ferramenta auxiliar verificou se os *datasets* possuíam informações sobre restrições ou direitos de acesso nos metadados, mas não foi capaz de encontrar essas informações nos metadados, mesmo caso da Unicamp, USP e UFABC.

Apesar desse resultado, em um *dataset* foi possível encontrar os termos de acesso e restrições na aba de "*terms*", conforme mostra a Figura 35. Caso semelhante foi visto no repositório da Unicamp (que também utiliza o *Dataverse*).

Figura 35 – Termos de acesso disponibilizado num conjunto de dados depositado no repositório da Unifesp



Fonte: *print screen* retirado do repositório da Unifesp (2022).

A ferramenta não foi capaz de recuperar essa informação, mas um usuário humano é capaz de encontrá-la e interpretá-la, entendendo que é possível solicitar o acesso aos arquivos. É fornecido, inclusive, um e-mail para contato em 'solicite acesso', o que não foi visto no caso da Unicamp. Cabe à equipe gestora do repositório da Unicamp determinar se vai aplicar melhorias neste aspecto, buscando fornecer os termos de acesso para todos os conjuntos de dados depositados. Vale lembrar também que, na publicação original dos princípios FAIR, Wilkinson *et al*. (2016) explicam que é preciso fornecer as condições exatas de acesso aos dados e, idealmente, a acessibilidade é especificada de forma que uma máquina possa entender automaticamente os requisitos e, em seguida, executá-los automaticamente ou alertar o usuário sobre os requisitos.

O repositório conta, inclusive, com um guia para deixar os arquivos restritos no Repositório de Dados[9]. O responsável pelo conjunto de dados pode fazer isso diretamente no repositório, editando os termos na hora de depositar os arquivos, ao contrário do caso de outras instituições da amostra onde era preciso contatar a equipe gestora para restringir acesso a um conjunto de dados.

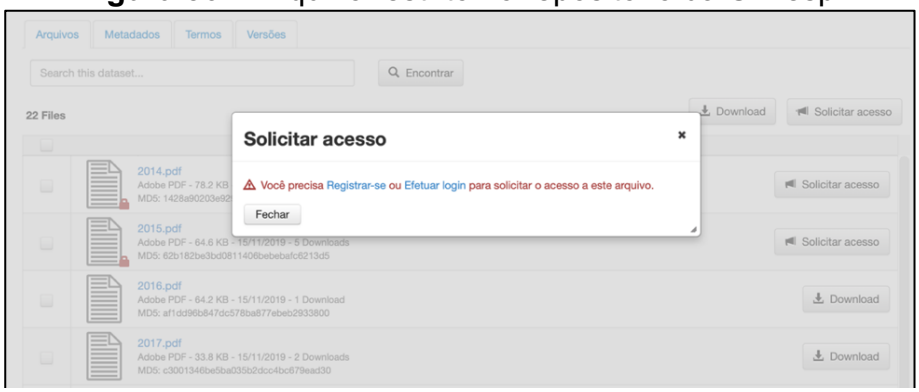
Na Figura 36 é possível observar o caso de um *dataset* restrito. Assim como no repositório da UNESP e da Unicamp, os metadados continuam disponíveis, mesmo que os arquivos não estejam, uma boa prática importante para o FAIR (e que é adotada pelo *Dataverse*). Nas perguntas frequentes (depósito de dados), inclusive, tem-se o seguinte aviso: "Uma vez publicados, não será possível excluir os dados. O sistema *Dataverse* permite apenas que fiquem em modo indisponível - desativado, tornando-os inacessíveis ao público. Isto garante o registro de todo histórico de edição do documento depositado, além de assegurar sua autenticidade. Antes de publicar seus dados, reveja todas as etapas no item PARA EDITAR UM CONJUNTO DE DADOS."

Assim como no caso da Unicamp, existe um botão para solicitar diretamente acesso aos arquivos, e somente após o *login* é possível fazer a solicitação (Figura 36). De acordo com Dias, Anjos e Rodrigues (2019), é essencial que essa prática seja adotada, fornecendo informações nos metadados que possibilitem identificar os indivíduos ou as instituições detentoras dos dados, além das informações sobre o

[9] Disponível em: https://bibliotecas.unifesp.br/images/servicos/Guia_%20deixar%20arquivos%20restritos%20no%20repositório%20de%20dados.pdf

conteúdo do arquivo. De tal modo o usuário pode decidir se aquele arquivo é útil para sua pesquisa, solicitando o acesso.

Figura 36 – Arquivo restrito no repositório da Unifesp



Fonte: *print screen* retirado do repositório da Unifesp (2022).

Já em relação ao segundo aspecto (FsF-A1-03D), que diz respeito aos dados serem acessíveis por meio de um protocolo de comunicação padronizado, a ferramenta examinou se os metadados incluíam um *link* para dados com base em protocolos de comunicação da *web* padronizados (*standard_data_protocol*), o que não foi encontrado, mesmo resultado dos outros cinco repositórios avaliados. Esse é, portanto, outro ponto que as instituições podem trabalhar. O conjunto de dados (https://doi.org/10.7910/DVN/WR4S9I) depositado no *Harvard Dataverse,* já citado neste trabalho, conseguiu aderir (avançado) a este aspecto, e pode ser usado para estudo.

Vale destacar, como feito para os outros repositórios, que todos os *datasets* da Unifesp são identificados por *links* que começam com https://. Essa questão também foi, inclusive, testada para os metadados (FsF-A1-02M), e a ferramenta F-UJI conseguiu localizar o protocolo padrão para acesso aos metadado: HTTPS.

Então, em “acessível”, sugere-se que as informações quanto aos níveis/condições de acesso sejam publicamente disponibilizadas para todos os conjuntos de dados, de modo legível por máquina. No repositório *Zenodo*, por exemplo, é possível verificar casos de sucesso como o uso do metadado ‘*dc:rights*’ com o seguinte valor: *info:eu-repo/semantics/openAccess*.

Fazendo uma comparação geral, é possível notar que todos os seis repositórios obtiveram notas semelhantes, variando entre 1 e 1,5. A única diferença foi que os conjuntos de dados de alguns repositórios (UFSCar e UNESP) pontuaram em nível inicial no aspecto FsF-A1-01M enquanto outros pontuaram como incompletos

(Unicamp, USP, UFABC, Unifesp). Para os outros dois aspectos avaliados (FsF-A1-03D e FsF-A1-02M), todos os conjuntos de dados dos seis repositórios pontuaram exatamente igual: incompleto e avançado, respectivamente.

A ferramenta F-UJI também avaliou os *datasets* quanto à **interoperabilidade,** e os conjuntos de dados da Unifesp não obtiveram bons resultados. Como visto no Gráfico 13 todos os *datasets* receberam nota 0 da ferramenta auxiliar. Foram avaliados três aspectos: 1) Os metadados são representados usando uma linguagem de representação de conhecimento formal (FsF-I1-01M); 2) Metadados usam recursos semânticos (FsF-I1-02M) e 3) Os metadados incluem *links* entre os dados e suas entidades relacionadas (FsF-I3-01M). Os resultados podem ser vistos no Quadro 28.

Quadro 28 - Aderência quanto à Interoperabilidade (Unifesp)

Código	FsF-A1-01M	FsF-A1-03D	FsF-A1-02M
VEPZWC	incompleto	incompleto	incompleto
H7UNRI	incompleto	incompleto	incompleto
BYDG95	incompleto	incompleto	incompleto
GNFXNB	incompleto	incompleto	incompleto
6TJHBJ	incompleto	incompleto	incompleto
QPODEL	incompleto	incompleto	incompleto
76JSFJ	incompleto	incompleto	incompleto
EXAOHP	incompleto	incompleto	incompleto
YRVDP8	incompleto	incompleto	incompleto
339GI2	incompleto	incompleto	incompleto
1CJNZG	incompleto	incompleto	incompleto
YZZVOL	incompleto	incompleto	incompleto
L9DH4T	incompleto	incompleto	incompleto
3HVW0Q	incompleto	incompleto	incompleto
1QOUY8	incompleto	incompleto	incompleto

Fonte: elaborado pela autora (2022).

Logo de cara é possível observar que todos os conjuntos de dados pontuaram como incompletos para todos os três aspectos, ou seja, não tiveram as características de interoperabilidade validadas pela ferramenta F-UJI. Quanto ao primeiro aspecto, por exemplo, que diz respeito a uma linguagem de representação de conhecimento formal, a ferramenta alegou não encontrar dados estruturados embutidos na página,

nem *typed links* baseados em RDF ou metadados RDF disponíveis por meio de negociação de conteúdo.

Era esperado que houvesse metadados estruturados e analisáveis (JSON-LD, RDFa) incorporados no código XHTML/HTML da página inicial. A equipe gestora do repositório pode avaliar a viabilidade de investimentos neste primeiro aspecto, considerando sua importância para que as máquinas sejam capazes de trocar e interpretar os dados. Essa qualidade está contemplada na terceira estrela da escala FAIR proposta por Hodson *et al.* (2018).

Já no aspecto FsF-I1-02M a ferramenta buscou por uso de *namespaces* de vocabulário e de recursos semânticos nos metadados, e todos os *datasets*, novamente, pontuaram como incompletos. Como já visto, o teste buscou extrair *namespaces* dos metadados baseados em RDF*,* que garantem que determinado conjunto de objetos tenha nomes exclusivos para que possa ser facilmente identificado. A ferramenta alegou que nenhuma correspondência de *namespace* de vocabulário foi encontrada, assim como nenhum *namespace* de vocabulários semânticos nos metadados. Um exemplo de vocabulário seria o *Schema.org.*

Os vocabulários usados para descrever conjuntos de dados precisam ser documentados e resolvíveis usando identificadores globalmente únicos e persistentes. Essa documentação precisa ser facilmente encontrada e acessível por qualquer pessoa que use o conjunto de dados. A equipe gestora pode avaliar a viabilidade de investimento para aderir a esta característica, que diz respeito à quarta estrela da escala proposta por Hodson *et al.* (2018).

Quanto ao último aspecto da interoperabilidade, que diz respeito às ligações entre os dados e suas entidades relacionadas, todos os conjuntos de dados pontuaram como “incompleto”, como já visto no Quadro 28. Era preciso indicar de forma explícita, nos metadados do item, os recursos relacionados (e preferencialmente por meio de *links* ou identificadores legíveis por máquina).

Uma referência qualificada é uma referência cruzada que explica sua intenção. O objetivo é criar tantos *links* significativos quanto possível entre recursos de (meta)dados para enriquecer o conhecimento contextual sobre os dados, e essa qualidade é indicada na quinta e última estrela da escala FAIR de Hodson *et al.* (2018).

Assim como no caso da USP e da UNESP, a ferramenta F-UJI alegou não encontrar nenhum recurso relacionado nos metadados *Dublin Core* ou *Datacite*. E, de fato, ao fazer uma análise manual nos registros dos conjuntos de dados não foi

possível encontrar recursos relacionados/referências qualificadas entre recursos, o que acaba por diminuir a interoperabilidade dos dados depositados. Importante relembrar que essas referências qualificadas ajudam a contextualizar os dados, que possuem um nível de abstração muito maior que publicações científicas. Sendo assim, implementar essa característica também aumentaria a reutilização dos dados depositados, auxiliando na sua interpretação.

Logo, os conjuntos de dados da Unifesp seguiram a tendência defendida por Henning *et al.* (2019a), que apontou a interoperabilidade como o elemento mais desafiador das práticas FAIR. Segundo os autores, os elementos da interoperabilidade são menos conhecidos e mais caros de implementar. Além disso, menos profissionais qualificados estão disponíveis para auxiliar na interoperabilidade.

A equipe gestora do repositório pode, então, adotar a escala de 5 estrelas de Hodson *et al*. (2018) para definir prioridades de investimento, uma vez que é de extrema importância que a instituição busque implementar tecnologias para aumentar a interoperabilidade de seus dados. Este trabalho pode ser feito em conjunto com outras instituições e até usar como base outros repositórios que utilizam o *Dataverse*, como o *Harvard Dataverse*.

Fazendo uma comparação geral, é possível notar que outros três repositórios obtiveram notas semelhantes, variando entre 0 e 1: UFSCar, USP e UNESP. No primeiro aspecto de interoperabilidade (FsF-I1-01M) todos esses três repositórios tiveram seus *datasets* avaliados como incompletos. Já o segundo aspecto (FsF-I1-02M) foi um ponto fraco geral, com todos os conjuntos de dados da amostra obtendo nível inicial (ou incompleto). Até mesmo a Unicamp, que obteve notas altas em interoperabilidade. Para o terceiro aspecto (FsF-I3-01M) os níveis de aderência variaram entre os repositórios, e três conseguiram atingir nível avançado: Unicamp, UFSCar, UFABC. Foi também o caso de alguns *datasets* da UNESP. Os melhores resultados foram os da Unicamp (a maioria dos *datasets* recebeu nota 3 de 4) e da UFABC (todos os *datasets* receberam nota 2/4).

Por fim, todos os *datasets* da Unifesp foram avaliados quanto à **reutilização**. Como já visto, a reutilização vêm sendo o ponto fraco dos repositórios da amostra, e o mesmo acontece com a Unifesp. Todos os *datasets* obtiveram nota 0 em reutilizável, assim como no caso da interoperabilidade.

Os dados foram testados em cinco aspectos pela ferramenta auxiliar: 1) Metadados especificam o conteúdo dos dados (FsF-R1-01MD); 2) Os metadados

incluem informações de licença sob as quais os dados podem ser reutilizados (FsF-R1.1-01M); 3) Os metadados incluem informações de proveniência sobre a criação ou geração dos dados (FsF-R1.2-01M); 4) Os metadados seguem um padrão recomendado pela comunidade de pesquisa-alvo dos dados (FsF-R1.3-01M) e 5) Os dados estão disponíveis em um formato de arquivo recomendado pela comunidade de pesquisa-alvo (FsF-R1.3-02D). O Quadro 29 apresenta os resultados encontrados.

Quadro 29 - Aderência quanto à Reutilização (Unifesp)

Código	FsF-R1-01MD	FsF-R1.1-01M	FsF-R1.2-01M	FsF-R1.3-01M	FsF-R1.3-02D
VEPZWC	incompleto	incompleto	incompleto	inicial	incompleto
H7UNRI	incompleto	incompleto	incompleto	inicial	incompleto
BYDG95	incompleto	incompleto	incompleto	inicial	incompleto
GNFXNB	incompleto	incompleto	incompleto	inicial	incompleto
6TJHBJ	incompleto	incompleto	incompleto	inicial	incompleto
QPODEL	incompleto	incompleto	incompleto	inicial	incompleto
76JSFJ	incompleto	incompleto	incompleto	inicial	incompleto
EXAOHP	incompleto	incompleto	incompleto	inicial	incompleto
YRVDP8	incompleto	incompleto	incompleto	inicial	incompleto
339GI2	incompleto	incompleto	incompleto	inicial	incompleto
1CJNZG	incompleto	incompleto	incompleto	inicial	incompleto
YZZVOL	incompleto	incompleto	incompleto	inicial	incompleto
L9DH4T	incompleto	incompleto	incompleto	inicial	incompleto
3HVW0Q	incompleto	incompleto	incompleto	inicial	incompleto
1QOUY8	incompleto	incompleto	incompleto	inicial	incompleto

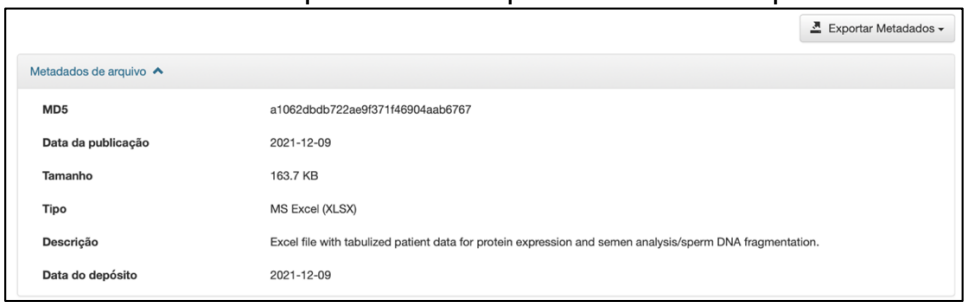
Fonte: elaborado pela autora (2022).

O primeiro aspecto (FsF-R1-01MD) avaliou pontos como a existência de informações mínimas fornecidas nos metadados sobre o conteúdo dos dados disponíveis, e todos os *datasets* pontuaram como “incompleto”. Ou seja, a ferramenta não conseguiu localizar, nos metadados, informações como tipo, tamanho ou variáveis medidas. Inclusive, a F-UJI não conseguiu recuperar nem o metadado “*object_type*” que foi encontrado para os outros repositórios, com a indicação de “*dataset*”.

A ferramenta não conseguiu extrair informações do arquivo como tamanho, assim como no caso de outras instituições da amostra: UFSCar e UNESP. Mas, ao entrar nos arquivos dos *datasets*, é possível encontrar o tamanho, o tipo, a descrição,

a data de depósito, a data de publicação, *etc*. (Figura 37). Ou seja, as informações estão declaradas, mas a ferramenta não foi capaz de extrair.

Figura 37 – Visualização do nome, tamanho e formato do arquivo de um conjunto de dados depositado no repositório da Unifesp



Fonte: *print screen* retirado do repositório da Unifesp (2022).

Isso pode estar ligado ao fato delas não estarem disponibilizadas nos metadados (do item) que podem ser extraídos em JSON, mesmo caso já comentado para outros repositórios como o da Unicamp. Essa hipótese se deve por causa do conjunto de dados do *Harvard Dataverse* (https://doi.org/10.7910/DVN/WR4S9I) que pontuou como avançado nesse aspecto. Nos metadados exportados do *dataset* em si é possível encontrar campos como "*name*", "*type*", "*format*" e "*size*" dos arquivos do conjunto de dados. Ou seja, como o repositório da Unifesp também utiliza *Dataverse*, pode adotar uma solução similar.

Além disso, a Unifesp não oferece a opção de exportar os metadados em JSON-LD, formato leve e facilmente interpretado por humanos e máquinas, o que é ideal para a interoperabilidade dos dados. O JSON-LD anota elementos em uma página, estruturando os dados, que podem ser usados pelos mecanismos de pesquisa para desambiguar elementos e estabelecer fatos em torno de entidades, que são então associados à criação de uma *web* mais organizada. Isso pode melhorar a contextualização dos dados, ajudando na reutilização.

Cabe então à equipe gestora do repositório avaliar as sugestões de acordo com as necessidades da comunidade. Isso porque, apesar de um usuário humano conseguir encontrar facilmente essas informações, um computador pode ter dificuldades de processar automaticamente, o que é preconizado pelo FAIR. Ou seja, não basta fornecer as informações sobre o *dataset*, elas precisam ser legíveis por máquina para serem FAIR.

O segundo aspecto (FsF-R1.1-01M) diz respeito aos metadados incluírem ou não informações de licença sob as quais os dados podem ser reutilizados. A licença é uma informação extremamente importante para que pesquisadores possam reutilizar os dados de pesquisa, mas a ferramenta F-UJI não conseguiu recuperá-la em um elemento de metadado apropriado para os *datasets* da Unifesp. O mesmo resultado foi encontrado para a Unicamp, USP, UFABC e UNESP, revelando um ponto fraco comum. No caso da UFSCar, apenas dois conjuntos de dados foram validados neste aspecto.

Conforme mostra a Figura 38, é possível identificar a aba “termos” com as informações sobre as renúncias de direito, ou seja, os termos de uso. Dos 15 *datasets*, 14 tem a licença CC0 declarada. É importante relembrar, porém, que essa é a licença padrão atribuída pelo *Dataverse* aos conjuntos de dados. Já o outro *dataset* possui a licença CC BY-NC-ND, com dois arquivos restritos.

Figura 38 – Termos de uso de um conjunto de dados depositado no repositório da Unifesp

Fonte: *print screen* retirado do repositório da Unifesp (2022).

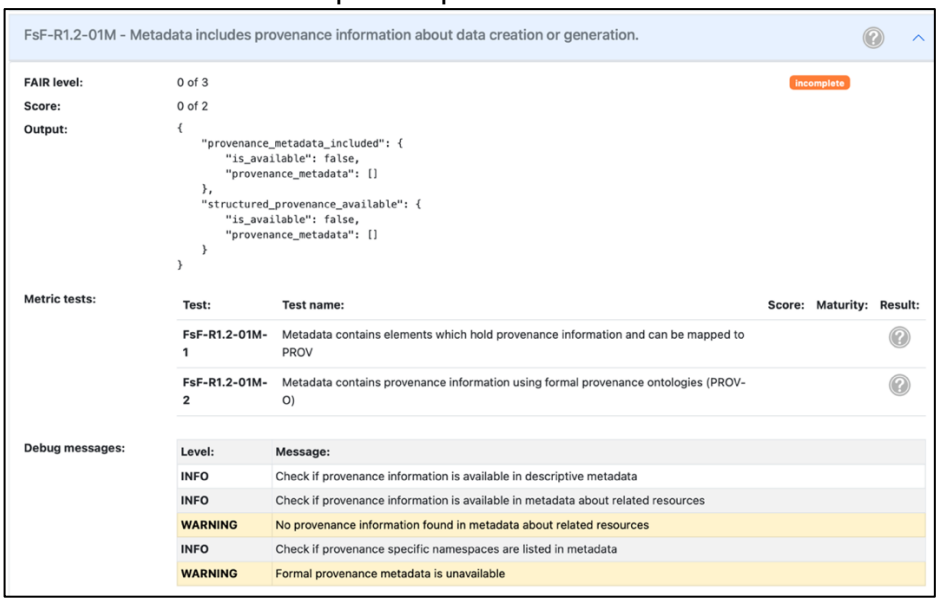
Pode-se citar novamente o caso do conjunto de dados (https://doi.org/10.7910/DVN/WR4S9I) depositado no *Harvard Dataverse*, que foi pontuado como “avançado” no aspecto FsF-R1.1-01M. Ao exportar os metadados em JSON percebe-se a indicação da licença da seguinte forma: *license":{"name":"CC0 1.0","uri":"http://creativecommons.org/publicdomain/zero/1.0"}.* A indicação da licença CC0, da forma como aparece para o usuário do repositório em *License/Data Use Agreement,* é seguida pelo *link* da respectiva licença, permitindo que o usuário a acesse e que a máquina a interprete. O mesmo acontece para os dois *datasets* do repositório da UFSCar que pontuaram como “avançado”.

Logo, essa é uma questão que a Unifesp pode avaliar a pertinência para melhorias, lembrando que para maximizar a reutilização é preciso atribuir uma licença bem definida e reconhecida internacionalmente, de modo que as condições de reutilização sejam claramente comunicadas (HODSON *et al.*, 2018), tanto para os

usuários humanos como para os agentes computacionais. A declaração de licença, inclusive, se encontra no segundo nível da escala FAIR dos autores, sendo, portanto, um item de prioridade.

O próximo aspecto (FsF-R1.2-01M) avaliou informações de proveniência nos metadados, e todos os 15 *datasets* foram novamente avaliados como incompletos. Conforme mostra a Figura 39, a ferramenta checou os metadados e não foi capaz de recuperar as informações necessárias. Todos os demais repositórios tiveram seus conjuntos de dados avaliados como “moderado” neste aspecto, indicando um ponto que a Unifesp pode investir para seguir a tendência regional.

Figura 39 – *Output* exibido pela ferramenta F-UJI ao avaliar um conjunto de dados da Unifesp no aspecto FsF-R1.2-01M



Fonte: *print screen* retirado do site da ferramenta F-UJI (2022).

De acordo com Hodson *et al.* (2018), os metadados básicos permitirão a descoberta de dados, mas são necessárias informações de proveniência muito mais ricas para entender como, por que, quando e por quem os dados foram criados. Isso inclui padrões da comunidade sobre como os dados foram criados (incluindo, por exemplo, protocolos de pesquisa, padrões mínimos de relatórios, processos experimentais, informações sobre calibração e localização do sensor).

Também deve incluir qualquer informação sobre redução de dados ou processos de transformação que são empregados em um determinado domínio para tornar os dados mais utilizáveis, compreensíveis ou "prontos para a ciência". A capacidade de humanos e máquinas de avaliar e selecionar dados com base em

critérios relacionados às informações de proveniência é essencial para a reutilização de dados (HODSON *et al*., 2018).

Deve haver clareza, de modo que qualquer pessoa que tente usar os dados tenha todas as informações de que precisa sem entrar em contato com os criadores. Os autores explicam que:

> O modelo *Open Archival Information System* (OAIS) apresenta um conceito de suma importância a esse respeito, argumentando que os dados devem ser "compreensíveis de forma independente". Isso significa que os dados devem ser acompanhados de informações suficientes 'para permitir que sejam interpretados, compreendidos e utilizados pela Comunidade Designada sem ter que recorrer a recursos especiais que não estão amplamente disponíveis, incluindo indivíduos nomeados (HODSON *et al*., 2018, p.14, tradução nossa).

Para permitir a reutilização mais ampla, os dados devem ser acompanhados por uma 'pluralidade de atributos relevantes', tornando sua interpretação clara pelos usuários. Essa característica se encontra no quarto nível da escala 5 estrelas dos autores. Sem essas informações sobre o contexto de criação dos dados de pesquisa, é menos provável que usuários reutilizem os *datasets* do repositório. Por isso a importância de investir nesta questão.

Quanto ao quarto aspecto (FsF-R1.3-01M), a ferramenta verificou se um padrão de metadados específico da comunidade é detectado usando *namespaces* e se está listado no registro Re3data. Esse foi o único aspecto de “reutilizável” em que os *datasets* da Unifesp foram validados. O nível inicial obtido diz respeito ao uso de um padrão de metadados multidisciplinar, endossado pela comunidade (*RDA Metadata Standards Catalog*), que é detectado através de *namespaces.* A ferramenta F-UJI detectou o padrão *Dublin Core*, assim como no caso dos demais repositórios, indicando que, de fato, existe essa tendência de escolha. O DC é, portanto, o padrão utilizado por todos os repositórios da amostra. Isso é um bom indicativo quando se pensa na interoperabilidade entre os sistemas.

Por fim, quanto ao último aspecto (FsF-R1.3-02D), a ferramenta verificou se o formato do arquivo de dados é aberto/de longo prazo/científico. Todos os conjuntos de dados foram avaliados como incompletos, mesmo resultado encontrado para todos os outros repositórios da amostra, indicando outro ponto fraco em comum.

A ferramenta F-UJI não conseguiu executar as verificações de formato de arquivo, uma vez que os identificadores de conteúdo de dados estavam

indisponíveis/inacessíveis. Mas, por conter apenas 15 conjuntos de dados, foi feita a análise manual dos formatos dos arquivos, para entender quais eram usados. No elemento "*type*" do arquivo foi possível identificar os seguintes formatos: Adobe PDF, MS Excel (XLSX), MS Word (docx), *image/bpm,* TIFF *Image*, vídeo/mpeg, vídeo/mp4, PNG *Image* e JPEG *Image.* São valores equivalentes aos encontrados para os repositórios que usam *Dataverse*.

Assim como no caso da UFSCar, da Unicamp, da UFABC, da USP e da UNESP, há o depósito de arquivos em formatos proprietários, ou seja, elaborados em *softwares* pagos e, portanto, que não estão disponíveis sem barreiras de acesso e custo. Isso pode acabar afetando tanto o acesso quanto a reutilização desses dados ao longo do tempo. Cabe à equipe gestora definir se deseja fazer um controle dos formatos depositados, incentivando o depósito dos não proprietários, garantindo que mais usuários tenham acesso aos dados de pesquisa. Esse é um dos pontos a ser avaliado para aumentar o nível de reutilização, junto com os demais que já foram comentados. Como visto, reutilização é um dos pontos fracos da instituição.

Sendo assim, os resultados obtidos por meio da ferramenta F-UJI demonstram uma falta de aderência geral aos princípios, com foco na interoperabilidade, em que todos os *datasets* receberam nota 0 (sendo o máximo 4), e reutilização, em que todos também receberam nota 0 (sendo o máximo 10). Com relação ao acesso, todos os *datasets* da amostra foram pontuados com a mesma nota: 1 de 3, exatamente igual ao repositório da Unicamp, da USP e da UFABC. Quanto à encontrabilidade dos dados, todos os conjuntos de dados receberam a nota 2,5 (sendo o máximo 7), um resultado abaixo da média obtida nos outros repositórios.

Fazendo uma comparação geral, é possível notar que o repositório da Unifesp obteve a menor pontuação dentre os seis da amostra. O repositório da UNESP, outro que obteve pontuações baixas para reutilizável, teve todos os seus conjuntos de dados avaliados com a nota 1. Ou seja, a reutilização pode ser vista como uma prioridade para a Unifesp. Mas não se pode deixar de relembrar que "reutilizável" foi a faceta mais desafiadora para todos os repositórios avaliados. Nenhum conjunto de dados obteve pontuação maior que 4, sendo essa a nota de apenas dois *datasets* da UFSCar. Existe, portanto, a necessidade geral de maiores investimentos para aumentar o potencial de reutilização dos dados de pesquisa depositados, uma vez que este é o objetivo final dos princípios FAIR.

No primeiro aspecto de reutilização (FsF-R1-01MD) quatro dos seis repositórios tiveram seus *datasets* validados, apesar de em nível inicial. Ou seja, é preciso aplicar melhorias para que os metadados especificam o conteúdo dos dados, fornecendo informações como tamanho, nome e tipo do arquivo. Já o segundo aspecto (FsF-R1.1-01M) foi um ponto fraco geral, com todos os conjuntos de dados da amostra obtendo nível incompleto. Ou seja, a licença de uso não está clara nos metadados dos *datasets*.

É de suma importância que os repositórios indiquem a licença num metadado adequado, o que deve ser feito com a sintaxe correta para processamento automático por máquinas. Esse caso foi visto com mais detalhes na seção da UFSCar, onde alguns *datasets* pontuaram como avançado. Para o terceiro aspecto (FsF-R1.2-01M), todos os repositórios tiveram seus conjuntos de dados avaliados como nível moderado, exceto a Unifesp (nível incompleto). Ou seja, a tendência é que os dados depositados tenham uma boa descrição quanto a sua proveniência, com uma pluralidade de atributos relevantes para contextualizar a origem dos dados.

Já no quarto aspecto (FsF-R1.3-01M) é possível observar outra tendência: todos os conjuntos de dados dos seis repositórios foram avaliados pela ferramenta como nível inicial, e todos utilizam o padrão DC. Como o padrão não foi detectado usando *namespaces* ou esquemas encontrados nos metadados fornecidos, foi atribuído o nível inicial. Por fim, quanto ao último aspecto (FsF-R1.3-02D), foi possível observar outra tendência: os conjuntos de dados de todos os repositórios da amostra foram avaliados como incompletos. Ou seja, os dados não estavam disponíveis em formatos recomendados pela comunidade de pesquisa. Alguns exemplos de formatos validados pela ferramenta são: csv, *tab-separated values* (tsv), GNU Zip (GZip) e *plain text*. Há, portanto, melhorias que podem ser implementadas quanto à reutilização, em todos os seis repositórios.

Sendo assim, após todas as seis avaliações dos repositórios de dados de pesquisa de São Paulo, é possível confirmar algumas questões como: as facetas de interoperabilidade e a reutilização foram, de fato, as mais difíceis de aderir, indo ao encontro dos resultados do cenário internacional (DUNNING; SMAELE; BÖHMERO, 2017). Além disso, os níveis de aderência gerais aos princípios FAIR foram, como defendido por Henning *et al*. (2019a), baixos. O maior nível de aderência foi de 50%, o que, para o cenário dos repositórios da amostra, representa um bom resultado. É importante lembrar que o FAIR é um espectro e dificilmente um conjunto de dados

obterá 100% de aderência. Outra questão a ser apontada é que o DC é, de fato, o esquema de metadados utilizado por todos os seis repositórios da amostra, indo de acordo com os estudos de Simionato (2017) e de Sanchez, Silva e Vechiato (2019).

Também é interessante mostrar que, conforme defenderam Silva e Rodrigues (2021), a plataforma *Dataverse* facilita a criação de dados de pesquisa FAIR. Os conjuntos de dados dos repositórios da Unicamp e da UFABC, ambos utilizando esse *software*, obtiveram as melhores pontuações gerais: 50% e 37% de aderência, respectivamente. Rodrigues, Dias e Lourenço (2022, p. 326) chegaram a uma conclusão similar ao afirmar que “Os repositórios em maior conformidade com os Princípios FAIR foram aqueles estabelecidos mediante o uso do *software Dataverse*”.

A falta de aderência geral pode também estar ligada com fatores como: 1) os repositórios são recentes, surgindo com o movimento de criação de uma rede de repositórios de dados das universidades públicas do Estado de São Paulo, sob a coordenação da FAPESP; 2) os repositórios são multidisciplinares, o que contribui para o alto grau de heterogeneidade dos dados e dificulta a implementação de algumas funcionalidades específicas de domínio. Existem vários tipos de dados de pesquisa, como visto, por exemplo, no Quadro 1.

Atender às especificidades de cada tipo de dados e de cada domínio pode se tornar um desafio para os repositórios. É preciso, portanto, considerar todas essas questões e investir em equipes multidisciplinares para a gestão dos repositórios de dados de pesquisa, permitindo que melhorias sejam aplicadas em prol de dados cada vez mais FAIR. Isso não só aumenta o potencial científico dos dados e auxilia na agilização da ciência, como também aumenta a visibilidade e prestígio da instituição e de seus pesquisadores.

5.2 Metadados descritivos para a representação de dados de pesquisa

Os metadados são essenciais quando se pensa em dados FAIR. Na escala cinco estrelas de Hodson *et al.* (2018), Figura 17, é possível observar que eles aparecem no núcleo básico. Ou seja, mesmo que em níveis iniciais de FAIR, é preciso estar atento aos metadados, descrevendo os conjuntos de dados com uma pluralidade de atributos ricos. Isso aumenta a encontrabilidade dos dados e permite sua citação, o que também afeta a reutilização.

Além disso, os metadados também são citados no segundo, terceiro, quarto e quinto nível da escala, demonstrando o quão indispensáveis são quando se fala de dados FAIR. No segundo nível, por exemplo, que diz respeito a aumentar o acesso, há a preocupação com a declaração das licenças de uso dos dados nos metadados. Já no terceiro nível, que diz respeito ao uso de padrões, é preciso que os metadados usem linguagens formais/acessíveis de representação do conhecimento; além de atenderem aos padrões do domínio e incluírem explicitamente o identificador dos dados que descrevem.

No quarto nível tem-se um foco grande nos metadados: metadados ricos e FAIR, trazendo a importância de questões como os dados serem ricamente descritos com uma pluralidade de atributos precisos/relevantes e metadados usarem vocabulários que seguem os princípios FAIR. Por fim, no quinto nível, que diz respeito às informações de proveniência e contextualização, tem-se a presença de metadados ao citar que eles devem estar associados com dados de proveniência; devem incluir referências qualificadas para outros (meta)dados e serem acessíveis, mesmo que os dados em si não estejam mais disponíveis.

Visto então a importância dos metadados para dados FAIR, optou-se por sugerir um conjunto mínimo de metadados descritivos para uma boa representação dos dados de pesquisa em repositórios. Assim as equipes gestoras têm uma base para investir em descrições mais ricas dos dados depositados em seus repositórios. Mas antes é importante comentar que foram encontrados diferentes padrões de metadados nos diferentes repositórios da amostra, e no Quadro 30 é possível visualizar os formatos presentes em cada um deles e o respectivo *software* utilizado.

Quadro 30 – Padrões de metadados adotados nos repositórios da amostra

Repositório	Padrão de metadados adotados	Software adotado
UFSCar	*Dublin Core*	*DSpace*
UNESP	*Dublin Core*	*DSpace*
Unicamp	*Dublin Core* *Data Documentation Initiative (DDI)* *Schema.org*	*Dataverse*
USP	*Dublin Core*	*DSpace*
UFABC	*Dublin Core* *Data Documentation Initiative (DDI)* *Schema.org*	*Dataverse*
Unifesp	*Dublin Core* *Data Documentation Initiative (DDI)*	*Dataverse*

Fonte: elaborado pela autora (2023).

Nota-se que os repositórios da Unicamp, da UFABC e da Unifesp permitiam a exportação dos metadados em diferentes formatos além do DC. Em comum, todos adotavam o *software Dataverse*. Nos repositórios de dados da UFSCar, USP e UNESP foi detectado apenas o uso do DC, padrão de metadados pré-definido do *DSpace*. O DDI, por exemplo, é um padrão internacional para descrever dados produzidos por pesquisas nas ciências sociais, comportamentais, econômicas e da saúde. Ele é gratuito e pode documentar e gerenciar diferentes estágios no ciclo de vida dos dados. Logo, percebe-se seu foco para a descrição dos dados de pesquisa. Outra vantagem é que esse padrão facilita a compreensão, interpretação e uso por pessoas, sistemas de *software* e redes de computadores, se alinhando com o que é proposto pelo FAIR.

Já o *Schema.org*, de acordo com Machado (2022, p. 11), é uma iniciativa que "[...] propõe criar, manter e promover esquemas para dados estruturados, visando favorecer a interpretação dos mecanismos de busca e assim propiciar melhores experiências aos usuários". Sua aplicação para a representação de recursos agrega valor semântico aos dados, o que favorece o processamento automático por máquina. Logo, também se alinha com o que é proposto pelo FAIR.

Considerando então os padrões adotados, seguiu-se para a indicação de um conjunto mínimo de metadados descritivos para a representação dos dados de pesquisa. O objetivo não era ser exaustivo, mas sim sugerir elementos de metadados a serem adotados de tal forma que os dados sejam devidamente encontrados, acessados e reutilizados. A sugestão desse conjunto foi baseada nas avaliações feitas pela ferramenta F-UJI, que ao avaliar um conjunto de dados, atribui níveis de conformidade para cada aspecto do FAIR (inicial, moderado e avançado).

Por meio de alguns conjuntos de dados que foram avaliados como avançados quanto aos metadados descritivos utilizados na sua representação (como os da Unicamp e da UFABC; e os utilizados como comparação do *Havard Dataverse* e *Zenodo*), extraiu-se alguns elementos que permitem uma boa descrição dos dados. Ou seja, ao usar esses elementos para a sua representação, os dados de pesquisa seriam avaliados como "avançados" pela F-UJI, indicando um alto nível de aderência, com todos os metadados essenciais para a devida citação do *dataset*. No Quadro 31 é possível observar esses elementos junto com uma breve definição e com a faceta do FAIR a qual corresponde.

Quadro 31 – Conjunto mínimo de metadados descritivos para dados de pesquisa

Elemento de metadados	Definição	Faceta do FAIR
creator	Entidade responsável principalmente por criar o recurso	Encontrabilidade e Reutilização
title	Nome dado ao recurso	Encontrabilidade
summary/abstract	Resumo do recurso	Encontrabilidade
publisher	Entidade responsável por disponibilizar o recurso	Encontrabilidade e Reutilização
keywords	Palavras que resumem os temas principais de um recurso	Encontrabilidade
description	Descrição resumida do propósito, natureza e escopo do recurso	Encontrabilidade e Reutilização
subject	Tópico do recurso	Encontrabilidade
publicationDate	Data que o recurso foi publicado	Encontrabilidade e Reutilização
type	Tipo do recurso	Encontrabilidade e Reutilização
identifier	Identificador do recurso	Encontrabilidade e Reutilização
format	Formato do arquivo, suporte físico ou dimensões do recurso	Encontrabilidade
size	Tamanho do recurso	Encontrabilidade e Reutilização
accessRights	Informações sobre quem pode acessar o recurso ou uma indicação de seu *status* de segurança	Acessibilidade
relation	Outros recursos relacionados ao recurso original	Interoperabilidade
rights	Informações sobre os direitos detidos no recurso/sobre o recurso	Reutilização

Fonte: elaborado pela autora (2023).

Os metadados descritivos citados no Quadro 31 foram mais bem comentados na seção 5.1, enquanto eram testados, extraídos (ou não) e avaliados pela ferramenta F-UJI. É interessante comentar que a maioria dos conjuntos de dados dos seis repositórios foram avaliados em um nível moderado de descrição (metadados descritivos), indicando que podem ser feitas representações mais ricas para os dados.

É importante também lembrar que, muitas vezes, as informações como “licença” não foram validadas pela F-UJI porque não se encontravam em metadados adequados, impossibilitando o processamento automático. Para que as máquinas consigam interpretar todos os dados sobre os dados é preciso se atentar ao uso dos elementos de metadados adequados, com a sintaxe correta.

O preenchimento dos valores foge do escopo deste estudo, mas indicar (por exemplo: no manual/guia institucional de autoarquivamento) os elementos mínimos de preenchimento para uma descrição mais rica dos dados de pesquisa em repositórios é de suma importância, aumentando sua encontrabilidade, acessibilidade, interoperabilidade e reutilização.

Como para todos os seis repositórios foi verificado o uso do *Dublin Core*, a indicação dos elementos de metadados do Quadro 31 se deu com base nesse padrão. As equipes gestoras podem entrar no *site* oficial para analisar outros elementos que podem ser empregados para tornar as descrições as mais ricas possíveis como: *language, identifierCitation, descriptionSponsorship, modifiedDate, etc.* Isso porque os metadados "[...] representam os recursos informacionais e nos repositórios de dados de pesquisa também precisam ser bem estruturados, de modo que possam atender à descrição, referência, significação, uso e reuso de dados em pesquisas científicas" (FELIPE; SANTOS, 2022, p. 2).

Sendo assim, é essencial que os repositórios busquem fornecer metadados o suficiente para que os usuários consigam encontrar, acessar e reutilizar os dados, tudo isso enquanto esses dados se conectam com outros recursos na *web* como as publicações relacionadas. Como os metadados se ligam a todas as facetas do FAIR, investir em representações ricas é importante.

Exatamente por isso objetivou-se indicar esses 15 elementos de metadados, buscando auxiliar as equipes gestoras na representação de seus dados de pesquisa. Esses metadados podem ser indicados em manuais de autoarquivamento ou preenchidos pela equipe, dando maior contexto para os dados depositados e aumentando seu potencial de reutilização, o que também contribui para a visibilidade da instituição. Ao investir em descrições ricas, os dados depositados serão cada vez mais FAIR.

6 CONCLUSÃO

No paradigma da *e-Science*, os dados ganham um papel central, deixando de ser meros subprodutos da pesquisa para se tornarem ativos competitivos que, quando compartilhados, trazem inúmeras vantagens. Racionalização de recursos, reprodutibilidade da pesquisa, maior visibilidade/transparência e agilização do ciclo científico são algumas delas. Cada vez mais partes interessadas como agências de financiamento, agências governamentais e editores estão defendendo, e até exigindo, o compartilhamento e gestão dos dados de pesquisa, demandando dos pesquisadores e das suas instituições novas habilidades e serviços.

Mas não basta simplesmente compartilhar os dados, é preciso se atentar às boas práticas internacionalmente adotadas para garantir que o máximo benefício possa ser extraído deles. É nesse cenário que se fala da importância dos princípios FAIR, que já são reconhecidos mundialmente como elementos-chave para boas práticas em todos os processos de gestão de dados. Por isso, o objetivo dessa pesquisa foi analisar o cenário regional quanto à adoção dos princípios em dados depositados em repositórios de São Paulo. Todos os repositórios da amostra estão vinculados a instituições de ensino superior, que estão em contato direto com os pesquisadores.

A partir das avaliações realizadas pela ferramenta auxiliar F-UJI, que testou automaticamente todos os conjuntos de dados depositados quanto aos princípios FAIR, foi possível analisar pontos fortes e fracos dos repositórios e identificar alguns padrões. O primeiro padrão diz respeito ao uso do *software Dataverse*, que realmente estava associado com os repositórios em maior conformidade com os princípios FAIR (Unicamp e UFABC), o que foi defendido por autores na literatura. Outro padrão interessante é o uso do *Dublin Core* em todos os repositórios. Mesmo quando os dados podiam ser extraídos em outros formatos como *Schema.org* ou DDI, havia a opção do DC. Há, portanto, essa tendência, como esperado.

Ainda no que se refere aos metadados, na seção 5.2 foi possível elaborar e sugerir um conjunto mínimo de elementos para a representação dos dados de pesquisa, baseado em resultados extraídos pela F-UJI. Logo, um bom conjunto mínimo de metadados descritivos para dados de pesquisa inclui: *creator, title, summary/abstract, publisher, keywords, publicationDate, type, identifier, description, subject, format, size, acessRights, rights* e *relation*. Essa indicação de um conjunto de

15 elementos pode auxiliar as equipes gestoras na elaboração de representações significativas dos dados de pesquisa, permitindo que eles sejam encontráveis, acessíveis, interoperáveis e reutilizáveis.

Como já esperado, a aderência dos conjuntos de dados dos repositórios da amostra aos princípios FAIR foi baixa. A maior pontuação geral 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. Quando comparados com *datasets* de outros repositórios como *Harvard Dataverse* e Zenodo, conhecidos internacionalmente, é possível notar que ainda há melhorias a serem aplicadas no cenário nacional em prol de dados FAIR.

Foi possível identificar que, de fato, a interoperabilidade e a reutilização foram as facetas mais difíceis de aderir. O único repositório que se destacou quanto a uma delas foi o da Unicamp, onde a maioria dos conjuntos de dados conseguiram pontuação 3 de 4 em interoperável. Apesar desse ótimo resultado, esses mesmos conjuntos de dados receberam notas baixas em reutilizável, indicando um ponto fraco comum entre todos os seis repositórios.

A melhor aderência foi em “encontrável”, onde conjuntos de dados conseguiram obter notas como 6 de 7 (Unicamp), 4 de 7 (UFSCar e UFABC) e 3,5 de 7 (UNESP e UFSCar novamente). Mas um ponto fraco comum foi a falta de validação pela ferramenta dos identificadores persistentes. PIDs estão, junto dos metadados, no núcleo básico da escala cinco estrelas FAIR e, portanto, é importante que as equipes avaliem essa questão, buscando adotar identificadores como o DOI.

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). Um ponto fraco comum que pode ser avaliado pelas equipes gestoras é a declaração nos metadados do nível e das condições de acesso aos dados. Apenas dois repositórios conseguiram pontuar neste aspecto, mas em nível inicial: UFSCar e UNESP.

No repositório da UFSCar, a ferramenta conseguiu recuperar uma licença CC. Entretanto, a informação não estava legível por máquina. O mesmo aconteceu no repositório da UNESP, com uma única diferença: a ferramenta recuperou a informação “acesso aberto” em vez de uma licença CC. Sendo assim, recomenda-se

que as instituições busquem declarar de forma explícita as condições de acesso, de tal modo que as máquinas consigam processar automaticamente (o que é fundamental para ser FAIR). Seria interessante, também, adotar indicações padronizadas das condições de acesso, usando um mesmo padrão para facilitar o entendimento dos usuários humanos e das máquinas.

Isso também vale para a questão da declaração nos metadados das licenças de uso dos dados, segundo aspecto testado em reutilização. Nenhum dos repositórios teve seus conjuntos de dados validado pela ferramenta F-UJI, exceto dois (dos 15 depositados) *datasets* da UFSCar. Como já visto, além de fornecer as licenças, é preciso adotar sintaxes válidas e investir em homogeneização e convergência, para que o processo de interoperabilidade entre as instituições possa acontecer de forma transparente, evitando esforços de integração adicionais.

Apesar dos pontos fracos em comum, indicando um nível ainda inicial de implementação dos princípios FAIR, cada repositório conta com suas particularidades. Alguns, por exemplo, são dedicados exclusivamente aos dados de pesquisa. Outros funcionam num esquema híbrido, com outros recursos (dissertações, artigos, teses, *etc*.) no mesmo sistema. Por isso optou-se por dedicar uma seção para cada repositório, trazendo sugestões individuais de acordo com os pontos fracos da instituição. Assim, a equipe gestora pode avaliar a pertinência e a aplicabilidade das sugestões, tendo um norte para investir em melhorias em seu repositório. Como já dito, é difícil que os dados 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.

Sendo assim, foi possível trazer um *feedback* para as instituições, que podem decidir como e se desejam adotar as sugestões. Entretanto, é importante que o Brasil siga o cenário internacional e invista cada vez mais no potencial dos dados de pesquisa. Os princípios FAIR têm um papel importante nisso.

Vale lembrar que os resultados aqui expostos se limitam ao que foi entregue automaticamente pela ferramenta auxiliar F-UJI. Todas as suas métricas estão disponíveis abertamente no *site*. Qualquer outro pesquisador pode, inclusive, reproduzir essa pesquisa utilizando a ferramenta e os mesmos conjuntos de dados.

As análises manuais foram feitas apenas em caráter adicional, para confirmar alguns resultados entregues pela F-UJI. Logo, todas as notas e níveis (inicial, moderado, avançado e incompleto) foram atribuídos pela ferramenta, e não pela

pesquisadora, que se responsabilizou por interpretar os dados, fazer comparações, indicar tendências e trazer *feedbacks* para as instituições da amostra.

Vale também destacar que os resultados aqui encontrados não podem ser generalizados, posto que a amostra investigada se refere a um pequeno recorte do todo, ou seja, da regionalização dos objetos estudados. Em estudos futuros pretende-se fazer uma comparação do cenário de São Paulo com outros cenários nacionais, buscando averiguar se existe uma tendência nos níveis de aderência ao FAIR no Brasil.

REFERÊNCIAS

ALVES, R. C. V. **Metadados como elementos do processo de catalogação**. 2010. 134 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010. Disponível em: https://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/alves_rachel.pdf. Acesso em: 30 jun. 2022.

AMARAL, F. **Introdução à ciência de dados: mineração de dados e Big Data**. Rio de Janeiro: Alta Books, 2016.

ARAKAKI, A. C. S.; ARAKAKI, F. A. Dados e metadados: conceitos e relações: concepts and relationships. **Ciência da Informação**, v. 49, n. 3, 2020. Disponível em: http://hdl.handle.net/20.500.11959/brapci/163406. Acesso em: 22 jul. 2022.

ARAÚJO, C. A. A. Teorias e tendências contemporâneas da Ciência da Informação. **Informação em Pauta**, Fortaleza, v. 2, n. 2, p. 9-34, 2017. Disponível em: http://www.periodicos.ufc.br/informacaoempauta/article/view/20162. Acesso em: 20 jan. 2021.

ARAÚJO, D. G.; DIAS, G. A.; AUTRAN, M. M. M. Compartilhamento de dados no contexto da ciência brasileira: um estudo integrativo. **Informação & Informação**, v. 26, n. 3, p. 100-124, 2021. Disponível em: https://brapci.inf.br/index.php/res/v/165618. Acesso em: 25 fev. 2022.

AUSTIN, C. C. et al. Key components of data publishing: using current best practices to develop a reference model for data publishing. **International Journal on Digital Libraries**, v. 18, n. 2, p. 77–92, 2016. Disponível em: https://link.springer.com/10.1007/s00799-016-0178-2. Acesso em: 27 jul. 2022.

BERNERS-LEE, T. **Linked Data**. 2009. Disponível em:https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 18 ago. 2021.

BERTIN, P.; VISOLI, M.; DRUCKER, D. A gestão de dados de pesquisa no contexto da e-science: benefícios, desafios e oportunidades para organizações de p&d. **PontodeAcesso**, Salvador, v. 11, n. 2, p. 34-48, 2017. Disponível em: https://portalseer.ufba.br/index.php/revistaici/article/view/21449. Acesso em: 04 jan. 2021.

BONETTI, L. G. **Serviços de gestão de dados de pesquisa em bibliotecas universitárias brasileiras**. 2019. 87 f., il. Trabalho de Conclusão de Curso

(Bacharelado em Biblioteconomia)—Universidade de Brasília, Brasília, 2019. Disponível em: https://bdm.unb.br/handle/10483/26427. Acesso em: 10 mai. 2022.

BORGMAN, C. L. **Big data, little data, no data: scholarship in the networked world**. Cambridge; London: The MIT Press, 2015.

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.

CÓRDULA, F. R.; ARAÚJO, W. J. O compartilhamento de dados científicos na era do e-Science. *In*: DIAS, G. A.; OLIVEIRA, B. M. J. F.(org.). Dados científicos: perspectivas e desafios. Paraíba: Editora UFPB, 2019. p. 177-187. Disponível em: http://www.editora.ufpb.br/sistema/press5/index.php/UFPB/catalog/book/359. Acesso em: 10 ago. 2021.
CURTY, R. Abordagens de reúso e a questão da reusabilidade dos dados científicos. **Liinc em Revista**, Rio de Janeiro, v. 15, n. 2, 2019. Disponível em: http://revista.ibict.br/liinc/article/view/4777. Acesso em: 17 ago. 2021.

COSTA, M. P. da; LEITE, F. C. L. Fatores que exercem influência na comunicação dos dados de pesquisa: uma revisão sistematizada da literatura no campo da Ciência da Informação. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 19., 2018. **Anais** [...] Londrina: Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual de Londrina (PPGCI/UEL), 2018. Disponível em: http://enancib.marilia.unesp.br/index.php/XIXENANCIB/xixenancib/paper/viewFile/1265/1 869. Acesso em: 26 nov. 2021.

DAVIS, H. M.; VICKERY, J. N. Datasets, a shift in the currency of scholarly communication: implications for library collections and acquisitions. **Serials Review**, v. 33, n. 1. p. 26-32, 2007. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0098791306001675. Acesso em: 12 jan. 2021.

DEVARAJU, A. *et al.* From Conceptualization to Implementation: FAIR Assessment of Research Data Objects. **Data Science Journal**, London, v. 20, n. 1, p. 1-14, 2021. Disponível em: https://datascience.codata.org/articles/10.5334/dsj-2021-004/. Acesso em: 15 jun. 2021.

DIAS, G. A.; ANJOS, R. L.; RODRIGUES, A. A. Os Princípios FAIR: viabilizando o reuso de dados científicos. *In*: DIAS, G. A.; OLIVEIRA, B. M. J. F.(org.). **Dados científicos:** perspectivas e desafios. Paraíba: Editora UFPB, 2019. p. 177-187.

Disponível em: http://www.editora.ufpb.br/sistema/press5/index.php/UFPB/catalog/book/359. Acesso em: 10 ago. 2021.

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.

FELIPE, C. B. M.; SANTOS, R. F. Avaliação de metadados em repositórios de dados de pesquisa sobre biodiversidade. **Em Questão,** Porto Alegre, v. 28, n. 3, p. 1-19, 2022. Disponível em: https://www.seer.ufrgs.br/index.php/EmQuestao/article/view/117591. Acesso em: 27 jul. 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.

GILLILAND, A. J. Setting the stage. *In*: BACA, Murtha (Ed.). **Introduction to metadata**. Los Angeles, CA: Getty Publications, 2016. Disponível em: http://www.getty.edu/publications/intrometadata/setting-the-stage/. Acesso em: 30 jun. 2019.

GOMES, F. A. **Padronização de metadados na representação da informação em repositórios institucionais de universidades federais brasileiras**. Dissertação (Mestrado) - Programa de Pós-Graduação em Ciência da Informação, Universidade Federal da Bahia (UFBA), Salvador, 2015. Disponível em: https://repositorio.ufba.br/handle/ri/18950. Acesso em: 29 jul. 2022.

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.

HENNING, P. C. *et al*. The FAIRness of data management plans: an assessment of some European DMPs. **Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, v. 15, n. 3, 2021. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/2270/0. Acesso em: 06 abr. 2022.

HENNING, P. C. *et al*. Go fair e os princípios fair: o que representam para a expansão dos dados de pesquisa no âmbito da ciência aberta. **Em Questão**, v. 25, n. 2, p. 389-412, 2019b. Disponível em: https://brapci.inf.br/index.php/res/v/113770. Acesso em: 01 ago. 2022.

HEY, T.; TREFETHEN, A. E-Science and Its Implications. **Philosophical Transactions of the Royal Society of London**, v. 361, n. 1809, p. 1809–25, 2003. Disponível em: https://doi.org/10.1098/rsta.2003.1224. Acesso em: 20 abr. 2022.

HIGMAN, R.; BANGERT, D.; JONES, S. Three camps, one destination: the intersections of research data management, FAIR and Open. **Insights**, v. 32, n. 1, 2019. Disponível em: http://insights.uksg.org/articles/10.1629/uksg.468/. Acesso em: 22 ago. 2022.

HODSON, S. *et al*. **Turning FAIR data into reality**. Interim report of the European Commission Expert Group on FAIR data. 2018. Disponível em: https://zenodo.org/record/1285272#.YuL_sC35RhA. Acesso em: 30 abr. 2022.

JUTY, N. *et al.* Unique, persistent, resolvable: Identifiers as the foundation of FAIR. **Data Intelligence**, v. 2, n. 1-2, p. 30-39, 2020. Disponível em: https://direct.mit.edu/dint/article/2/1-2/30/9992/Unique-Persistent-Resolvable-Identifiers-as-the. Acesso em: 30 jun. 2022.

KALINAUSKAITĖ, D. **To be findable, accessible, interoperable and reusable:** language data and technology infrastructure for supporting the FAIR data approach. 2017. Disponível em: http://ceur-ws.org/Vol-1856/p04.pdf. Acesso em: 10 jun. 2022.

KOSCHTIAL, C. Understanding e-Science: What Is It About?. *In*: KOSCHTIAL, C.; KÖHLER, T.; FELDEN, C. (ed.). **E-Science: Open, Social and Virtual Technology for Research Collaboration**. Springer International Publishing, 2021. Disponível em: https://doi.org/10.1007/978-3-030-66262-2. Acesso em: 15 nov. 2022.

LÖFFLER, F. *et al*. Dataset search inbiodiversity research: Do metadata in data repositories reflect scholarlyinformation needs?. **PLOS ONE**, San Francisco, v. 16, n. 3, p. 1-36, 2021. Disponível

em:https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0246099.Acesso em: 20 jun. 2021.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. **Data on the Web best practices**. W3C Working Draft, World Wide Web Consortium (W3C), 2017. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em: 30 ago. 2022.

MACHADO, D. O. F. **Schema.org para representação nos catálogos digitais: a descoberta de recursos informacionais na Web**. 2022. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de São Carlos, São Carlos, 2022. Disponível em: https://repositorio.ufscar.br/handle/ufscar/16358.

MARÍN-ARRAIZA, P.; HEREDIA, A. FAIR PIDs: O papel da ORCID no fortalecimento dos Princípios FAIR. *In*: SALES, L. F.; VEIGA, V. dos S.; HENNING, P.; SAYÃO, L. F. (org.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. p. 23 -30. Disponível em: 10.22477/9786589167242.cap2. Acesso em: 20 jul. 2021.

MCQUILTON, P. *et al*. Helping the Consumers and Producers of Standards, Repositories and Policies to Enable FAIR Data. **Data Intelligence**, v. 2, n. 1-2, p. 151-157, 2020. Disponível em: https://direct.mit.edu/dint/article/2/1-2/151/9991/Helping-the-Consumers-and-Producers-of-Standards. Acesso em: 20 jul. 2021.

MONS, B.; SCHULTES, E.; LIU, F.; JACOBSEN, A. The FAIR Principles: First Generation Implementation Choices and Challenges. **Data Intelligence**, v. 2, n. 1–2, p. 1–9, 2020. Disponível em: https://direct.mit.edu/dint/article/2/1-2/1-9/10016. Acesso em: 27 jul. 2022.

MONTEIRO, E. C. S. A.; SANT'ANA, R. C. G.; PÉREZ, A. H. Direitos autorais de dados científicos no contexto da ciência aberta: estudo do repositório de dados do consórcio madroño. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 20, Florianópolis, 2019. **Anais** [...]. Florianópolis: Universidade Federal de Santa Catarina, 2019. Disponível em: https://brapci.inf.br/index.php/res/v/122765. Acesso em: 10 jan. 2022.

MUSTAFEE, N. *et al*. Co-citation Analysis of Literature in E-science and E-infrastructures. **Concurrency and Computation**: Practice and Experience, v. 32, n. 9, 2019. Disponível em: https://doi.org/10.1002/cpe.5620. Acesso em: 20 abr. 2022.

NATIONAL INFORMATION STANDARDS ORGANIZATION. **A Framework of Guidance for Building Good Digital Collections**. 3. ed. Baltimore: NISO, 2007. Disponível em: http://www.niso.org/publications/rp/framework3.pdf. Acesso em: 30 abr. 2015.

NEUMANN, J.; BRASE, J. DataCite and DOI names for research data. **Journal of Computer-Aided Molecular Design**, v. 28, n. 10, p. 1035-1041, 2014. Disponível em: https://link.springer.com/article/10.1007/s10822-014-9776-5#citeas. Acesso em: 20 jun. 2022.

OECD, O. FOR E. C. AND D.-. OECD. **Principles and Guidelines for Access to Research Data**. [*s.l*: [*s.n.*], 2007. Disponível em: http://www.oecd.org/sti/inno/38500813.pdf. Acesso em: 12 jan. 2021.

OUCHI, M. T.; SIMIONATO, A. C. Descrição de conjuntos de dados na web com schema.org. **Informação & Tecnologia**, v. 5, n. 1, p. 128-140, 2018. Disponível em: http://hdl.handle.net/20.500.11959/brapci/110403. Acesso em: 12 jan. 2023.

PAGANINE, L. N.; AMARO, B. Características dos repositórios de dados científicos no brasil. **BIBLOS** - Revista do Instituto de Ciências Humanas e da Informação, v. 34, n. 1, p. 176-188, 2020. Disponível em: https://brapci.inf.br/index.php/res/v/146046. Acesso em: 21 set. 2022.

PAMPEL, H. *et al*. Making Research Data Repositories Visible: The Re3data.org Registry. **PLoS ONE**, v. 8, n. 11, 2013. Disponível em: 10.1371/journal.pone.0078080. Acesso em: 15 jun. 2022.

PAVÃO, C. G. *et al*. Metadados e Repositórios Institucionais: uma relação indissociável para a qualidade da recuperação e visibilidade da informação. **PontodeAcesso**, [*S. l.*], v. 9, n. 3, p. 103–116, 2015. Disponível em: https://periodicos.ufba.br/index.php/revistaici/article/view/15163. Acesso em: 28 jul. 2022.

PIWOWAR, H. A.; VISION, T. J. Data reuse and the open data citation advantage. **PeerJ PrePrints**, p. 1-25, 2013. Disponível em: https://peerj.com/articles/175/. Acesso em: 10 fev. 2022.

POMERANTZ, J. **Metadata**. Cambridge: The MIT Press, 2015.

PRODANOV, C. C.; FREITAS, E. C. **Metodologia do Trabalho Científico:** métodos e técnicas da pesquisa e do trabalho acadêmico. Novo Hamburgo: Feevale, 2013.

REIS, M. J.; SENA, N. C. D. S. Biblioteconomia de dados e ciência de dados no contexto da e-science. **Revista Fontes Documentais**, v. 4, n. ed., p. 51-64, 2021. Disponível em: http://hdl.handle.net/20.500.11959/brapci/193856. Acesso em: 06 dez. 2022.

ROCHA, R. P. *et al*. **Acesso aberto a dados de pesquisa no Brasil: soluções tecnológicas: relatório 2018**. Porto Alegre, RS: UFRGS, 2018. Disponível em: http://hdl.handle.net/10183/185126. Acesso em: 19 out. 2020.

RODRIGUES, E. **A pandemia e a emergência da ciência aberta**. *In*: A Universidade do Minho em Tempos de Pandemia. UMinho Editora: Braga, Portugal, 2020. Disponível em: https://repositorium.sdum.uminho.pt/handle/1822/68390. Acesso em: 20 jul. 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.

SALES, L. F.; SAYÃO, L. F. Há futuro para as bibliotecas de pesquisa no ambiente de escience? **Informação & Tecnologia**, v. 2, n. 1, p. 30-52, 2015. Disponível em: http://hdl.handle.net/20.500.11959/brapci/41568. Acesso em: 20 jan. 2023.

SANCHEZ, F. A.; SILVA, N. B. P.; VECHIATO, F. L. Padrões de metadados para representação e organização da informação em repositórios de dados de pesquisa. **Informação & Tecnologia**, v. 5, n. 1, p. 37-51, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/110395. Acesso em: 01 fev. 2021.

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. Disponível em: https://periodicos.ufrn.br/informacao/article/view/12280. Acesso em: 6 jun. 2022.

SANTOS, P. L. V. A. da C.; SANT'ANA, R. C. G. Dado e Granularidade na perspectiva da Informação e Tecnologia: uma interpretação pela Ciência da Informação. **Ciência da Informação**, [*S.l.*], v. 42, n. 2, jan. 2013. Disponível em: http://revista.ibict.br/index.php/ciinf/article/view/228. Acesso em: 10 jan. 2022.

SAYÃO, L. F. Padrões para bibliotecas digitais abertas e interoperáveis. Encontros **Bibli:** Revista Eletrônica de Biblioteconomia e Ciência da Informação, n. esp. 1. sem., p. 18-47, 2007. Disponível em: https://brapci.inf.br/index.php/res/v/91544. Acesso em: 08 abr. 2022.

SAYÃO, L. F.; SALES, L. F. Afinal, o que é dado de pesquisa?. **BIBLOS** - Revista do Instituto de Ciências Humanas e da Informação, v. 34, n. contexto, 2020. Disponível em: http://hdl.handle.net/20.500.11959/brapci/162776. Acesso em: 22 jun. 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.

SAYÃO, L. F.; SALES, L. F. Dados abertos de pesquisa: ampliando o conceito de acesso livre. **RECIIS** - Revista Eletrônica de Comunicação, Informação e Inovação em Saúde, Rio de Janeiro, v. 8, n. 2, p. 76-92, 2014. Disponível em: https://www.arca.fiocruz.br/handle/icict/17102. Acesso em: 15 jun. 2022.

SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro: CNEN, 2015. Disponível em: http://carpedien.ien.gov.br:8080/handle/ien/1624. Acesso em: 15 jan. 2021.

SAYÃO, L. F.; SALES, L. F. Um modelo de implementação para a internet de dados & serviços FAIR. *In:* SALES, L. F.; VEIGA, V. S. de O.; HENNING, P. *et al*. (Eds.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021, p. 223–250. Disponível em: https://ridi.ibict.br/bitstream/123456789/1182/2/IBICT_Principios%20FAIR%20aplicados%20a%20gest%c3%a3o%20de%20dados%20de%20pesquisa_2021.pdf. Acesso em: 16 ago. 2022.

SILVA, F. C. C.; RODRIGUES, M. M. Implementação dos princípios FAIR em repositórios de dados científicos: uma análise comparativa das infraestruturas de software do DSpace e Dataverse. *In:* SALES, L. F.; VEIGA, V. S. de O.; HENNING, P. *et al*. (Eds.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021, p. 123–136. Disponível em: https://ridi.ibict.br/bitstream/123456789/1182/2/IBICT_Principios%20FAIR%20aplicados%20a%20gest%c3%a3o%20de%20dados%20de%20pesquisa_2021.pdf. Acesso em: 16 ago. 2022

SILVA, L. C.; SANTAREM SEGUNDO, J. E.; SILVA, M. F. Princípios fair e melhores práticas do linked data na publicação de dados de pesquisa. **Informação & Tecnologia**, v. 5, n. 2, p. 81-103, 2018. Disponível em: https://brapci.inf.br/index.php/res/v/120679. Acesso em: 07 abr. 2021.

SILVA, M. F.; MARTINS, D. L.; SIQUEIRA, J. Web semântica em repositórios: ontologia para representação de bibliotecas digitais. **Ciência da Informação em Revista**, Maceió, v. 6, n. 1, p. 99-113, 2019. Disponível em:

https://periodicos.ufpb.br/ojs/index.php/pbcib/article/view/49529. Acesso em: 10 jun. 2022.

SILVEIRA, L. *et al*. Ciência aberta na perspectiva de especialistas brasileiros: proposta de taxonomia. **Encontros Bibli**: Revista eletrônica De Biblioteconomia E Ciência Da informação, 26, 1-27, 2021. Disponível em: https://brapci.inf.br/index.php/res/v/160597. Acesso em: 10 abr. 2021.

SIMIONATO, A. C. Mapeamento dos metadados para dados científicos. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 18, 2017. **Anais** [...]. Marília, SP: Universidade Estadual Paulista em Franca, 2017. Disponível em: http://enancib.marilia.unesp.br/index.php/XVIII_ENANCIB/ENANCIB/paper/viewFile/563/874. Acesso em: 20 jan. 2021.

SOUSA, B. A. Proposta de criação de um repositório institucional para o instituto federal de educação, ciência e tecnologia da paraíba – ifpb. **Revista Brasileira de Biblioteconomia e Documentação**, v. 8, n. 1, p. 66-84, 2012. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5000. Acesso em: 26 jul. 2022.

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.

STRECKER, D. *et al*. **Metadata Schema for the Description of Research Data Repositories: version 3.1**. Re3data, 2021. Disponível em: https://gfzpublic.gfz-potsdam.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_5007395. Acesso em: 21 fev. 2022.

TARTAROTTI, R. C.; DAL'EVEDOVE, P. R.; FUJITA, M. S. L. Biblioteconomia de dados em repositórios de pesquisa: perspectivas para a atuação bibliotecária. **Informação & Informação,** v. 24, n. 3, p. 207, 2019. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/38732. Acesso em: 28 jul. 2022.

TORINO, E.; TREVISAN, G. L.; VIDOTTI, S. A. B. G. Dados abertos capes: um olhar à luz dos desafios para publicação de dados na web. **Ciência da Informação**, v. 48, n. 3, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/136339. Acesso em: 03 ago. 2022.

TORINO, E.; VIDOTTI, S. A. B. G. Boas práticas para dados na web: análise do portal Dados Abertos Capes. **Informação & Sociedade**: Estudos, v. 31, n. 1, p. 1–25, 2021. Disponível em:

https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/50790. Acesso em: 10 jul. 2022.

VEIGA, V. S. O. *et al*. Compartilhamento de dados de pesquisa na fiocruz: diagnóstico e percepção do pesquisador. **Ciência da Informação**, v. 48, n. 3, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/136386. Acesso em: 06 jun. 2022.

VIDOTTI, S. A. B. G; TORINO, E.; CONEGLIAN, C. S. #SejaJUSTOeCUIDADOSO: princípios FAIR e CARE na gestão de dados de pesquisa. *In:* SALES, L. F.; VEIGA, V. S. de O.; HENNING, P. *et al*. (Eds.). **Princípios FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021, p. 209–222. Disponível em: https://ridi.ibict.br/bitstream/123456789/1182/2/IBICT_Principios%20FAIR%20aplicados%20a%20gest%c3%a3o%20de%20dados%20de%20pesquisa_2021.pdf. Acesso em: 16 ago. 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.

ZENG, M. L.; QIN, J. **Metadata**. 2.ed. Chicago, IL: ALA Neal-Schuman, 2016.