Revista Científica da Universidade Eduardo Mondlane, Série: Letras e Ciências Sociais ISSN 2307-3918

Pecha Kucha

PRINCÍPIOS FAIR ALIADOS AO ACESSO E REUSO DE DADOS: análise de conjuntos de dados sobre Covid-19 presentes no Repositório PubChem

Tainá Regly[1,2], Viviane Santos de Oliveira Veiga[3] e Aline da Silva Alves[3]

[1]Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasil
Universidade Federal de Rio de Janeiro (UFRJ), Brasil
[3] Fundação Oswaldo Cruz (Fiocruz), Brasil

RESUMO: Este artigo objetiva verificar o alinhamento aos princípios FAIR no compartilhamento de dados de pesquisa em saúde e analisar seu potencial reuso na crise sanitária da Covid-19. Utiliza uma abordagem teórico-descritiva, bibliográfica e exploratória para estabelecer critérios de seleção do tema-piloto, da fonte de dados e da ferramenta de análise. Optou-se pelo tema piloto, um medicamento que foi amplamente debatido no campo científico e midiático, para o tratamento da Covid-19, a Cloroquina. Como fonte de análise utiliza o repositório disciplinar químico PubChem e o FairDataBR como ferramenta de análise para verificar o nível de aderência dos conjuntos de dados selecionados aos princípios FAIR de maneira semiautomatizada. Como resultado, aponta que os princípios Acessível e Reusável (notas 8.33 e 8.20) obtiveram as maiores notas, principalmente devido à facilidade de acesso, download e compartilhamento. Por outro lado, os princípios Encontrável e Interoperável (notas 5.20 e 7.75) apresentaram menor desempenho por não utilizarem identificadores persistentes em todos os conjuntos disponibilizados e não fazer uso de um esquema de metadados formalizado pela comunidade química. Conclui que, em relação aos princípios FAIR, existe um alinhamento parcial dos conjuntos de dados sobre Cloroquina depositados no PubChem. Desta forma, a análise mostra um potencial de reuso de (média das notas 7.37), pois todos os princípios apoiam o reuso. Por fim, reafirma que para que dados sejam FAIR, o ecossistema precisa ser FAIR, portanto, algumas medidas devem ser tomadas pelo PubChem para apoiar a geração de dados, como a designação ou a inserção obrigatória de identificadores persistentes.

Palavras-chave: Covid-19. FAIR, dados de pesquisa, PubChem.

FAIR PRINCIPLES ALLIED TO DATA ACCESS AND REUSE: analysis of Covid-19 datasets from PubChem repository

ABSTRACT: This article aims to verify alignment with FAIR principles in health research data sharing and analyze its potential reuse in the Covid-19 health crisis. It uses a theoretical-descriptive, bibliographic, and exploratory approach to establish criteria for selecting the pilot theme, the data source, and the analysis tool. The pilot theme was chosen, a drug that has been widely debated in the scientific and media field, for the treatment of Covid-19, Chloroquine. As a source of analysis it uses the chemical disciplinary repository PubChem and FairDataBR as an analysis tool to check the level of adherence of the selected datasets to the FAIR principles in a semi-automated way. As a result, it points out that the Accessible and Reusable principles (scores 8.33 and 8.20) obtained the highest scores, mainly due to ease of access, downloading and sharing. On the other hand, the Findable and Interoperable principles (scores 5.20 and 7.75) performed less well because they do not use persistent identifiers in all the available sets and do not make use of a metadata schema formalized by the chemical community. It concludes that, with respect to FAIR principles, there is partial alignment of the Chloroquine datasets deposited in PubChem. Thus, the analysis shows a reuse potential of (average of scores 7.37), as all principles support reuse. Finally, it reaffirms that for data to be FAIR, the ecosystem needs to be FAIR, so some measures should be taken by PubChem to support data generation, such as the designation or mandatory insertion of persistent identifiers.

Keywords: Covid-19, research data. FAIR, PubChem.

Correspondência para: (correspondence to:) taina.regly@gmail.com

INTRODUÇÃO

O movimento de Ciência Aberta vem se tornando, cada vez mais, aceito pela comunidade científica que, por sua vez, é uma grande incentivadora do compartilhamento de dados de pesquisa justamente pelo seu potencial de reprodutibilidade de estudos e aceleração das descobertas. Para além de uma grande concentração de esforços no estudo do vírus SARS-CoV-2, da doença Covid-19 e da investigação do seu tratamento, prevenção e vacinas, a pandemia resultou também na adoção de preceitos da Ciência Aberta, tal como a partilha de dados, a publicação de *preprints* e o acesso aberto às publicações.

Para que esses preceitos sejam colocados em prática é necessário que se faça uso de ambientes digitais que ofereçam recursos que garantam o acesso a longo prazo e de maneira aberta dos dados oriundos de pesquisas científicas. Assim, os repositórios de dados são um excelente ambiente que propicia esses benefícios, mais a visibilidade, compartilhamento e curadoria desses dados. O repositório PubChem foi selecionado para fazer parte deste trabalho por deter essas características importantes para a Ciência Aberta e devido à sua capacidade de ser uma importante fonte de informação mantida pelo National Institutes of Health (NIH) que proporciona acesso aos usuários que desejam ter acesso a dados químicos.

Apesar de a disponibilização dos dados ser uma questão essencial para a promoção da Ciência Aberta, existem outras características que os dados devem possuir para que possam ser compartilhados e reusados de maneira plena, tendo todo o seu potencial passível de utilização. Com isso, foram desenvolvidos os princípios FAIR, que consistem em quinze cânones que alicerçam e garantem a qualidade de conjuntos de dados para que esses sejam encontráveis, acessíveis, interoperáveis e reusáveis em pesquisas (WILKINSON *et al*., 2016).

A aplicação desses princípios tem como intuito servir como diretriz para que os dados de pesquisa ganhem contexto, sejam entendidos e possam ser reusados. Dessa forma, sua utilização impulsiona descobertas científicas mais velozes e fundamentadas, incluindo a leitura automatizada de máquinas tanto dos dados como dos metadados. Fato esse que contribui para uma melhor gestão e disseminação das descobertas científicas (SIMÕES; ANJOS e DIAS, 2021).

Tendo isso em vista, o objetivo desta proposta consiste em avaliar e verificar o nível de aderência de conjuntos de dados sobre o medicamento Cloroquina, oferecidos pelo PubChem, aos princípios FAIR.

METODOLOGIA

Utilizamos uma abordagem teórico-descritiva, bibliográfica e exploratória, uma vez que, com base em estudos e investigações teóricas, buscamos identificar conceitos e elementos para estabelecer critérios aptos a descrever características referentes aos conjuntos de dados selecionados.

Para viabilizar a análise desses conjuntos de dados, fizemos uso da ferramenta FairDataBR[1], desenvolvida por pesquisadores vinculados à Universidade Federal da Paraíba (PPGCI/MPGOA - UFPB). Esse instrumento possibilita, através da aplicação de um questionário, a avaliação e a verificação do nível de aderência de conjuntos de dados aos princípios FAIR de maneira automatizada, conferindo uma nota geral para a aplicação das especificações e uma para cada categoria FAIR: encontrável, acessível, interoperável e reutilizável.

Os conjuntos de dados selecionados e o resultado de sua análise com a ferramenta

FairDataBR podem ser encontrados publicados no repositório Zenodo, em Regly (2022).

RESULTADOS

No princípio Encontrável (*findable*), o conjunto de dados selecionado apresentou o menor alinhamento com os princípios FAIR e obteve nota 7.60. Creditamos esse desempenho ao fato de, apesar de fazer uso de identificadores persistentes, os recursos digitais não poderem ser encontrados através de mecanismos de busca na web. Outros pontos abordados na avaliação referem-se aos metadados utilizados que são estruturados e ao conjunto de dados que está publicado em um repositório específico de domínio.

O princípio Acessível (*accessible*) atingiu a maior nota de seus semelhantes, ficando com 8.33. Os dados são acessíveis ao público, inclusive através do identificador fornecido, podendo ser baixados sem a necessidade de cadastro. Outras características construtivas do "dataset" se referem ao uso de protocolos HTTP e ao fato de estarem acessíveis online sem a necessidade de intermediação de protocolos ou ferramentas especializadas e fazer uso de API de Web Service padrão. O único ponto limitativo deste princípio está atrelado ao aspecto de que o repositório não deixa claro se os metadados estarão acessíveis mesmo quando os dados não mais o estiverem.

Em relação à Interoperabilidade (*interoperable*) a nota alcançada foi 7.75. Tal nota foi obtida devido aos conjuntos de dados estarem disponíveis em formatos preferenciais. No caso do repositório, os formatos XML e JSON. Também foi identificado o uso de instrumentos de controle terminológico com identificador global, tal como ontologias e RDF voltados para a descrição do conteúdo e do contexto dos registros disponibilizados. Além disso, os conjuntos de dados estão vinculados a outros dados a partir de identificadores, o que torna as coleções de dados formadas pelos repositórios ainda mais enriquecidas de informações. Schymanski e Bolton (2021) contam que se as estruturas químicas são fornecidas de maneira a serem legíveis por máquinas, essas se tornam mais interoperáveis e, consequentemente, reutilizáveis. Uma vez que o potencial de interoperabilidade e localização de um conjunto de dados é aperfeiçoado, os obstáculos que impedem o reuso desses dados são comprimidos. O único ponto desfavorável em relação à interoperabilidade do repositório se refere ao fato de o esquema de metadados utilizado para descrever os dados não ser formalizado ou aprovado pela comunidade. O PubChem faz uso de um esquema de metadados próprio, o que prejudica a interoperabilidade entre diferentes sistemas e enfraquece o compartilhamento de dados.

Por fim, o princípio Reusável (*reusable*) obteve 8.20 como nota. O PubChem não exemplifica de maneira clara se os dados são licenciados, deixando para os responsáveis pelos dados a incumbência de revelar a licença preferida, que nem sempre é indicada. Nos conjuntos de dados analisados foram encontradas menções a licenças que, por sua vez, são abertas, o que favorece o reuso dos dados, apesar de estes não estarem de acordo com padrões de metadados para o domínio da Química elencados no FAIRsharing[2]. Os conjuntos de dados possuem proveniência detalhada através do redirecionamento para a página onde estão contidos os dados originais e, em alguns casos, a proveniência é acompanhada pelo seu histórico de versões.

DISCUSSÃO

A avaliação feita pela ferramenta FairDataBR acerca da aplicação dos princípios FAIR no conjunto de dados selecionado no repositório PubChem

alcançou como média a nota 7.97. Nessa aferição, o aspecto da acessibilidade obteve o melhor desempenho enquanto o da encontrabilidade o pior. Apesar da performance ter deixado a desejar nesse último tópico, consideramos que o "dataset" está, majoritariamente, alinhado aos princípios FAIR e seus dados estão aptos a serem compartilhados e reutilizados por pesquisadores.

O uso dos princípios FAIR na Química necessita que sejam fornecidos tanto padrões como orientações para que os pesquisadores se tornem competentes para tratar seus dados de modo que estes sejam interoperáveis e possam ser localizados. A partir dessas ações, são gerados benefícios, tais como o aumento da visibilidade das pesquisas científicas, melhoria na legibilidade dos dados pelas máquinas, auxílio aos autores para entenderem os requisitos de compartilhamento de dados relacionado ao financiamento de suas pesquisas e aumentar a visibilidade das informações produzidas (SCHYMANSKI e BOLTON, 2021)

Este estudo teve como intuito analisar o alinhamento dos conjuntos dados sobre o medicamento Cloroquina aos princípios FAIR. Como resultado da avaliação, os cânones Acessível e Reusável foram os que obtiveram as maiores notas, principalmente devido à facilidade de acesso, download e compartilhamento. Por outro lado, os princípios Encontrável e Interoperável apresentaram menor desempenho pelo fato de os dados não poderem ser encontrados através de mecanismos de busca na web e não fazer uso de um esquema de metadados formalizado pela comunidade química. O resultado obtido pelo pode ser aperfeiçoado caso determinadas modificações sejam implementadas no repositório, tal como a formalização do esquema de metadados utilizado, o uso de licenciamento nesses campos, a inclusão de identificadores em todos os dados e a garantia de que os metadados estarão disponíveis mesmo quando o conjunto de dados não estiver mais acessível.

REFERÊNCIAS

SCHYMANSKI, E. L., & EVAN E. B. "FAIR chemical structures in the Journal of Cheminformatics." Journal of cheminformatics v.13, n50, 2021, p.1-3. https://doi.org/10.1186/s13321-021-00520-4

SIMÕES, R. C.; ANJOS, R. L.dos & Dias, G. A. (). Análise dos conjuntos de dados disponíveis no repositório COVID-19 Data Sharing/BR à luz dos princípios FAIR. In Princípios SALES, L.F. et al. (Org). **FAIR aplicados à gestão de dados de pesquisa**. Rio de Janeiro: Ibict, 2021. pp. 91–102.. https://doi.org/10.22477/9786589167242.cap7

REGLY, T. Dados provenientes da análise do repositório PubChem com a ferramenta FairDataBR [Data set]. Zenodo. 2022. https://doi.org/10.5281/zenodo.7071895

WILKINSON, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v.3, n.160018, 2016. https://doi.org/10.1038/sdata.2016.18

NOTAS

[1] Disponível em: https://wrco.ufpb.br/fair/index.html. Acesso em: 22 mar. 2022.

[2] FAIRsharing. Disponível em: https://fairsharing.org/. Acesso em: 15 abr. 2022.