Pecha Kucha

DOI: 10.21680/2447-7842.2023v9n2ID33832

Padrões de metadados de proveniência considerados relevantes pelos pesquisadores no reuso dos dados em COVID-19

Provenance metadata standards considered relevant by investigators in the reuse of COVID-19 data

Anderson Silva de Araujo [1]

Viviane Santos de Oliveira Veiga [2]

Carlos Henrique Marconde [3]

Submetido em: 17/04/2023	Aprovado na ConfOA: 14/06/2023	Publicado em: 04/12/2023

Resumo: O artigo investiga a percepção de pesquisadores sobre metadados de proveniência necessários para o reúso de dados de pesquisa em COVID-19. A metodologia envolveu uma revisão de fontes bibliográficas e documentais, nacionais e internacionais e a aplicação de um instrumento de coleta de dados. Os resultados revelaram que a maioria dos repositórios que armazenam dados de pesquisa em COVID-19 opta por não adotar os padrões de metadados amplamente reconhecidos pela comunidade científica. Os participantes destacaram a importância de informações claras nos metadados, como tema ou palavras-chave, informações

[1] Graduação em Teologia - Seminário Teológico Batista do Sul do Brasil, graduação em Letras - Português e Grego pela Universidade do Estado do Rio de Janeiro e graduação em Biblioteconomia pela Universidade Federal do Estado do Rio de Janeiro. Especialização em Informação científica e tecnológica em saúde (2019). Mestrado em Informação e Comunicação em Saúde (PPGICS-Fiocruz).
[2] Doutora em Ciências - área de concentração: Informação e Comunicação em Saúde pelo Programa de Pós-Graduação em Informação e Comunicação em Saúde/Fiocruz , com período sanduíche na Universidade de Coimbra. Mestre em Ciências - área de concentração: Gestão da Informação e Comunicação em Saúde pela Escola Nacional de Saúde Pública Sérgio Arouca-Fiocruz. Bacharel em Biblioteconomia e Documentação pela Universidade Federal do Estado do Rio de Janeiro.
[3] Graduação em Arquitetura e Urbanismo pela Universidade Federal Fluminense, mestrado em Ciência da Informação pela Universidade Federal do Rio de Janeiro e doutorado em Ciência da Informação pela Universidade Federal do Rio de Janeiro.

sobre licenças de uso, detalhes sobre os coletores de dados e estratégias de preservação em longo prazo. Os metadados de proveniência atualmente em uso não oferecem robustez necessária para garantir o reúso eficaz dos dados de pesquisa em COVID-19, cruciais para evitar ou responder a outra emergência sanitária.

Palavras-chave: metadados de proveniência; COVID-19; dados de pesquisa.

Abstract: The article investigates researchers' perception of provenance metadata necessary for the reuse of COVID-19 research data. The methodology involved a review of national and international bibliographic and documentary sources and the application of a data collection instrument. The results revealed that most repositories that store COVID-19 research data choose not to adopt metadata standards widely recognized by the scientific community. Participants highlighted the importance of clear information in metadata, such as theme or keywords, information on usage licenses, details about data collectors, and long-term preservation strategies. The provenance metadata currently in use does not offer the necessary robustness to ensure effective reuse of COVID-19 research data, which is important for avoid or responde another emergency health.

Keywords: provenance metadata; COVID-19; research data.

1 INTRODUÇÃO

A proveniência de dados pode ser considerada um requisito importante para estabelecer confiabilidade e prover segurança em sistemas computacionais de informação, facilitando dessa forma o reúso dos dados (Freund; Sembay & Macedo, 2019). Um padrão de metadados serve para especificar regras de como os dados devem ser criados ou incluídos (por exemplo, como identificar o título principal), regras de representação para conteúdo (por exemplo, padrões de representação do tempo) e valores de conteúdo admissíveis (isto é, se os termos devem ser tomados

a partir de um vocabulário controlado específico ou podem ser providos pelo autor, derivados do texto, ou aditados pelo trabalho de criadores de metadados sem uma lista de termos controlados). Pode haver ainda regras de sintaxe para a codificação dos elementos e seu conteúdo (Chan & Zeng, 2006; National Information Standards Organization, 2004). Os princípios FAIR, um acrônimo para Findable, Accessible, Interoperable e Reusable são norteadores na gestão de dados de metadados. (Henning *et al*., 2018).

Com o surto do novo coronavírus, várias organizações uniram esforços para impedir o avanço da pandemia e entender o desenvolvimento e as implicações da doença, de forma que o trabalho conjunto e contínuo permitisse uma reação mais rápida e coordenada (OPAS & OMS, 2020). Apesar da necessidade de políticas baseadas em evidências e tomada de decisões médicas, não há padrão internacional ou sistema coordenado para coletar, documentar e disseminar dados e metadados relacionados à COVID-19. Desta forma, seu uso e reutilização para análise epidemiológica oportuna é uma questão desafiadora devido a problemas com documentação, interoperabilidade, integridade, heterogeneidade metodológica e qualidade dos dados. O presente estudo tem como objetivo identificar quais são os metadados de proveniência que seriam fundamentais para o reúso de dados no contexto da pesquisa em COVID-19, segundo a percepção de pesquisadores.

2 METODOLOGIA

Foi elaborado instrumento de coleta de dados, composto pelo Termo de Consentimento Livre e Esclarecido (TCLE) e pelo questionário[4]. Foi aplicado no período de 03 de janeiro de 2023 a 08 de fevereiro de 2023, aos pesquisadores que receberam fomento para pesquisa da Fiocruz nos editais sobre a COVID-19[5],[6]. O questionário foi aplicado por plataforma de correio eletrônico, para os 135

[4] O questionário pode ser acessado neste link: https://forms.gle/g4Db2i3Jb6TUX3TK7.
[5] Disponível em: https://portal.fiocruz.br/edital-ideias-e-produtos-inovadores-covid-19-encomendas-estrategicas
[6] Disponível em: https://portal.fiocruz.br/edital-geracao-de-conhecimento-enfrentamento-da-pandemia-e-pos-pandemia-covid-19-encomendas

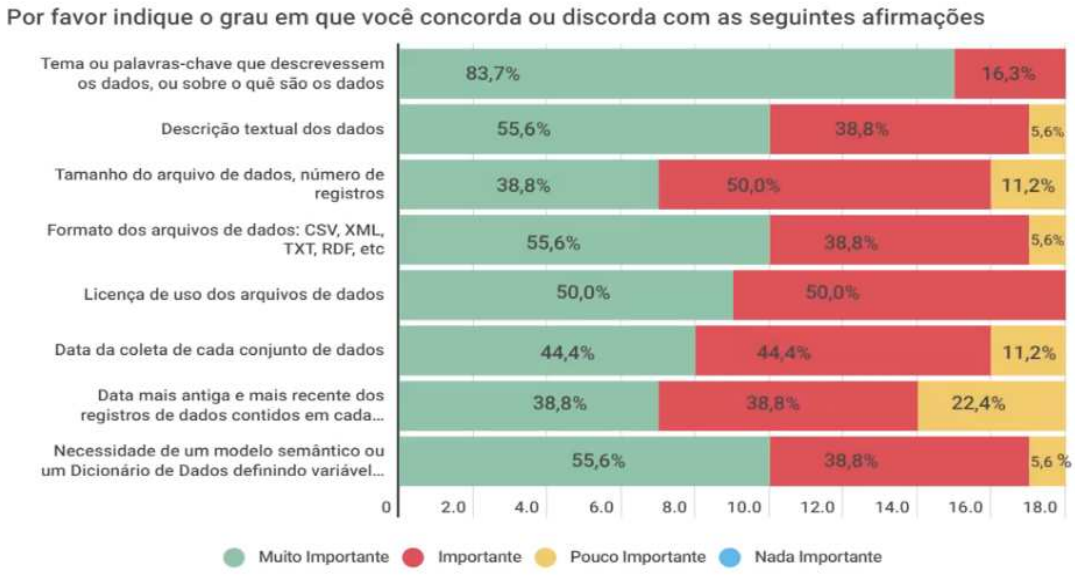
coordenadores identificados no levantamento. Foram recebidas 18 respostas, cerca de 23,4% do total de pesquisas financiadas.

3 RESULTADOS E ANÁLISES

A informação sobre os dados é importante, conforme estabelecido pelos princípios FAIR, para que os dados sejam localizáveis, acessíveis, interoperáveis e reutilizáveis. Nessa categoria foram formuladas oito questões para identificar os quesitos que os respondentes consideram como mais importantes sobre a informação sobre os dados; e que consideram mais relevantes ou importantes para o compartilhamento e para o reúso. Neste artigo são apresentados os resultados obtidos através de três questões do formulário (4, 5 e 6).

O gráfico 1 a seguir apresenta quais metadados ou informação os respondentes consideram **muito importante**, **importante**, **pouco importante** ou **nada importante** para viabilizar o reúso dos dados de pesquisa em COVID-19.

Gráfico 1 – Informações sobre os dados (Questão 4)

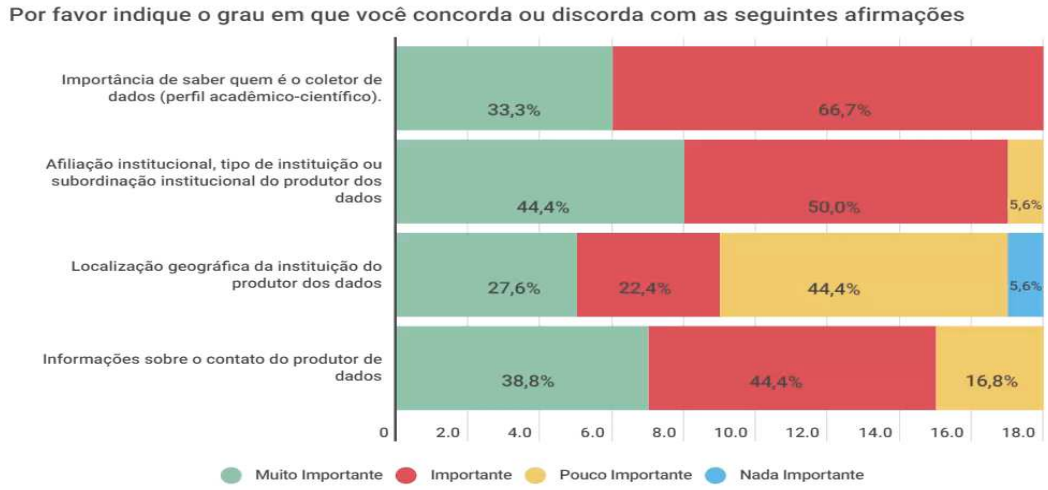


Fonte: Dados da pesquisa (2023).

Verificou-se que 100% dos respondentes indicaram ser **muito importante** e importante a *descrição do tema ou palavras-chave* que descrevem os dados ou sobre o quê são os dados. Lembrando que a descrição temática é um dos fatores que compõe o princípio *Findable*, do FAIR. Quanto à importância da *licença de uso dos arquivos de dados*, 100,0% dos respondentes consideram **muito importante** e **importante.** A informação referente à data da coleta de cada conjunto de dados foi vista em sua maioria como **importante**, porém alguns respondentes não o consideram um fator importante para o reúso dos dados de pesquisa.

A informação sobre o produtor de dados é importante, para isso, os metadados devem ter a sua proveniência detalhada. Deve-se saber de onde os dados vieram, esclarecer a origem, afiliação, localização geográfica etc. Este fluxo de informação deve ser descrito em um formato legível por mecanismos automatizados. O gráfico 2 a seguir apresenta quanto às informações sobre o produtor de dados que os respondentes consideram **importante** para viabilizar o reúso dos dados de pesquisa.

Gráfico 2 – Informação sobre o produtor de Dados (Questão 5)

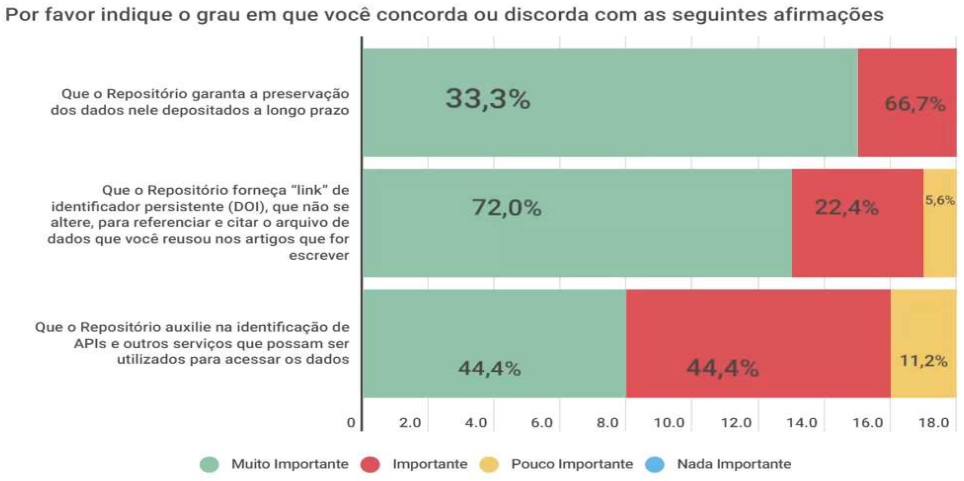


Fonte: Dados da pesquisa (2023).

Quanto à importância de saber *quem é o coletor de dados* (perfil acadêmico-científico), 100% dos respondentes consideram como **importante** e **muito importante**. Em termos quantitativos, a informação sobre a localização geográfica dividiu a opinião dos respondentes com metade considerando como **importante** e a outra metade como **não sendo importante**. Esta questão foi a única que recebeu a resposta de **nada importante** por parte de um respondente. Quanto à importância de saber informações sobre o *contato do produtor de dados*, 44,4% dos respondentes consideram como **importante**, 38,8% dos respondentes como **muito importante**. Esta questão apoia o reúso dos dados para uma averiguação direta com o produtor dos dados, em caso de dúvida e necessidade de mais informação.

A informação sobre os repositórios onde os dados estão disponibilizados é importante para uso e reúso de dados e precisa ser otimizada consonante os princípios FAIR (Wilkinson *et al.*, 2016). O gráfico 3 a seguir apresenta quais informações, sobre o repositório onde os dados estão disponibilizados, os respondentes consideram muito importante, importante, pouco importante ou nada importante, para viabilizar o reúso dos dados de pesquisa em COVID-19.

Gráfico 3 - Informações sobre o Repositório onde os dados estão disponibilizados (Questão 6)



Fonte: Dados da pesquisa (2023).

Quanto à questão da importância de que o *repositório garanta a preservação dos dados nele depositado*, a longo prazo, 33,3% dos respondentes consideram como **muito importante** e 66,7% dos respondentes consideram como **importante**. O princípio FAIR da acessibilidade (*Acessible*) requer que os dados estejam disponíveis por um longo tempo. Os conjuntos de dados tendem a degradar-se ou a desaparecer completamente tornando um desafio a preservação. Este princípio diz que os dados devem estar em um formato que permita o acesso ao longo dos anos.

Quanto à questão sobre a importância dos *repositórios de fornecer um "link" de identificadores persistente (DOI)* para referenciar e citar o arquivo de dados que foram reusados, 72,0% dos respondentes consideram como **muito importante**, 22,4% dos respondentes consideram **importante**. Para atender o princípio FAIR de dados localizáveis (*Findable*) deve ser atribuído aos dados, dentre outros quesitos, um identificador globalmente exclusivo e persistente. Quanto à questão que avalia a importância de que *o repositório auxilia na identificação de APIs* (Application Programming Interface, ou interface de programação de aplicações) *e de outros serviços para identificar os dados*, 44,4% dos respondentes consideram como **muito importante**, 44,4% dos respondentes consideram como **importante**. O uso de APIs permite a comunicação com outros dados facilitando o princípio FAIR da interoperabilidade.

4 CONSIDERAÇÕES

Verificou-se nesta pesquisa a necessidade de aprimoramento na escolha dos esquemas de metadados na descrição dos dados para apoiar o reúso dos dados de pesquisa em COVID-19. É necessário alinhar os metadados de proveniência dos dados ao domínio e às necessidades e percepções dos pesquisadores, para gerar conjuntos de dados de pesquisa com alto potencial de reúso, principalmente em um cenário de emergência sanitária. Um fator que os pesquisadores consideraram como mais relevante é a parte da descrição do tema ou palavras-chave que descrevem os dados ou sobre o quê são os dados. Esses dados levam a compreender a importância do uso das ontologias e dos vocabulários controlados para a

comunidade disciplinar, para a descrição temática do conjunto de dados, otimizando assim o reúso dos dados de pesquisa em COVID- 19 e facilitando a acessibilidade e a reutilização desses dados cruciais para a pesquisa em saúde pública e, assim, salvar vidas.

REFERÊNCIAS

Chan, L. M. & Zeng, M. L. (2006). Metadata interoperability and standardization: a study of methodology part i: achieving interoperability at the schema level. *D-Lib Magazine*, 12(6). http://www.dlib.org/dlib/june06/chan/06chan.html.

Freund, G. P.; Sembay, M. J. & Macedo, D. D. J. de (2019). Proveniência de Dados e Segurança da Informação: relações interdisciplinares no domínio da Ciência da Informação. *Revista Ibero-Americana De Ciência Da Informação*, *12*(3), 807–825. https://doi.org/10.26512/rici.v12.n3.2019.21203.

Henning, P. C. *et al.* (2018). Desmistificando os princípios fair: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados fair. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 19. http://hdl.handle.net/20.500.11959/brapci/103243.

National Information Standard Organization. (2004) *Understanding Metadata*. Bethesda, MD: NISO Press. https://www.niso.org/publications/understanding-metadata-2017.

OPAS & OMS. (2020). *OMS declara emergência de saúde pública de importância internacional por surto de novo coronavírus.* https://www.paho.org/pt/news/30-1-2020-who-declares-public-health-emergency-novel-coronavirus.

Wilkinson, M. D. *et al*. (2016). Os Princípios Orientadores FAIR para gestão e administração de dados científicos. *Scientific Data*, 3, 160018. https://doi.org/10.1038/sdata.2016.18.