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

Comunicação

METADADOS PARA REPRESENTAÇÃO DE DADOS EM COVID-19: um estudo exploratório

Anderson Silva de Araujo[1], Viviane Santos de Oliveira Veiga[1], Isabella Henrique Lima Pereira[1] e Mylena Cristhina Araujo de Oliveira[2,3]

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

RESUMO: Este artigo objetiva apresentar os padrões de metadados utilizados pelos repositórios de dados de pesquisa que disponibilizam conjuntos de dados em Covid-19 em acesso aberto. Adota uma abordagem descritiva e exploratória, por meio de levantamentos bibliográficos e documental. Utiliza como fontes de informação as bases *Web of Science e PubMed, e o diretório* RE3DATA. Identifica os principais padrões de metadados utilizados para representar dados em Covid-19 encontrados na literatura e os padrões e os esquemas de metadados utilizados nos repositórios de dados com conjuntos de dados em COVID-19. Verificou-se que os repositórios registrados no Re3data com conjuntos de dados em Covid-19 não utilizam os padrões de metadados específicos do campo da saúde identificados na literatura, o que pode comprometer o alinhamento destes dados aos princípios FAIR, principalmente no quesito interoperabilidade semântica. A maioria destes repositórios possui um esquema de metadados próprio. Recomenda-se um estudo mais aprofundado sobre os esquemas de metadados elaborados por estes repositórios e seu alinhamento com os princípios FAIR e as demandas do campo.

Palavras-chave: Covid-19, dados de pesquisa, padrões de metadados, repositórios digitais.

METADATA FOR DATA REPRESENTATION IN COVID-19: an exploratory study

ABSTRACT: This article aims to present the metadata standards used by research data repositories that make Covid-19 datasets available in open access. It adopts a descriptive and exploratory approach, through bibliographical and documentary surveys. It uses the Web of Science and PubMed databases and the RE3DATA directory as sources of information. Identifies the main metadata patterns used to represent COVID-19 data found in the literature and the metadata patterns and schemas used in data repositories with COVID-19 datasets. It was verified that the repositories registered in Re3data with Covid-19 datasets do not use the specific metadata standards of the health field identified in the literature, which may compromise the alignment of these data with the FAIR principles, mainly in terms of semantic interoperability. Most of these repositories have their own metadata schema. Further study on the metadata schemes elaborated by these repositories and their alignment with FAIR principles and field demands is recommended.

Keywords: Covid-19, research data, metadata standards. digital repositories.

Correspondência para: (correspondence to:) moranderson0182@gmail.com

INTRODUÇÃO

Com o surto do novo coronavírus várias organizações uniram esforços para impedir o avanço da pandemia e entender o desenvolvimento e implicações da doença, de forma que o trabalho conjunto e contínuo permitisse uma reação mais rápida e coordenada. A fidedignidade, acesso e potencial reuso de dados em Covid e relacionados aos diversos aspectos da Covid-19 tornaram-se fundamentais. Neste contexto, a transformação destes dados em dados FAIR motivou a criação da Rede VODAN (Virus Outbreak Data Network

https://www.go-fair.org/implementation-networks/overview/vodan/) e no Brasil, a Rede VODAN, https://portal.fiocruz.br/en/vodan-brazilBR.

Os princípios FAIR (*Findable, Accessible, Interoperable e Reusable*) foram desenvolvidos para orientar as boas práticas na pesquisa científica, de modo a facilitar a localização, o acesso, a interoperabilidade e o reuso de dados de pesquisa. Para que conjuntos de dados sejam FAIR, dados e metadados precisam estar alinhados a estes princípios. Desta forma, metadados e padrões de metadados são importantes para garantir que as plataformas digitais disponibilizem dados encontráveis, acessíveis, interoperáveis e reutilizáveis.

Metadados são conceituados como dados sobre dados (CAMPOS, 2007). A função dos metadados é garantir a padronização dos recursos informacionais, pautados em esquemas e regras internacionais na tentativa de facilitar a identificação, a busca, a localização, a recuperação, a preservação, o uso e o reuso.

Esquema de metadados é uma lista de propriedades principais de metadados escolhidas para uma identificação consistente de um recurso para fins de citação e recuperação, juntamente com instruções de uso recomendado (DATACITE METADATA WORKING GROUP, 2017).

Os bibliotecários produzem e padronizam metadados há séculos, desde as primeiras tentativas de organização da informação a partir da descrição de documentos. Profissionais de diversas áreas estão buscando criar instrumentos de descrição da informação, mas seu desconhecimento dos métodos, processos e peculiaridades característicos do campo da documentação, da Biblioteconomia, e da Ciência da Informação, tem gerado uma variedade de padrões que muitas vezes não atendem satisfatoriamente às exigências de uma lógica descritiva estabelecida, e que dê conta da complexidade da caracterização desse material e que atenda às necessidades informacionais atuais. (MILSTEAD; FELDMAN, 1999; ALVES, 2005; CASTRO; SANTOS, 2007 & CASTRO, 2008).

No contexto da saúde, faz-se necessário o uso estratégico das tecnologias disponíveis, atrelado à adoção de metadados e padrões de metadados de domínio especializado. Como a área é extremamente dinâmica, a adoção de determinados modelos e padrões podem garantir um alto índice de revocação, bem como a recuperação da informação mais bem estruturada e efetiva.

Desta forma, este artigo objetiva apresentar os padrões de metadados utilizados pelos repositórios de dados de pesquisa que disponibilizam, em acesso aberto, conjuntos de dados em Covid-19.

Os dados oriundos de pesquisas podem ser armazenados, preservados e acessados, contribuindo para o reuso e a reprodutibilidade do conhecimento científico. Diversos financiadores e editores científicos têm determinado que os dados de pesquisa devem ser acessíveis tanto para a comunidade científica, quanto para a sociedade em geral.

PADRÕES DE METADADOS PARA DADOS DE PESQUISA EM COVID-19

Para conhecer os metadados utilizados para representar dados de pesquisa em Covid19 foi aplicada uma abordagem descritiva e exploratória. Foram realizados levantamentos bibliográficos e documental. Para o levantamento bibliográfico recorreu-se às bases *Web of Science e PubMed,* nas quais foram aplicados os seguintes termos: covid; interoperability; covid-19; interoperable; sars vírus; metadata standards; sars-cov-2; metadata; sars; medical records; coronavírus; coronavirus disease.

Ao analisar literatura levantada sobre metadados e covid-19, identificou-se a presença de 1 (um) padrão de metadados em

artigo indexado na Web of Science e 3 (três) iniciativas de padrões de metadados em artigos indexados noPubMed. Na *Web of Science o metadado foi encontrado no artigo "COVID-19 pandemic reveals the peril of ignoring metadata standards". O* Genomic Standards Consortium (GSC, www.gensc.org) que foi fundado há 15 anos por cientistas que observaram que os dados de sequência do genoma, ainda uma novidade na época, raramente tinham os metadados mais básicos prontamente disponíveis em um formato estruturado.

As primeiras listas de verificação elaboradas pelo GSC se concentraram em orientar os cientistas a adicionar as informações mínimas necessárias para permitir a reutilização de seus dados em estudos futuros.

Na *PubMed* foram encontradas três iniciativas o Outbreak.info no artigo "Outbreak.info Research Library: A standardized, searchable platform to discover and explore COVID-19 resources and data", o PHA4GE no artigo "GA4GH: Políticas e padrões internacionais para compartilhamento de dados em pesquisa genômica e saúde" , e o GISAID no artigo "Interoperable medical data: The missing link for understanding COVID-19".

O Outbreak.info é um projeto dos laboratórios Su, Wu e Andersen da Scripps Research para unificar a epidemiologia e dados genômicos de COVID-19 e SARS-CoV-2, pesquisas publicadas e outros recursos.

Pesquisadores, autoridades de saúde e o público podem rastrear no Outbreak.info a pandemia usando dados sobre casos, mortes e variantes genômicas e manter-se atualizados sobre pesquisas relacionadas por meio de visualizações interativas, uma biblioteca pesquisável e dados brutos para download.

O PHA4GE identificou a necessidade de uma especificação de dados contextuais SARS-CoV-2 de código aberto e adequada à finalidade que possa ser usada para estruturar informações consistentemente como parte de boas práticas de gerenciamento de dados e para compartilhamento de dados com parceiros confiáveis e/ou repositórios públicos.

A especificação foi desenvolvida por consenso entre especialistas do domínio e incorporou os padrões existentes da comunidade com ênfase nas necessidades de saúde pública do SARS-CoV-2, garantindo a privacidade, maximizando o conteúdo das informações e a interoperabilidade entre conjuntos de dados e bancos de dados para melhor permitir análises para combater o COVID-19. O pacote de especificações também contém vários materiais de acompanhamento, como procedimentos operacionais padrão, ferramentas, um guia de referência e protocolos de envio de repositório (protocols.io) para ajudar a colocar o padrão em prática.

E o GISAID que tem como objetivo facilitar o compartilhamento de sequências do genoma viral e metadados clínicos e epidemiológicos relacionados para ajudar os pesquisadores a entender como os vírus evoluem e se espalham durante epidemias e pandemias.

Após esta etapa foi realizado um mapeamento, no Re3Data, dos repositórios de dados com conjuntos de dados de pesquisa em COVID-19. O Re3data, é um diretório global de repositórios de dados de pesquisa que abrange repositórios de dados de pesquisa de diferentes disciplinas acadêmicas. É mantido financeiramente pela Fundação Alemã de Pesquisa e coordenado por instituições científicas e acadêmicas na Alemanha.

Nesta ferramenta, identificou-se os repositórios de dados com conjuntos de dados em COVID-19, utilizando a seguinte estratégia: utilização do termo "data repository", selecionamos o filtro ***Data access***, no qual optamos pelo campo ***Open*** que nos deu o resultado 843[IL1] [VSDOV2]. Em seguida, utilizamos o

campo das **palavras chave** (**Covid-19**) e obtivemos um resultado final de 18 repositórios.

O idioma adotado pelos repositórios é majoritariamente inglês, sendo 15 (quinze) unicamente de língua inglesa, 1 (um) em inglês, mas que permite a tradução para outras línguas, 1(um) em alemão, e 1 (um) em Português (Figura 1).

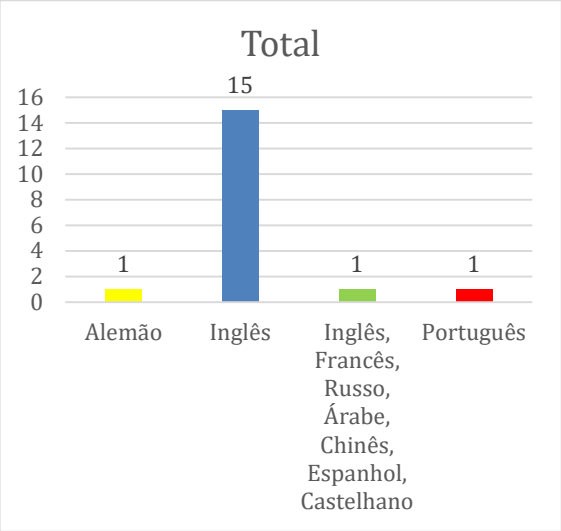

FIGURA 1: Idioma dos repositórios

Na Figura 2, a maioria dos repositórios pertence aos Estados Unidos que possui um total de 8 repositórios no qual 5 pertencem unicamente ao país e 3 são fruto de parcerias com outros países. O Brasil aparece com um repositório.

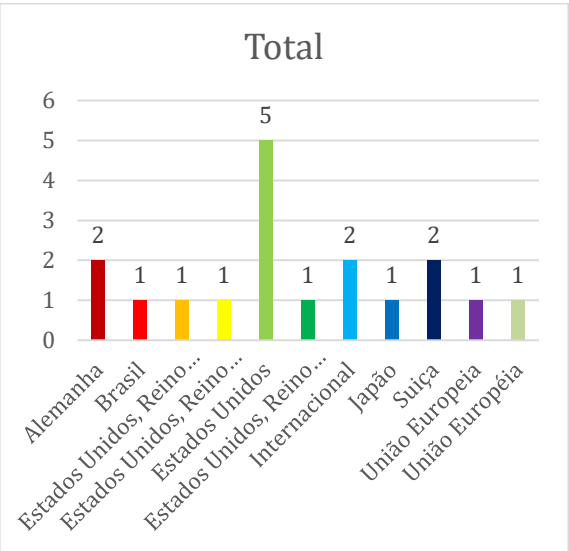

FIGURA 2: Países dos repositórios

Para entendermos melhor o tipo de pesquisa disponibilizada nos repositórios, bem como comparar o tipo de metadados utilizados, foi criada uma nuvem de palavras com os assuntos indicados nos repositórios. Com base na Figura 3, explicitada abaixo, as palavras mais recorrentes foram: Ciência da vida aparece 175 vezes, pesquisa biológica 7 e ciência sociais 5.

FIGURA 3: Nuvem de palavras

Em relação ao padrão de metadados explicitado na figura 4, dos 18 Repositórios, 8 (oito) indicaram que utilizam esquemas de metadados desenvolvidos pelo próprio repositório (Repository-Developed Metadata Schema), 2 (dois) o Dublin Core 1 (um) informou o RDF Data Cube Vocabulary. 7 (sete) repositórios não informaram o padrão de metadados adotado.

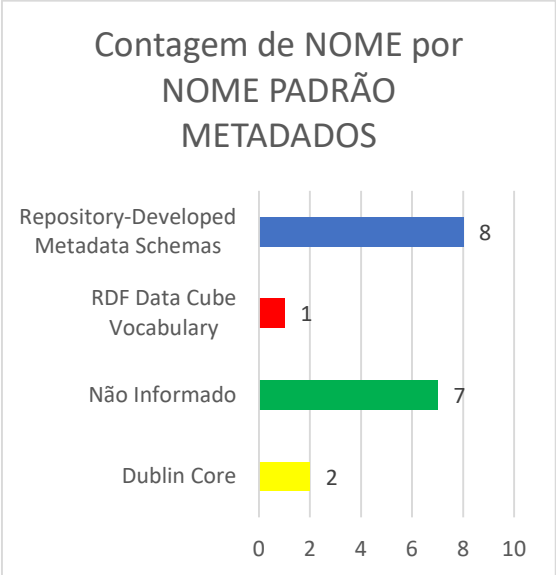

FIGURA 4: Padrão de Metadados

O RDF Data Cube Vocabulary fornece um meio de publicar dados multidimensionais, como estatísticas, na web de forma que possam ser vinculados a conjuntos de dados e conceitos relacionados usando o padrão W3C RDF (Resource Description Framework).

O Dublin Core é um esquema de metadados que visa descrever objetos digitais, tais como, vídeos, sons, imagens, textos e sites na web. Aplicações de Dublin Core utilizam XML e o RDF (Resource Description Framework).

A Dublin Core Metadata Initiative (DCMI) (Iniciativa Dublin Core Metadados) é uma organização dedicada a promover a adoção de padrões de interoperabilidade de metadados e desenvolver vocabulários especializados para descrever fontes e recursos da Web para que os sistemas de busca e recuperação de informações sejam mais rápidos e flexíveis.

Como apresentado anteriormente, a maioria dos repositórios analisados identificou a necessidade de desenvolver metadados próprios para representar os conjuntos de dados em seus repositórios. Isto aponta a necessidade da criação de padrões de metadados, analisados e validados pela comunidade de domínio e da comunidade científica geral, para garantir a interoperabilidade no campo científico.

CONSIDERAÇÕES FINAIS

Verificou-se que os repositórios do campo da saúde, principalmente aqueles cadastrados no Re3data na categoria Ciências da Vida foram os que mais armazenaram dados de pesquisa em COVID-19. Os Estados Unidos, foi identificado como o país com o maior número de repositórios que disponibilizam dados de pesquisa em covid-19. Quanto ao padrão de metadados verificamos que a maioria destes repositórios possui um esquema de metadados próprio, o que pode significar que os esquemas atuais não estão atendendo as demandas de representação descritiva e temática dos dados de pesquisa.

Também constatou-se que os repositórios analisados não adotaram os padrões de metadados específicos do campo da saúde, o que pode por um lado comprometer o alinhamento destes dados aos princípios FAIR, e por outro talvez favorecer a interoperabilidade com outras disciplinas.

Por fim, constatamos que é necessário um estudo mais aprofundado para verificar a contribuição dos novos esquemas de metadados criados pelos repositórios para a descrição dos conjuntos de dados em Covid-19, e o grau de FAIR destes esquemas, visto que um subprincípio do FAIR recomenda o uso de padrões da comunidade, isto é, do domínio científico.

REFERÊNCIAS

ALVES, R. C. V. ***Web Semântica**: uma análise focada no uso de metadados*. 2005. 180 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia Ciências, Universidade Estadual Paulista - UNESP, Marília, 2005.

CAMPOS, L. F. B. Metadados digitais: revisão bibliográfica da evolução e tendências por meio de categorias funcionais. ***Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação***, v.12, n. 23, p. 16-46, 2007.

CASTRO, F. F. **Padrões de representação e descrição de recursos informacionais em bibliotecas digitais na perspectiva da ciência da informação**: uma abordagem do MarcOntinitiative na era da web semântica. 2008. 201 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista - UNESP, Marília, 2008.

CASTRO, F. F.; SANTOS, P. L. V. A. C. Os metadados como instrumentos tecnológicos na padronização e potencialização dos recursos informacionais no âmbito das bibliotecas digitais na era da web semântica. **Informação & Sociedade**, v. 17, n. 2, p.

13-19, maio/ago. 2007.

DATACITE METADATA WORKING GROUP. **DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1. DataCite e.V. 10.5438/0014**. 2017. https://schema.datacite.org/meta/kernel-4.1/doc/DataCite-MetadataKernel_v4.1.pdf

MILSTEAD, J.; FELDMAN, S. ***Metadata**: cataloging by any other name*. [S. l.: s. n.], 1999.