INSTITUTO DE COMUNICAÇÃO E INFORMAÇÃO CIENTÍFICA E TECNOLÓGICA EM SAÚDE
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMAÇÃO E COMUNICAÇÃO EM SAÚDE

Padrões de metadados de proveniência para reuso de dados de pesquisa em Covid 19 alinhados aos princípios FAIR

Anderson Silva de Araujo

Rio de Janeiro
2023

FUNDAÇÃO OSWALDO CRUZ

INSTITUTO DE COMUNICAÇÃO E INFORMAÇÃO CIENTÍFICA E TECNOLÓGICA EM SAÚDE
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMAÇÃO E COMUNICAÇÃO EM SAÚDE

Anderson Silva de Araujo

Padrões de metadados de proveniência para reuso de dados de pesquisa em Covid 19 alinhados aos princípios FAIR

Dissertação de Mestrado apresentado ao Programa de Pós-Graduação *Stricto Sensu* do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, como requisito parcial para obtenção do grau de Mestre em Informação e Comunicação em Saúde, na área de concentração Configurações e Dinâmicas da Informação e Comunicação em Saúde.

Orientadora: Prof. Dra. Viviane Santos de Oliveira Veiga
Coorientador: Prof. Dr. Carlos Henrique Marcondes.

Rio de Janeiro
2023

Araujo, Anderson Silva de.

Padrões de metadados de proveniência para reuso de dados de pesquisa em Covid 19 alinhados aos princípios FAIR / Anderson Silva de Araujo. - Rio de Janeiro, 2023.
111 f.

Dissertação (Mestrado) – Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Pós-Graduação em Informação e Comunicação em Saúde, 2023.

Orientadora: Viviane Santos de Oliveira Veiga.
Coorientador: Carlos Henrique Marcondes.

Bibliografia: f. 81-87

1. Metadados. 2. Reuso de dados de pesquisa. 3. Metadados de Proveniência. 4. COVID-19. 5. Dados de pesquisa científica. 6. Princípios FAIR I. Título.

Elaborado pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca de Manguinhos/Icict/Fiocruz com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Igor Falce Dias de Lima - CRB-7/6930.

ANDERSON SILVA DE ARAUJO

Padrões de metadados de proveniência para reuso de dados de pesquisa em Covid 19

Aprovado em: _____

Banca examinadora:

_____

Drª. Viviane Santos de Oliveira Veiga – Orientadora
PPGICS/ICICT/Fiocruz

_____

Dr. Carlos Henrique Marcondes – Coorientador
PPGCI da Universidade Federal Fluminense

_____

Drª. Cícera Henrique da Silva – Titular Interno
PPGICS/ICICT/Fiocruz

_____

Dr. Claudio Jose Silva Ribeiro – Titular Externo
Professor UNIRIO - Universidade Federal do Estado do Rio de Janeiro

Dedico este trabalho

À minha esposa e aos meus filhos pelo amor, carinho e incentivo na minha caminhada.

AGRADECIMENTOS

Agradeço a Deus por me sustentar, me guiar, me amar e ter enviado o seu filho Jesus Cristo para me salvar.

À minha família pela compreensão e companheirismo, em especial, à minha mãe por me acolher na fase final da dissertação.

Aos meus orientadores, Viviane Veiga e Carlos Marcondes, pela dedicação, paciência e tempo despendido em muitas horas de zoom e Meet. Sou grato pelo compartilhar de conhecimento e exemplo de profissionalismo.

Aos professores e membros da banca, Cícera Henrique, Cláudio Jose, Rosane Abdala e Naira Silveira que aceitaram ao convite para compor a banca de defesa me deixando muito honrado.

À minha esposa por se companheira na minha caminhada acadêmica e mesmo nas dificuldades permanecer ao meu lado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e Vice-presidência de Educação, Informação e Comunicação da Fiocruz (VPEIC) pela bolsa concedida para que eu pudesse me dedicar integralmente à pesquisa.

Pois dele, por ele e para ele são todas as coisas. A ele seja a glória para sempre! Amém.

Romanos 11:36

Padrões de metadados de proveniência para reuso de dados de pesquisa em Covid 19 alinhados aos princípios FAIR

RESUMO

O presente estudo teve como objetivo identificar as percepções dos pesquisadores quanto aos metadados de proveniência necessários para o reuso de dados de pesquisa de COVID-19. A metodologia consistiu na pesquisa bibliográfica em fontes de informação nacionais e internacionais que forneceu as bases teóricas da pesquisa e foi realizado um mapeamento das diretrizes internacionais para padrão de metadados para dados de pesquisa nas iniciativas internacionais Fairsharing (https://fairsharing.org/), Digital Curation Centre (DCC) (https://www.dcc.ac.uk/) e Research Data Alliance (RDA) (https://www.rd-alliance.org/), permitindo identificar os padrões de metadados utilizados na pesquisa em COVID-19 e a mapear as principais diretrizes apontadas na literatura para o tema. Foi elaborado instrumento de coleta de dados (questionário online) para verificar quais os padrões de metadados de proveniência apoiam os pesquisadores no reuso dos dados em COVID-19. Foram analisados os dados bibliográficos, as pesquisas dos repositórios, a percepção dos pesquisadores obtidos através do instrumento de coleta para identificar os padrões de metadados de proveniência que podem ser utilizados nas pesquisas sobre COVID-19. Verificou-se que realmente só temos padrões para determinada área ou campo de conhecimento e não diretrizes internacionais gerais. Constatou-se que que há um consenso por parte dos respondentes nas questões da importância da informação sobre tema ou palavra-chave; licença de uso, coletor de dados e preservação dos dados a longo prazo por parte dos repositórios, quanto aos metadados de proveniência necessários para o reuso dos dados de pesquisa em Covid 19.

Palavras-chaves: Padrões de metadados. Dados de pesquisa. COVID-19. Princípios FAIR.

ABSTRACT

The present study aimed to identify researchers' perceptions of the provenance metadata required for the reuse of COVID-19 research data. The methodology consisted of bibliographical research in national and international information sources that provided the theoretical bases of the research and a mapping of international guidelines for metadata standards for research data in international Fairsharing initiatives was carried out (https://fairsharing.org/) , Digital Curation Center (DCC) (https://www.dcc.ac.uk/) and Research Data Alliance (RDA) (https://www.rd-alliance.org/), allowing to identify the metadata standards used in research on COVID-19 and to map the main guidelines indicated in the literature for the subject. A data collection instrument (online questionnaire) was developed to verify which metadata patterns of provenance support researchers in the reuse of data on COVID-19. Bibliographic data, repositories searches, researchers' perception obtained through the collection instrument were analyzed to identify provenance metadata patterns that can be used in research on COVID-19. It was verified that we really only have standards for a certain area or field of knowledge and not general international guidelines. It was found that there is a consensus on the part of the respondents on the issues of the importance of information on a topic or keyword; use license, data collector and long-term preservation of data by the repositories, regarding the provenance metadata necessary for the reuse of research data in Covid 19.

Keywords: Metadata standards. Search data. COVID-19. FAIR Principles.

LISTA DE TABELAS

Tabela 1 – Perfil dos pesquisadores.....68
Tabela 2 – Laboratório, Departamento ou unidade.....69

LISTA DE ILUSTRAÇÕES

Quadro 1 – Síntese da metodologia.....43
Figura 1 **-** Padrão de Metadados utilizados pelos repositórios.....46
Figura 2 – Campo de pesquisa Fairsharing.....47
Figura 3 – Policies Fairsharing.....48
Figura 4 – Filtros Fairsharing.....49
Figura 5 – Pesquisa sem filtro Fairsharing.....50
Quadro 2 – Resultado Fairsharing.....51-52
Figura 6 - Guidance DCC.....52
Figura 7- Metadados Disciplinares/DCC.....52
Figura 8 - Metadados Disciplinares/Disciplinas DCC.....53
Figura 9 – Pesquisa por Disciplina DCC.....53
Figura 10 – Dados gerais de pesquisa.....54
Quadro 3 – Padrões de metadados/Disciplinas DCC.....54-57
Quadro 4 - Padrão de metadados gerais/DCC.....58-60
Quadro 5 - Grupos padrões de metadados RDA.....62-63
Gráfico 1 - Brasil/Estado/Laboratórios Fiocruz.....70
Gráfico 2 – Informações sobre os dados.....72
Gráfico 3 – Informação sobre o produtor de Dados.....74
Gráfico 4 - Informações sobre o repositório onde estão os dados disponibilizados..... 75
Quadro 6 – Informações Relevantes.....76
Quadro 7 - Síntese/Questões e princípios FAIR.....78

LISTA DE SIGLAS E ABREVIATURAS

ABCD *Access to Biological Collection Data*

API *Application Programming Interface*

BAV *Biblioteca Apstolica Vaticana*

BDTD *Biblioteca Digital Brasileira de Teses e Dissertações*

BRAPCI *Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação*

BSLIS *School of Library and Information Science*

CDTS *Centro de Desenvolvimento Tecnológico em Saúde*

CERIF *Common European Research Information Format*

ConfOA *Conferência Lusófona de Ciência Aberta*

CRF *Case Report Form*

CSV *Comma Separated Values*

DCAT *Data Catalog Vocabulary*

DCC *Digital Curation Centre*

DeCS *Descritores em Ciências da Saúde*

DFG *Fundação Alemã de Pesquisa*

DIF *Directory Interchange Format*

DIHS *Departamento de Direitos humanos, Saúde e Diversidade*

DOI *Digital Object Identifier*

ENSP *Escola Nacional de Saúde Pública*

EOSC *European Open Science Cloud*

FAIR *Findable, Accessible, Interoperable e Reusable*

FAPESP *Fundação de Amparo à Pesquisa do Estado de São Paulo*

Fiocruz *Fundação Oswaldo Cruz*

IBICT *Instituto Brasileiro de Informação em Ciência e Tecnologia*

ICICT *Instituto de Comunicação e Informação Científica e Tecnológica em Saúde*

INI *Instituto Nacional de Infectologia*

IOC *Instituto Oswaldo Cruz*

ITA *Instituto Tecnológico de Aeronáutica*

KIT *Karlsruhe Institute of Technology*

LACES *Laboratório de Comunicação e Saúde*

LGPD *Lei Geral de Proteção de Dados Pessoais*

LIS *Library and Information Services*

LIS *Laboratório de Informação em Saúde*

MEDLINE *Medical Literature Analysis and Retrieval System Online*

MeSH *Medical Subject Headings*

MIBBI *Minimum Information for Biological and Biomedical*

MIG *Metadata Interest Group*

MRC *UK Medical Research Council*

MSDWG *Metadata Standards Directory Working Group*

NIH *National Institutes of Health*

NSB *National Science Board*

NSF *National Science Foundation*

O&M *Observations and Measurements*

OAI-ORE - *Open Archives Initiative Object Reuse and Exchange*

OAI-PMH *Open Archives Initiative Protocol for Metadata Harvesting*

OECD *Organisation for Economic Co-operation and Development*

OMS *Organização Mundial da Saúde*

OpenAIRE *Open Access Infrastructure for Research in Europe*

OSC *Office of Strategic Coordination*

PDBx/mmCIF *Protein Data Bank Exchange Dictionary*

PGD *Plano de Gestão de Dados*

PMC *PubMed Central*

PPGICS *Programa de Pós-Graduação em Informação e Comunicação em Saúde*

RC-UEM *Revista Científica da Universidade Eduardo Mondlane*

RDA *Research Data Alliance*

RDPIG *Research Data Provenance group*

Re3data.org *Registry of Research Data Repositories*

SDMX *Statistical Data and Metadata Exchange*

TCLE *Termo de Consentimento Livre e Esclarecido*

TI *Tecnologia da Informação*

TICs *Tecnologias de Informação e Comunicação*

UFABC *Universidade Federal do ABC*

UFSCar *Universidade Federal de São Carlos*

Unesp *Universidade Estadual Paulista*

Unicamp *Universidade Estadual de Campinas*

Unifesp Universidade *Federal de São Paulo*

USP *Universidade de São Paulo*

VODAN *Virus Outbreak Data Network*

W3C World *Wide Web Consortium*

WDC *World Data Center*

WHO *World Health Organization*

Sumário

1 INTRODUÇÃO 16

1.1 OBJETIVOS 19

1.2 JUSTIFICATIVA 19

2 REFERENCIAL TEÓRICO: Conceito e definições 23

2.1 DADOS DE PESQUISA 23

2.2 GESTÃO, COMPARTILHAMENTO E ABERTURA DE DADOS. 29

2.3 METADADOS 33

3 PROCEDIMENTOS METODOLÓGICOS 39

4 APRESENTAÇÃO DOS RESULTADOS 44

4.1 ARTIGO - METADADOS PARA REPRESENTAÇÃO DE DADOS EM COVID-19: UM ESTUDO EXPLORATÓRIO. 44

4.2 DIRETRIZES INTERNACIONAIS PARA PADRÕES DE METADADOS PARA DADOS DE PESQUISA. 46

4.2.1 Fairsharing 47

4.2.2 DCC 51

4.2.3 RDA 61

4.3 ANÁLISE DO QUESTIONÁRIO 65

4.3.1 Coleta de dados e análise dos dados 65

5 CONSIDERAÇÕES FINAIS 79

REFERÊNCIAS 81

APÊNDICE A – Plano de Gestão de Dados 88

APÊNDICE B – ARTIGO 96

APÊNDICE C – Questionário 106

1 INTRODUÇÃO

O presente trabalho tem como tema os Padrões de metadados de proveniência para reuso de dados de pesquisa em COVID-19 (Do inglês Coronavirus Disease 2019 Denominação da doença causada pelo novo coronavírus SARS-CoV-2) alinhados aos princípios FAIR (Findable, Acessible, Interopable, Reusable).

A escolha do tema foi motivada pela mobilização tanto nacional como internacional para o enfrentamento da COVID-19. Destacam-se os desafios apresentados tanto pelo Governo quanto pela comunidade científica no acesso e uso de dados no tema. Agências de saúde do mundo todo se esforçaram para produzir pesquisas e encontrar soluções rápidas no enfrentamento à crise sanitária.

A premissa era a coleta, o compartilhamento e a integração dos dados. “A pandemia da COVID-19 desafiou sistemas de saúde e pesquisas em todo o mundo. Os dados são coletados em todo o mundo e precisam ser integrados e disponibilizados a outros pesquisadores rapidamente”. (QUERALT-ROSINACH et al, 2022, p.1, tradução nossa[1]).

Esses dados, se compartilhados, podem apoiar na rapidez da resposta da ciência no tratamento de uma crise sanitária e na minimização das suas consequências na sociedade. Por isto, é de suma importância o tratamento dos dados em saúde, a fim de que as informações possam ser facilmente acessadas e recuperadas pelos diversos tipos de usuários.

A Organização Mundial da Saúde (OMS) em 30 de janeiro de 2020, declarou o surto do coronavírus ou novo coronavírus, como emergência de saúde pública de interesse internacional e em 11 de março de 2020, como pandemia (WHO, 2020).

Agências de saúde de todo o mundo produzem diagnósticos e tratamentos dos casos, com isso gerando dados importantes para o estudo da doença.

Como esses dados podem ser usados, compartilhados e reusados? Como esses dados podem apoiar os pesquisadores a formular novos tipos de indagações, hipóteses no estudo de questões cruciais para a ciência e para a sociedade?

[1] No original: “The COVID-19 pandemic has challenged healthcare systems and research worldwide. Data is collected all over the world and needs to be integrated and made available to other researchers quickly” (QUERALT-ROSINACH et al, 2022, p. 1).

Dados compartilhados de forma adequada têm um grande potencial para responder questões científicas e proporcionarem uma resposta adequada para a sociedade.

Dagliati *et al* (2021, p.1, tradução nossa[2]) afirmam que “a pandemia de COVID-19 mostrou claramente que os principais desafios e ameaças para a humanidade precisam ser abordados com respostas globais e decisões compartilhadas”.

Diversas instituições, as agências de fomento e pesquisadores da área da ciência da informação compreendem que o compartilhamento de dados de pesquisa pode constituir uma fonte riquíssima de recursos para a pesquisa científica (NASSI-CALÒ, 2010), para isso é indispensável dispor de ambientes informacionais, de plataformas, de sistemas, que assegurem o armazenamento dos dados, a fim de disponibilizar informações seguras e confiáveis aos diversos tipos de usuários (especialistas e não especialistas).

Nesse contexto os padrões ou esquemas de metadados são indispensáveis. São eles que determinarão a descrição, a representação, o acesso, a confiabilidade e a persistência do recurso/objeto digital no ambiente informacional, além de definir a interoperabilidade entre sistemas.

É necessária uma organização de uma coleta e transmissão desses dados para que sejam acessíveis e até reutilizados em outras pesquisas. Um exemplo disso é o projeto VODAN. A rede VODAN - Virus Outbreak Data Network foi criada para implantar uma infraestrutura de nós federados para a captura e uso de dados e publicações, seguindo os princípios FAIR (WILKINSON *et al*., 2016) para gestão de dados de pesquisa.

A Rede tem como proposta a organização de todo um sistema de dados e procedimentos de coleta para que eles possam ser reusados. O padrão de metadados se torna um componente dessa infraestrutura que apoia o reuso dos dados, com o objetivo de colocar esses recursos disponíveis tanto para uso humano quanto para sua exploração por agentes de software, estimulando reuso e reprodutibilidade na área científica.

Cabe aos profissionais envolvidos com esse tipo de informação proporcionar a curadoria dos dados de pesquisa para que eles possam ser reusados em larga escala.

[2] No original: “The coronavirus disease 2019 (COVID-19) pandemic has clearly shown that major challenges and threats for humankind need to be addressed with global answers and shared decisions” (DAGLIATI, et al, 2021, p. 1)

O reuso de dados de pesquisa é um conceito que envolve a utilização de ferramentas tecnológicas para o compartilhamento de dados de forma livre e sem restrições. (MARTINS; PERLIN, 2020)

Pasquetto, Randles e Borgman (2017) definem o reuso de dados como:

> O problema mais fundamental na compreensão da reutilização de dados é distinguir entre “uso” e “reutilização”. Na situação mais simples, os dados são coletados por um indivíduo, para um projeto de pesquisa específico, e o primeiro “uso” desse indivíduo é fazer uma pergunta de pesquisa específica. Se esse mesmo indivíduo retornar ao mesmo conjunto de dados posteriormente, seja para o mesmo projeto ou para um projeto posterior, isso geralmente seria considerado um “uso”. Quando esse conjunto de dados é contribuído para um repositório, recuperado por outra pessoa e implantado para outro projeto, geralmente seria considerado uma “reutilização”. Na linguagem comum das práticas de dados, a reutilização geralmente implica o uso de um conjunto de dados por alguém que não seja o responsável pela coleta (PASQUETTO; RANDLES; BORGMAN, 2017, p.6, tradução nossa[3]).

Instituições acadêmicas, financiadores da pesquisa em saúde e os principais periódicos da área enfatizaram a importância do compartilhamento, acesso e reuso de dados de pesquisa, de modo a maximizar o conhecimento e os seus benefícios para a saúde (WELLCOME TRUST, 2012).

Nesse contexto o pesquisador é ator central para realizar a adequada gestão dos dados de pesquisa, pois há diversos benefícios ao disponibilizarem suas bases de dados em repositórios de dados para a utilização livre por parte de outros pesquisadores.

Um dos objetivos principais do reuso é usar dados gerados por outros pesquisadores, reduzindo o tempo (sem necessidade de nova coleta) e custos na etapa de coleta dos dados.

O compartilhamento de dados favorece a reprodutibilidade da ciência, o aumento de citações da pesquisa original, a melhora na gestão de dados por parte da comunidade acadêmica e, talvez mais relevante no contexto econômico atual, a

[3] No original: “The most fundamental problem in understanding data reuse is to distinguish between a “use” and a “reuse.” In the simplest situation, data are collected by one individual, for a specific research project, and the first “use” is by that individual to ask a specific research question. If that same individual returns to that same dataset later, whether for the same or a later project, that usually would be considered a “use.” When that dataset is contributed to a repository, retrieved by someone else, and deployed for another project, it usually would be considered a “reuse.” In the common parlance of data practices, reuse usually implies the usage of a dataset by someone other than the originator” (PASQUETTO; RANDLES; BORGMAN, 2017, p. 6).

diminuição dos custos associados à condução de uma pesquisa (MARTINS; PERLIN, 2020).

Garantir que os dados de pesquisa, juntamente com suas descobertas publicadas, sejam amplamente disponibilizados para a comunidade de pesquisa leva a mais descobertas e maior eficácia.

Esta situação suscita as seguintes questões. Que informações são importantes para que o pesquisador reutilize os dados? O que faz o pesquisador considerar se esses dados estão bem descritos ou não para reuso? Qual a importância de se conhecer quais os metadados de proveniência que são fundamentais para garantir que os pesquisadores venham a reutilizar esses dados em pesquisa sobre COVID-19?

Responder essas perguntas é fundamental. Diante do exposto esse projeto busca responder a seguinte questão: Quais os metadados de proveniência necessários para garantir o reuso dos dados de pesquisa?

1.1 OBJETIVOS

Identificar metadados de proveniência fundamentais para o reuso de dados no contexto da pesquisa em COVID-19 alinhados aos princípios FAIR. Para tanto, foram definidos os seguintes objetivos específicos:

1- Identificar os padrões de metadados mais utilizados para descrição de conjuntos de dados de pesquisa de COVID-19.

2- Mapear as diretrizes internacionais para padrões de metadados para dados de pesquisa alinhados aos princípios FAIR.

3- Verificar que metadados de proveniência são considerados relevantes pelos pesquisadores para viabilizar seu reuso em dados de pesquisa em COVID-19.

1.2 JUSTIFICATIVA

Os dois principais produtos da ciência, reconhecidos atualmente na maioria dos domínios científicos, são os artigos científicos e os dados de pesquisa.

A comunicação cientifica tem sido um dos temas mais estudado pela Ciência da Informação (COSTA, 2005). Recentemente os dados digitais têm recebido grande

atenção devido a sua quantidade crescente (“Big Data”), ao custo de gerá-los (“eScience”) e ao seu potencial de uso.

A pesquisa é importante pois os dados têm sido considerados a moeda mais valiosa da ciência (DAVIS; VICKERY, 2007). A ampla disponibilidade e acessibilidade dos dados são itens fundamentais da agenda da ciência aberta.

Dados sobre a epidemia de COVID-19 vêm sendo coletados de forma descentralizada em todo o mundo. A OMS propôs uma padronização destes dados através de sua coleta como o CRF (Case Report Form). Estes dados, coletados a nível mundial, são importantes pois podem revelar novas ideias e caminhos que ajudem a combater a pandemia (WHO, 2021).

Em busca da ampla disponibilidade de dados de pesquisa e do cumprimento das novas diretivas governamentais, uma série de agências de financiamento, editores de periódicos científicos, instituições acadêmicas e organizações de pesquisa iniciaram uma chamada intensiva para o compartilhamento de dados nos últimos anos.

A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) reconhece a importância da gestão adequada dos dados de pesquisa como parte essencial das boas práticas de pesquisa. Para tanto, considera necessário que os dados resultantes de projetos financiados pela Fundação sejam gerenciados e compartilhados de forma a garantir o maior benefício possível para o avanço científico, tecnológico, socioeconômico e cultural. Além de racionalizar recursos, a gestão apropriada de tais dados facilita a reprodutibilidade da pesquisa e permite promover novas pesquisas, graças à possibilidade de reuso e compartilhamento (www.fapesp.br/gestaodedados).

Também a National Science Foundation (NSF), uma das principais agências federais de fomento à pesquisa estadunidense, em 2011, implementou sua política para compartilhamento de dados (data sharing mandates) (CURTY, 2018)

À medida que novos repositórios de dados são criados para abrigar os dados da pesquisa, e mais dados se acumulam em seus servidores, a atenção se desloca para encontrar maneiras de sustentar o valor destes resultados da investigação e maximizar a sua reutilização, uma vez que os benefícios do compartilhamento de dados só podem ser mensurados por meio de seu reuso efetivo.

Neste sentido, a sustentabilidade do ciclo de vida da ciência aberta depende de condições propícias para a maximização do reuso de dados de pesquisa, ao invés de meramente acumulá-los em repositórios de dados.

Tempo, dinheiro e economia de esforço são frequentemente reconhecidos como principais motivadores para o reuso de dados de pesquisa, não adianta compartilhar dados se esses dados não forem reutilizados (HYMAN, 1972; KIECOLT; NATHAN, 1985; CASTELO, 2003; LAW, 2005)

A pandemia da COVID-19 levou a pesquisadores do mundo todo a se empenharam para produzir estudos emergenciais para o tratamento dos dados em saúde.

É imprescindível assegurar que os dados armazenados sejam disponibilizados de forma segura e confiável para os diversos tipos de usuários, sejam especialistas e não especialistas, para isso, é fundamental que os metadados associados a estes dados possam ser acessados tanto por máquinas como por humanos.

A interoperabilidade é essencial para que os dados não sejam perdidos. Estudar as iniciativas internacionais e nacionais que tratam da adoção de metadados e padrões de metadados nos ambientes, sistemas e plataformas que operam com conteúdos digitais para dados e informações referente ao COVID-19 se torna de suma importância.

Os pesquisadores e instituições de pesquisa devem estar atentos, dentre outros aspectos, às questões que envolvem ao compartilhamento dos dados de pesquisa. Os dados precisam ser coletados, organizados, preservados e compartilhados permitindo com isso a possibilidade do reuso dos dados de pesquisa.

Os metadados (que de modo simplista podem ser definidos como dados sobre dados) precisam estar alinhados a princípios que permitam a localização, a acessibilidade, a interoperabilidade e o reuso dos dados.

Como respostas a estas questões foram criados os princípios FAIR, um acrônimo para Findable, Accessible, Interoperable e Reusable (localizáveis, acessíveis, interoperáveis e reusáveis), criados para proporcionar o grau máximo de reuso de dados de pesquisa, a partir da adoção de padrões, metadados, vocabulários controlados, ontologias e identificadores persistentes que proporcionam significado preciso aos dados e aos demais objetos a eles vinculados (VEIGA *et al*, 2019).

Para o reuso dos conjuntos de dados de pesquisa eles precisam estar de acordo com os princípios FAIR, justificando assim a importância do uso de metadados

padronizados para o acesso em geral aos dados usados nas pesquisas de COVID-19.

É importante que se compartilhe os dados segundo os princípios FAIR para que eles possam ser reutilizados. O que é importante para que o pesquisador reutilize um dado? Quem vai dizer se um dado está bem descrito ou não para reuso é o pesquisador. Quais são os metadados de proveniência fundamentais para garantir que os pesquisadores consigam reutilizar dados de pesquisa em COVID-19.

No Brasil o projeto VODAN-BR faz parte dessa rede e tem como objetivo coletar dados de pacientes com COVID-19 padronizados pelo CRF a partir de hospitais, formatá-los segundo os princípios FAIR e disponibilizá-los para reuso. Neste contexto a confiabilidade e a proveniência destes dados são importantes.

Yoon e Lee ao examinarem quantitativamente fatores de confiança no reaproveitamento de dados a partir das perspectivas dos usuários constataram que:

> O fator Produtor de Dados (H1) e a Qualidade dos Dados (H3) foram significativos, conforme previsto, enquanto a Comunidade Acadêmica (H3) e o Intermediário de Dados (H4) não foram significativamente relacionados com a confiança dos usuários em dados (YOON; LEE, 2019, p. 1245, tradução nossa[4]).

A proposta deste estudo é identificar metadados de proveniência fundamentais para o reuso de dados no contexto da pesquisa em COVID-19 alinhados aos princípios FAIR.

[4] No original: “This study found that the Data Producer (H1) and Data Quality (H3) were significant, as predicted, while Scholarly Community (H3) and Data Intermediary (H4) were not significantly related to reusers’ trust in data (YOON; LEE, 2019, p.1245)”.

2 REFERENCIAL TEÓRICO: Conceito e definições

2.1 DADOS DE PESQUISA

Os dados estão presentes no nosso dia a dia. Coletamos dados sobre qualquer coisa, desde em relação ao tempo, à economia, aos esportes, à política etc.

Porém, muitas das vezes uma pergunta fica no ar: o que são dados? E por que dados de pesquisa são importantes?

Ao falarmos sobre dados de pesquisa se torna imprescindível definirmos o que são dados.

O avanço da tecnologia digital trouxe grandes transformações desde a disponibilização do resultado da pesquisa (periódico eletrônico), quanto ao processo de produção e organização da ciência.

As Tecnologias de Informação e Comunicação (TICs) proporcionaram a criação de redes de colaboração e de compartilhamento entre pesquisadores em diferentes partes do mundo, a chamada pesquisa eletrônica, ou e-Science. (VEIGA, 2017).

Hey e Hey (2006) definem a e-Science como uma “ciência em rede e baseada em dados”, que traz desafios tanto para os cientistas quanto para bibliotecários que precisarão de habilidades e tecnologias para gerenciar, pesquisar e fazer a curadoria desses novos recursos de dados, para a realização de pesquisa colaborativa.

Jonh Taylor foi o primeiro a utilizar o termo, em 2001, ele percebeu que muitas áreas da ciência poderiam se beneficiar de uma infraestrutura comum de tecnologia da informação (TI). Ele articulou uma visão para esse tipo de ciência distribuída e colaborativa e introduziu o termo "e-Science” (TAYLOR, 2001 apud HEY; HEY, 2006).

Importante ter em mente que a e-Science não significa uma nova disciplina científica, mas é a abreviatura para o conjunto de ferramentas e tecnologias necessárias para apoiar a ciência colaborativa e em rede (HEY; HEY, 2006).

Cunha (2010) apontou a e-Science como um fator de relevância para o surgimento de novo acervo ligado aos dados de pesquisa (científicos).

Segundo Hjørland (2018, tradução nossa[5]) o termo dado ao ser usado em diversos campos “pode aparecer muitas vezes em termos compostos como banco de

[5] No original: “Data is a much-used concept in many fields, including LIS, in particular in composite terms such as database, data archive, data mining, descriptive data, metadata, linked data and now big data.” (HJØRLAND, 2018)

dados, arquivo de dados, mineração de dados, dados descritivos, metadados, dados vinculados e agora big data".

A falta de um entendimento e de uma conceituação do que venha a ser dado pode causar confusão para muitos. A etimologia da palavra vem do latim "*datum"* (latim, particípio passado de dare, 'dar') o que para Jensen (1950, p. ix, tradução nossa[6]) "é uma infelicidade, pois o termo que definiria melhor seria caftum (latim, particípio passado de cafere^ 'tomar')", esse sim para ele, deveria ter passado a simbolizar o fenômeno-unidade na ciência, "pois ciência lida, não com 'aquilo que foi dado' pela natureza ao cientista, mas com 'aquilo que foi tomado' ou selecionado da natureza pelo cientista de acordo com seu propósito" (JENSEN,1950).

O termo pode aparecer tanto no singular (dado), no plural (dados) ou na sua versão em inglês *data*.

> Em latim, data é o plural de datum. Historicamente em áreas científicas especializadas, o termo data, em inglês, é tratado como plural, tendo um verbo correspondente no plural, como na frase: "dados foram coletados e classificados". Em se tratando, no entanto, de uso moderno não científico, data geralmente não é empregado como um termo plural. Em vez disso, é tratado como um substantivo incontável (mass noun ou uncountable noun), semelhante a uma palavra como informação, levando um verbo correspondente no singular (SEMIDÃO, 2014, p. 71).

Hjørland (2018) ao abordar diversas definições de dados considera que a melhor seja de Fox e Levitin:

> Dentro desta estrutura, definimos um dado ou item de dados, como um triplo <***e, a, v***>, onde ***e*** é uma entidade em um modelo conceitual, ***a*** é um atributo da entidade ***e***, e ***v*** é um valor do domínio de atributo ***a***. Um dado afirma que a entidade ***e*** tem valor ***v*** para o atributo ***a***. Dados são os membros de qualquer coleção de itens de dados (HJØRLAND, 2018, tradução nossa). [7]

[6] No original: "It is an unfortunate accident of history that the term datum (Latin, past participle of doy "to give") rather than caftum (Latin, past participle of cafere^ "to take") should have come to symbolize the unit-phenomenon in science. For science deals, not with "that which has been given" by nature to the scientist but with "that which has been taken" or selected from nature by the scientist in accordance with his purpose[...]" (JENSEN, 1950, p. ix)

[7] No original: "Within this framework, we define a datum or data item, as a triple <e, a, v>, where e is an entity in a conceptual model, a is an attribute of entity e, and v is a value from the domain of attribute a. A datum asserts that entity e has value v for attribute a. Data are the members of any collection of data items" (HJØRLAND, 2018).

Para entender melhor essa definição podemos tomar como exemplo uma adaptação baseada no exemplo de Marcondes (2021). Observemos o número “1974”, considerando 1974 como um dado, então na afirmação “Anderson nasceu em 1974”, encontramos a entidade de quem se fala “Anderson”, um atributo ou propriedade dessa entidade “nasceu” e o valor desse atributo ou propriedade “1974”. “Dados, inclusive o Big Data, não fazem sentido, sem estarem referenciados à entidade e ao atributo ou propriedade desta – o metadado” (MARCONDES,2021, p. 5).

Resumindo: Dado (1974, uma unidade de data) é a representação simbólica ou o símbolo do valor (o ***v*** da tripla <***e, a, v***>) de uma propriedade observável/observada (o ***a*** da tripla <***e, a, v***>) do fenômeno ou objeto que está sendo analisado (o ***e*** da tripla <***e, a, v***>). “Os símbolos fazem sentido quando fazem referência a uma entidade e às suas correspondentes propriedades, isto é, aos metadados de uma entidade, na forma de triplas <***e, a, v***>” (MARCONDES; RAMOS JUNIOR; MARTINS, p. 154, 2021).

Com o avanço das tecnologias digitais, as coisas do mundo natural são cada vez mais representadas na forma de dados e compartilhadas em redes como a Internet. Os dados que são inseridos, gerados e criados no computador, tornam-se cada vez mais diversos e complexos (ZHU; XIONG, 2015).

Os dados podem ser primários, dados que não manipulados ou alterados por pesquisadores. Ao serem manipulados por pesquisadores, ou seja, os dados primários, depois de filtrados, analisados ou organizados, tornam-se dados secundários ou transformados (NATIONAL RESEARCH COUNCIL, 1999; PIORUN, 2013).

Quanto aos dados de pesquisa eles podem ser definidos como:

> Registros factuais (pontuações numéricas, registros textuais, imagens e sons) usados como fontes primárias para pesquisa científica e que são comumente aceitos na comunidade científica como necessárias para validar os resultados da pesquisa. Um conjunto de dados de pesquisa constitui uma representação sistemática e parcial do assunto que está sendo investigado (US NATIONAL, 1997, tradução nossa[8]).

Segundo o National Science Board (NSB) (2011), essa definição inclui tanto os dados analisados quanto os metadados, que abordaremos mais tarde, que descrevem

[8]No original: “research data are defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated" (US NATIONAL, 1997).

como esses dados foram gerados. Os dados analisados são limitados a informações digitais que descrevem os resultados de pesquisa financiada, incluindo imagens digitais, tabelas publicadas e tabelas usadas para criação de visualizações gráficas (PAMPEL et. al., 2013; NATIONAL SCIENCE BOARD, 2011).

O National Institutes of Health (NIH) vê os dados de pesquisa como material fatual registrado e aceito comumente por uma comunidade científica como evidência para documentar e apoiar resultados de pesquisa (NATIONAL INSTITUTES OF HEALTH, 2003).

Os dados de pesquisa podem ser físicos ou digitais. A cada dia cresce o número de dados de pesquisa coletados digitalmente ou representados em meio digital, produzidos por pesquisadores e cientistas do mundo todo com o apoio das tecnologias da informação.

O Interagency Working Group on Digital Data do National Science and Technology Council nos USA define dados de pesquisa em formato digital como:

> [...] quaisquer dados digitais, bem como os métodos e as técnicas utilizados na criação e na análise desses dados, que um pesquisador precisar para verificar resultados ou ampliar conclusões científicas, incluindo dados digitais associados a informação não digital, como metadados associados a amostras físicas (NATIONAL SCIENCE BOARD, 2011, p. 1-33, tradução nossa [9])

Segundo Semeler e Pinto (2019, p.1) “os dados de pesquisa são o resultado de qualquer investigação sistemática que envolva processos de observação, experimentação ou simulação de procedimentos de pesquisa científica”.

Os dados de pesquisa podem compreender:

> [...] resultados de pesquisas originais [...] dados de pesquisa: os dados, os registros, arquivos ou outras evidências, independentemente do seu conteúdo ou forma (por exemplo, material impresso, digital, físico ou de outras formas), que compreendem observações de pesquisa, descobertas ou resultados, incluindo

[9] No original: “[...] for use or repurposing for scientific or technical research and educational applications [...] It refers to the full range of data types and formats relevant to all aspects of science and engineering research and education in local, regional, national, and global contexts with the corresponding breadth of potential scientific applications and uses [...]. Digital research data is any digital data, as well as the methods and techniques used in the creation and analysis of that data, that a researcher needs to verify results or extend scientific conclusions, including digital data associated with non-digital information, such as the metadata associated with physical samples” (NATIONAL SCIENCE BOARD, 2011, p. 1-33).

> materiais primários e analisados (RICE; SOUTHALL, 2016, p. 20, tradução nossa [10]).

Os dados de pesquisa podem ser classificados como dados observacionais, dados computacionais, dados experimentais. Podem ser também identificados como dados brutos (raw data) ou preliminares, dados derivados, dados referenciais ou canônicos. (SAYÃO; SALES, 2015).

Green; Macdonald; Rice (2009), definem dados observacionais, computacionais ou experimentais da seguinte forma:

Dados observacionais – São dados essencialmente históricos e não podem ser reproduzidos (por exemplo, observações da temperatura do oceano em uma data específica). Por serem datados no tempo e no espaço, pois não podem ser coletados uma segunda vez, devem ser preservados por tempo indeterminado.

Dados computacionais – São dados gerados a partir de modelos computacionais e eles podem exigir arquivamento completo das informações e a execução (por exemplo, hardware, software, dados de entrada). Diferente dos dados observacionais a sua preservação por longo prazo pode não ser necessária, visto que os dados podem ser replicados ao longo do tempo.

Dados experimentais – São dados provenientes de situações controladas em bancadas de laboratórios, como por exemplo, medidas de uma reação química, cromatogramas, microensaios. Os dados experimentais podem não ser facilmente reproduzidos, dadas as considerações de custo e a complexidade de todas as variáveis experimentais (GOLD, 2007; GREEN; MACDONALD; RICE, 2009).

Quanto aos dados brutos, derivados e canônicos segundo (GOLD, 2007; GREEN; MACDONALD; RICE, 2009; SAYÃO; SALES, 2015) eles podem ser definidos como:

Dados brutos (ou crus) - São dados que vêm diretamente dos instrumentos científicos.

[10] No original: “The recorded information necessary to support or validate a research project’s observations, findings or outputs’ [...] ‘That which is collected, observed, or created in a digital form, for purposes of analysing to produce original research results’ [...] ‘Research data: The data, records, files or other evidence, irrespective of their content or form (e.g. in print, digital, physical or other forms), that comprise research observations, findings or outcomes, including primary materials and analysed” (RICE; SOUTHALL, 2016, p. 20).

Dados derivados ou compilados - São resultados da transformação ou combinação de dados brutos ou de outros dados. Ex.: mineração de texto, dados de censo agregados.

Dados canônicos ou dados referenciais - São coleções de dados consolidados e arquivados geralmente em grandes centros de dados. Ex.: sequência genética, estrutura química.

Os dados de pesquisas necessitam ser preservados e compartilhados, porém apresentam especificidades que precisam de reflexão e análise. Dentre elas se destaca o fator de que os dados de pesquisa não possuem uma formalização estabelecida por serem de uma tipologia heterogênea em seu formato e em seu objeto. O seu compartilhamento envolve questões como ética e integridade na pesquisa, dados sensíveis, entre outros (VEIGA, 2017).

Os dados de pesquisa precisam ser identificáveis, citáveis, visíveis, recuperáveis, interpretáveis, contextualizáveis, interoperáveis e reutilizáveis onde quesitos de consistência e procedência são considerados. Eles são objetos digitais e não digitais, como documentos, questionários, avaliações, registros de casos, protocolos de estudo, planilhas, notas de laboratório, notas de campo, diários, filmes, imagens, arquivos digitais de áudio e vídeo, sequências genéticas, coordenadas geográficas, banco de dados, algoritmos, metodologias, protocolos, entre outros tipos de manifestação de pesquisa (ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2007; XIA; WANG, 2014; SAYÃO; SALES, 2015; DUDZIAK, 2016; KOLTAY, 2017; INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICES AND TECHNOLOGY, 2006; HENDERSON, 2017).

Podemos observar exemplos desta enorme quantidade de dados de pesquisa digitais nos dados gerados pelo telescópio espacial Hubble, **https://www.nasa.gov/mission_pages/hubble/main/index.html**, nos dados gerados pelo projeto de pesquisa Genoma Humano, **https://www.genome.gov/human-genome-project** **,** ou nos dados gerados pelo Large Hadron Collider, “o maior e mais poderoso acelerador de partículas do mundo, **https://home.cern/science/accelerators/large-hadron-collider**. Estes dados de pesquisa digitais são parte do fenômeno do Big Data” (MARCONDES, 2021).

Na ciência aberta podemos nos deparar tanto com o *Big Data* quanto com o *Small Data*. O termo *Big Data* pode ser aplicado quando dispomos de um grande

volume de dados, produzido e atualizado em alta velocidade apresentando questões específicas sobre sua veracidade, consistência e confiabilidade, tanto como o valor que os dados detêm, medido pela capacidade de estes gerarem novos conhecimentos e avanços (DEMCHENKO *et al*., 2013).

Entretanto, não se deve somente se preocupar em coletar grandes volumes de dados, mas também em busca transformar os dados em valor real (DAVENPORT, 2014; BURLINGAME; NIELSEN, 2014; VAN DER AALST, 2014; FEDERER, 2016)

Segundo Marcondes; Ramos Junior; Martins (2001) “a essência do fenômeno Big Data são os dados, uma vez que não existe Big Data sem dados”.

O *Small data*, ao contrário, são dados pequenos o suficiente para serem convenientemente armazenados em uma única máquina, particularmente em servidores locais ou em um desktop. Normalmente, estes dados que são coletados pelo pesquisador individualmente ou com um pequeno grupo de pesquisadores, normalmente formados por alunos de pós-doutorado, doutorado, mestrado e graduação (VEIGA, 2017).

2.2 GESTÃO, COMPARTILHAMENTO E ABERTURA DE DADOS.

O acesso aos dados de pesquisa é um dos pilares do movimento em prol da ciência aberta. Segundo Veiga os princípios da ciência aberta se baseiam no acesso aberto aos dados de pesquisa e às publicações científicas, principalmente às financiadas com recursos públicos; ferramentas e métodos de pesquisa abertos; processos de investigação colaborativos; a implementação de uma ciência cidadã; e a inovação aberta (VEIGA, 2017, p. 44).

Medeiros (2018) define dados abertos “como conjuntos de dados cujos metadados sejam obrigatoriamente públicos”, mas cujos dados por questões de ética e privacidade podem ser restritos.

A proveniência dos dados pode ser definida como o registro que descreve pessoas, instituições, entidades e atividades envolvidas na produção, influência ou entrega de um dado ou objeto (FREUND *et al,* 2019).

O termo proveniência de dados diz respeito à origem ou procedência dos dados. A proveniência também está relacionada à auditoria, triagem, linhagem e origem do dado (DAVIDSON; FREIRE, 2008).

A proveniência ajuda a responder questões sobre os dados, tais como: quem criou este dado e quando, o momento em que o dado foi modificado, e por quem e qual foi o processo usado para criar o dado.

De acordo com Davidson; Freire (2008), a proveniência pode ser dividida em três tipos:

• Prospectiva: trata-se da sequência de processos utilizados (receita) para a geração do dado, ou seja, captura os passos que devem ser seguidos para a geração de um dado produto.

• Retrospectiva: trata-se das informações obtidas durante a execução dos processos de geração do dado. Compreende desde o tempo de duração de cada atividade executada até a origem dos dados de entrada. Além disso, não depende do tratamento da proveniência prospectiva para ser utilizado. Em outras palavras, é como se fosse um log detalhado da execução de uma tarefa.

• Dados definidos pelo usuário: qualquer dado que o usuário julgar necessário para futura utilização. Como exemplo, pode-se citar anotações, conclusões a respeito do processo e, até mesmo, observações sobre parâmetros utilizados.

Saber como os problemas foram resolvidos durante os processos de coleta (ou experimento) também ajudarão os reutilizadores dos dados a identificarem se os dados foram devidamente processados (FANIEL; JACOBSEN, 2010).

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 reuso dos dados (FREUND *et al*, 2008).

Com o aumento da quantidade e variedade de dados de pesquisa, é importante ter uma gestão adequada dos dados para garantir sua qualidade, segurança e acessibilidade. A gestão de dados de pesquisa inclui a coleta, organização, documentação, armazenamento, compartilhamento e preservação de dados.

Um incentivo para os pesquisadores compartilharem dados de pesquisa vem da visibilidade, já que estudos recentes demonstraram que pesquisadores que compartilham dados em arquivos públicos recebem mais citações (PIWOWAR; VISION, 2013).

O incentivo para o compartilhamento de dados tem crescido através de políticas públicas de ciência e tecnologia, internacionais e nacionais. Financiadores internacionais e nacionais passaram a estabelecer recomendações sobre políticas e disponibilidades de dados. Iniciativas internacionais como National Science

Foundation https://www.nsf.gov/bfa/dias/policy/, World Bank Open Data https://data.worldbank.org/, e iniciativas nacionais como Open Knowledge Foundation https://ok.org.br/; Portal Brasileiro de Dados Abertos https://wiki.dados.gov.br/ e Fiocruz: https://www.arca.fiocruz.br/handle/icict/46408 são exemplos disso.

Cabe aos pesquisadores e as instituições o cuidado para que os seus dados de pesquisa sejam organizados, armazenados e disponibilizados com isso proporcionando a integridade e a reprodutibilidade da pesquisa (DUDZIAK, 2018).

A gestão de dados de pesquisa "cobre todo ou parte do ciclo de vida dos dados - desde a descoberta, coleta e organização de dados". (MICHENER, 2015, tradução nossa [11]), para isto, é fundamental para os pesquisadores e instituições terem um plano de Gestão de Dados (PGD), que é um documento que visa descrever o tratamento dos dados durante um projeto de pesquisa e o que ocorrerá com esses dados após a finalização da pesquisa.

Um importante guia para a gestão de dados de pesquisa é o documento da associação Science Europe, https://www.scienceeurope.org/our-resources/practical-guide-to-the-international-alignment-of-research-data-management/. Esse guia é basilar para estabelecer o que é essencial para compor o PGD. Importante ressaltar que os dados a serem disponibilizados precisam estar certificados com uma licença aberta para garantir o seu reuso de forma legal.

A Fiocruz elaborou um panorama geral das perguntas que podem vir a ser requisitadas em um PGD e que se encontra em, https://www.arca.fiocruz.br/handle/icict/54805, foram consideradas neste guia:

> a) a descrição dos dados; b) a descrição da documentação e dos metadados; c) as questões éticas e de conformidade legal; d) o armazenamento e backup; e) a política de preservação; f) o compartilhamento; g) as responsabilidades e os recursos financeiros (VEIGA et al, 2022).

A preservação digital dos dados de pesquisas permite que eles tenham uma longevidade e uma garantia de que permaneçam disponíveis, recuperáveis e compreensíveis ao longo do tempo (NASCIMENTO; ARAÚJO; ARELLANO, 2020).

O repositório digital confiável é um instrumento estratégico para a preservação dos dados de pesquisa, pois, segundo Sayão e Sales (2015, p. 23), os dados ao serem

[11] No original: "[...]cover all or portions of the data life cycle—from data discovery, collection, and organization[...]" ((MICHENER, 2015).

“depositados em ambientes que garantam sua preservação ativa por longo prazo, mantendo as suas características de autenticidade, integridade e proveniência, de forma que eles estejam sempre disponíveis e prontos para serem usados”, permitindo com que os dados sejam recuperados, usados, reusados, compartilhados e citados por pesquisadores.

Repositório digital pode ser definido como um arquivo digital que reúne uma coleção de documentos digitais com a finalidade de organizar, reunir e facilitar o acesso à produção científica de instituições de ensino e pesquisa. (WEITZEL, 2006).

> Os repositórios digitais que adotam o modelo OAI, isto é, que adotam o protocolo OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting), compartilham os mesmos metadados, tornando os seus conteúdos interoperáveis entre si. De um modo geral, os termos “repositórios institucionais” ou “temáticos” são adotados para caracterizar os repositórios digitais que reúnem respectivamente a produção científica de uma instituição ou área. (WEITZEL, 2006, p.139).

Em resumo podemos afirmar que o repositório temático prioriza determinado domínio do conhecimento, enquanto o institucional prioriza a produção de uma instituição específica.

Os repositórios promovem o aumento da visibilidade do impacto dos resultados de pesquisa através de adequado planejamento, implementação e adoção (LEITE, 2009).

Como citado por Weitzel (2006), a adoção do protocolo OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) foi fundamental diante do crescimento dos repositórios. Segundo Marcondes; Sayão (2002, p. 46) OAI-PMH é “um conjunto de especificações técnicas e princípios organizacionais bastante simples, porém potencialmente poderosos e de grande alcance, no objetivo de integração desses arquivos” o que possibilita os repositórios tornarem-se interoperáveis.

Com a disseminação das tecnologias de informação os dados de pesquisa digitais proliferaram, sendo o uso crescente de tecnologias uma das principais fontes geradoras de dados. Os dados de pesquisa são o resultado dos mais diferentes processos de investigação através de dispositivos técnicos, transformando todas as etapas da pesquisa, desde um experimento que gera dados ao um estudo empírico

sobre a observação de fenômenos culturais, até a publicação de resultados de pesquisa em um repositório de dados de pesquisa (PAMPEL, 2013).

Para se obter informações sobre repositórios de dados é possível consultar o Re3Data – Registry of Research Data Repositories – um diretório de registro de repositórios de dados de pesquisa.

O Re3Data é um cadastro global de repositórios de pesquisa. Foi fundado em 2012 e é financiado pela Fundação Alemã de Pesquisa (DFG) (https://www.re3data.org/about).

O Re3data congrega iniciativas da Library and Information Services (LIS) of the GFZ German Research Centre for Geosciences, da Library of the Karlsruhe Institute of Technology (KIT), da School of Library and Information Science (BSLIS) na Humboldt-Universität, em Berlin e das Libraries of the Purdue University, na Alemanha.

O Re3data oferece a pesquisadores, organizações de financiamento, bibliotecas e editores uma visão geral sistemática da paisagem heterogênea dos repositórios de dados de pesquisa (PAMPEL et al., 2013; COSTA; BRAGA, 2016).

O Re3data abrange diferentes disciplinas acadêmicas, apresenta uma relação de repositórios que servem para o armazenamento permanente e o acesso a conjuntos de dados de pesquisa em que órgãos de financiamento, editores e instituições acadêmicas promovem uma cultura de compartilhamento, acesso e visibilidade aos dados de pesquisa (RE3DATA, 2018).

2.3 METADADOS

Vimos sobre os dados de pesquisa e seu armazenamento, mas como se ter acesso aos dados? Segundo Veiga (2017, p. 48) “o compartilhamento de dados é fundamental para prover acesso a esses dados”. Porém este compartilhamento precisa ser realizado de forma adequada.

> [...] os dados devem estar acessíveis e prontamente localizados; eles devem ser inteligíveis para aqueles que desejam examiná-los; os dados devem ser avaliados para que possam ser feitos julgamentos sobre sua confiabilidade e a competência daqueles que os criaram; e eles devem ser utilizáveis pelos outros. Para que os dados atendam a esses requisitos, eles devem ser suportados por metadados explicativos (dados sobre dados). (ROYAL SOCIETY, 2012, p.7, tradução nossa[12], grifo nosso).

[12] No original: “data must be accessible and readily located; they must be intelligible to those who wish to scrutinise them; data must be assessable so that judgments can be made about their reliability and the

Os metadados são fundamentais no contexto da gestão, compartilhamento e reuso de dados.

> 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. Os metadados cumprem essa tarefa, porque eles são a documentação dos dados. Os metadados que são usados para descreverem os dados, permitem que eles estejam autodocumentados agora e no futuro (SAYÃO; SALES, 2015, p. 23).

Metadado é comumente definido como sendo dados sobre dados, mas possui uma definição mais abrangente. São atributos que representam uma entidade (objeto do mundo real) em um sistema de informação (HJORLAND 2018).

O objetivo e a função dos metadados se encontra nos princípios da catalogação, garantindo assim a padronização dos recursos informacionais, através de regras, códigos e esquemas internacionais que facilitam identificar, buscar, localizar, recuperar, preservar, o uso e o reuso dos recursos informacionais (ALVES, 2005; CASTRO, 2008).

“Esquemas”, “padrões”, “formato”, “sistema”, e “conjuntos de elementos” de metadados são termos que são usados sem distinção na literatura para se referir a padrões de metadados (CHAN; ZENG, 2006; NATIONAL INFORMATION STANDARDS ORGANIZATION, 2004).

Segundo o WIKI do IBICT (INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIAS) “metadados são agrupados em esquemas, que os organizam, normalizam e os descrevem, criando padrões” (WIKI, 2018). Estes padrões de metadados facilitam a interoperabilidade entre sistemas.

Os padrões de metadados possibilitam que os dados de pesquisa possam ser descritos, obtendo dentre outras informações, as de sua proveniência.

Segundo Alves:

> [...] o termo metadados está intimamente associado com a definição de padrões de metadados. Para que eles possam existir, os

competence of those who created them; and they must be usable by others. For data to meet these requirements it must be supported by explanatory metadata (data about data)” (ROYAL SOCIETY, 2012).

> metadados (metadata) devem estar codificados em estruturas padronizadas de descrição denominadas como padrões de metadados (metadata statement) (ALVES, 2010, p. 48).

De acordo com Woodley (2016), os padrões de metadados são reflexos da funcionalidade da informação e do conhecimento que ficam armazenados e são expressos no processamento em sistemas e nos mecanismos de busca, possibilitando a descoberta da informação por meio dos mecanismos de busca.

Segundo Formenton *et al* (2017, p.6):

> O “esquema” é uma entidade total contendo os componentes semânticos e de conteúdo (tidos como um “conjunto de elementos”), bem como a codificação dos elementos com uma sintaxe ou linguagem de marcação, como Standard Generalized Markup Language (SGML) e Extensible Markup Language (XML). Assim, um conjunto de elementos de metadados dispõe de dois componentes básicos:
>
> 1) Semântica – as definições ou os significados dos elementos e seus refinamentos;
>
> 2) Conteúdo – as declarações ou as instruções de quais e como os valores devem ser atribuídos para os elementos.

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 codificação dos elementos e seu conteúdo (NATIONAL INFORMATION STANDARDS ORGANIZATION, 2004; CHAN; ZENG, 2006).

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 (OMS,2020).

Uma compreensão imediata da epidemiologia da doença COVID-19 é crucial para retardar infecções, minimizar mortes e tomar decisões informadas sobre quando e em que medida impor medidas de mitigação e quando e como reabrir a sociedade.

Apesar de nossa 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 ao 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.

Há uma necessidade premente de um sistema global coordenado que englobe preparação, detecção precoce e resposta rápida a novas ameaças emergentes, como tem sido o vírus SARS-CoV-2 e a doença COVID-19 que ele causa.

O público-alvo das recomendações e diretrizes epidemiológicas são agências governamentais e internacionais, formuladores de políticas e decisores, epidemiologistas e especialistas em saúde pública, especialistas em preparação e resposta a desastres, financiadores, fornecedores de dados, professores, pesquisadores, médicos e outros usuários em potencial.

A fidedignidade, acesso e potencial reuso de dados em COVID-19 e relacionados aos diversos aspectos da doença são 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 são:

> [...] um acrônimo para Findable, Accessible, Interoperable e Reusable, está presente nas discussões e práticas contemporâneas da ciência de dados, desde o início de 2014, 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. Desde então, tais princípios passaram a serem norteadores da descoberta, do acesso, da interoperabilidade, do compartilhamento e da reutilização dos dados de pesquisa. (HENNING; et al, 2018, p.1)

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 (FORMENTON et al, 2017).

Segundo Marcondes (2021, p.4):

> Os princípios FAIR enfatizam também o uso das tecnologias DAI, ou seja, os conjuntos de dados devem ser identificados e acessíveis através de identificadores/*links* persistentes – URI -, dados, devem estar associados a licenças explícitas que permitam seu reuso, dados e seus respectivos metadados devem utilizar vocabulários padronizados internacionalmente e amplamente aceitos e reconhecidos. Os princípios FAIR visam garantir, além da abertura de modo a facilitar o reuso dos dados de pesquisa, também o princípio M4M[13] – *Metadata for Machines* - (metadados para máquinas): "There is no FAIR data without machine-actionable metadata", que os dados de pesquisa possam ser "inteligíveis", "compreensíveis" por computadores

GO FAIR é uma iniciativa internacional que tem por objetivo promover o desenvolvimento coerente da Internet global de serviços e dados FAIR, seguindo as diretrizes do European Open Science Cloud (EOSC). Há escritórios regionais da rede GO-FAIR em diversos países, como Alemanha, Áustria, Brasil, China, Dinamarca, França, Estados Unidos e Holanda (DRUCKER et al, 2021, p. 167)

O GO FAIR tem por prioridade tornar todos os tipos de dados, que se encontram fragmentados e desconectados, mais facilmente localizáveis, acessíveis, interoperáveis e reutilizáveis, ou seja, FAIR, facilitando, dessa forma, o seu reconhecimento por máquinas e pessoas. (VEIGA; QUEIROZ, 2019)

A Rede de Implementação GO FAIR Brasil Saúde (https://www.go-fair-brasil.org/) é uma rede temática responsável pela elaboração de estratégias de implementação dos princípios FAIR no campo da saúde. Sua coordenação está sob a responsabilidade do Instituto de Comunicação e Informação Científica e Tecnologia em Saúde (ICICT) da Fundação Oswaldo Cruz (Fiocruz), que conta com a participação de diversas instituições das áreas de Saúde Pública, Vigilância Sanitária, Informação e Comunicação em Saúde, História do Patrimônio Cultural das Ciências e da Saúde, Oncologia, Enfermagem e Educação Profissional em Saúde. (VEIGA; QUEIROZ, 2019).

[13] Ver em https://www.go-fair.org/resources/go-fair-workshop-series/metadata-for-machines-workshops/m4m/

Diante do exposto verifica-se a importância de se identificar quais são os metadados de proveniência que seriam fundamentais para o reuso de dados no contexto da pesquisa em COVID-19.

3 PROCEDIMENTOS METODOLÓGICOS

O estudo se caracteriza como pesquisa exploratória, descritiva, de caráter qualitativo.

As pesquisas exploratórias pretendem observar e compreender os mais variados aspectos relativos ao fenômeno estudado pelo pesquisador. As pesquisas exploratórias mais comuns são os levantamentos bibliográficos (GIL, 2017)

A pesquisa descritiva descreve uma realidade, visa descrever um determinado evento, realidade ou situação (GIL, 2017). Pesquisa qualitativa examina evidências baseadas em dados verbais e visuais para entender um fenômeno em profundidade. Portanto, seus resultados surgem de dados empíricos, coletados de forma sistemática. Nas pesquisas qualitativas, você pode utilizar entrevistas, grupos focais ou observações (GIL, 2017).

Para alcançar os objetivos propostos foram realizados os seguintes passos metodológicos:

- **Para identificar os padrões de metadados utilizados para dados de pesquisa e mapear suas diretrizes internacionais (objetivos 1 e 2)** foi realizado levantamento bibliográfico e documental, em duas etapas:

Etapa 1 - Levantamento bibliográfico

No contexto da revisão bibliográfica, segundo Lakatos e Marconi (2001, p. 183) a pesquisa bibliográfica “[...] abrange toda bibliografia já tornada pública em relação ao tema estudado, desde publicações avulsas, boletins, jornais, revistas, livros, pesquisas, monografias, teses, materiais cartográficos, etc. [...]”, proporcionando ao pesquisador contato direito com contribuições de diversos autores sobre determinado assunto.

Para a pesquisa bibliográfica foram selecionadas as seguintes fontes de informações internacionais e nacionais: SCOPUS (via portal Capes), Medical Literature Analysis and Retrieval System Online (Medline) via PubMED, Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI) e a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD). O Google Acadêmico (Google Scholar) também foi utilizado como fonte de informação complementar.

Após a revisão da literatura que suportou a elaboração do referencial teórico do projeto, foram selecionados alguns termos que apoiassem o levantamento

bibliográfico da etapa 1 da metodologia. Foram identificados descritores no Descritores em Ciências da Saúde (DeCS) e no Medical Subject Headings (MeSH).

Desta forma, a busca envolveu tanto termos controlados utilizados pelas bases de dados, bem como linguagem natural, na ausência de vocabulário controlado.

Após a seleção dos termos mais representativos do tema da pesquisa para a construção das estratégias de busca, os procedimentos subsequentes foram buscas simples e combinadas de acordo com as especificidades de cada fonte de informação selecionada.

Nas bases Web of Science e PubMed e Google acadêmico procederam-se buscas simples e, quando possíveis, avançadas utilizando-se operadores booleanos com os termos a partir do campo específico assunto, título, resumo, palavras-chave e resumo, com o uso de filtro de data de publicação de 5 anos.

Nas bases Web of Science, PubMed e Google acadêmico foram aplicados os seguintes termos: covid; interoperability; COVID-19; interoperable; sars vírus; metadata standards; sars-cov-2; metadata; sars; medical records; coronavírus; coronavirus disease.

Efetuada a pesquisa bibliográfica, os dados coletados foram descarregados em formato CSV (Comma Separated Values) e exportados para o formato XLS na ferramenta Excel.

A etapa seguinte consistiu na limpeza dos dados a partir de: 1) remoção de duplicatas; 2) análise do título, resumo e palavras-chave.

Em seguida, foi realizada a leitura integral dos documentos selecionados que apoiou na complementação do referencial teórico desta pesquisa, a identificar os padrões de metadados utilizados na pesquisa em COVID-19 e a mapear as principais diretrizes apontadas na literatura para o tema.

Etapa 2 - Levantamento documental

Esse procedimento distancia-se do caráter bibliográfico no que tange à natureza das fontes, pois “vale-se de materiais que não receberam ainda um tratamento analítico, ou que ainda podem ser reelaborados de acordo com os objetivos da pesquisa” (GIL, 2012, p. 51), como políticas, editais, relatórios, procedimentos, diretrizes, entre outros documentos dessa natureza.

Para isso foi realizado um mapeamento das diretrizes internacionais para padrão de metadados para dados de pesquisa nas iniciativas Fairsharing

(https://fairsharing.org/), Digital Curation Centre (DCC) (https://www.dcc.ac.uk/) e Research Data Alliance (RDA) (https://www.rd-alliance.org/).

O FAIRSHARING é um recurso de suporte ao FAIR que fornece um registro informativo e educacional sobre padrões de dados, bancos de dados, repositórios e políticas, juntamente com ferramentas e serviços de pesquisa e visualização que interoperam com outros recursos que estão de acordo com os princípios FAIR.

O FAIRsharing orienta os consumidores a descobrir, selecionar e usar padrões, bancos de dados, repositórios e políticas com confiança, e os produtores a tornar seus recursos mais detectáveis, mais amplamente adotados e citados.

O GRUPO DO DCC para estudo dos padrões de metadados é um centro de especialização reconhecido internacionalmente em curadoria digital, com foco na criação de capacidades e habilidades para o gerenciamento de dados de pesquisa.

A Research Data Alliance (RDA) é uma iniciativa para construir conexões técnicas e sociais para viabilizar o compartilhamento aberto de dados de pesquisa. A visão da RDA é tornar viável que pesquisadores possam compartilhar abertamente seus dados entre diferentes tecnologias, disciplinas e países, de forma a endereçar os grandes desafios da sociedade em escala global.

- **Para verificar quais os padrões de metadados de proveniência são considerados relevantes pelos pesquisadores no reuso dos dados em COVID-19 (Objetivo 3)** foi elaborado instrumento de coleta de dados (questionário online).

Etapa 3 – Pesquisa empírica.

A pesquisa empírica serve para ancorar e comprovar no plano da experiência aquilo apresentado conceitualmente, investiga um fenômeno contemporâneo dentro do seu contexto da vida real, especialmente quando os limites entre o fenômeno e o contexto não estão claramente definidos (YIN, 2001 p. 33). A observação e experimentação empíricas oferecem dados para sistematizar a teoria.

Nesta etapa foi dada voz também ao ator principal que é o pesquisador, para que informe o que o predisporia a reusar dados de pesquisa em COVID-19.

O corpus da pesquisa foi formado por pesquisadores que receberam financiamento, via edital da Fiocruz, para investigação em COVID-19, no período de 2020 a 2021.

A Fiocruz é uma destacada instituição de ciência e tecnologia em saúde da América Latina reconhecida como importante na pesquisa no campo da COVID-19. É

referência internacional em pesquisa no campo da saúde pública, presente fisicamente em todas as regiões do Brasil, e lidera no país os esforços mundiais contra o novo coronavírus.

O instrumento de coleta de dados foi composto pelos seguintes elementos: e-mail de apresentação expondo os objetivos da pesquisa, instruções para preenchimento, prazos, identificação e contato dos responsáveis; Termo de Consentimento Livre e Esclarecido (TCLE); e questionário.

O questionário foi elaborado observando o referencial teórico e os objetivos da pesquisa projetado com 23 perguntas divididas em 5 partes, sendo: (i) Pergunta sobre conhecimento do Repositório FAPESP e seu uso, 2 questões; (ii) Informações sobre os dados com 9 questões; (iii) Informações sobre o produtor de dados com 4 questões; (iv) Informações sobre o repositório onde os dados estão disponibilizados com 3 questões e (v) Informações sobre o perfil do pesquisador com 5 questões.

O formato das questões foi com diferentes opções, tais como: do tipo aberta e fechada com múltiplas escolhas e escalonadas. As alternativas escalonadas estão organizadas em escala, de maneira que o respondente indique o seu posicionamento diante da pergunta. Com quatro pontos a fim de atribuir valores quantitativos para respostas qualitativas, sendo: (1) Nada Importante; (2) Pouco Importante; (3) Importante; (4) Muito Importante.

O instrumento foi criado com o aplicativo Google Forms e pode ser visto no Apêndice C.

Também foram utilizados como fontes de informação para esta pesquisa os insumos e sugestões do Grupo GOFAIR Brasil Saúde e do projeto VODAN-Br, GT-Metadados e proveniência de pesquisa.

O processo de gestão dos dados coletados nesta pesquisa está descrito no Plano de Gestão de Dados, elaborado a partir do modelo simplificado fornecido pela Fiocruz (apêndice A). O conjunto de dados gerados durante a etapa da pesquisa documental bem como o PGD da pesquisa foram depositados na comunidade da Fiocruz, disponível no repositório multidisciplinar de acesso aberto Zenodo e estão registrados sob o DOI: 10.5281/zenodo.7643810.

Segue um quadro síntese da metodologia adotada para atingir os objetivos apresentados.

Quadro 1 – Síntese da metodologia

Objetivos específicos	Procedimentos metodológicos	Página
1- Identificar os padrões de metadados mais utilizados para representação de conjuntos de dados de pesquisa em COVID-19.	Levantamento Bibliográfico e Levantamento documental	46 a 69
2 - Mapear as diretrizes internacionais para padrões de metadados para dados de pesquisa, alinhados aos princípios FAIR.		
3 - Verificar quais os padrões de metadados de proveniência apoiam os pesquisadores no reuso dos dados em COVID-19	Elaboração e aplicação de instrumentos de coleta de dados (questionário online semiestruturado)	70 a 81

Fonte: Elaboração própria

4 APRESENTAÇÃO DOS RESULTADOS

Nesta seção serão apresentados os resultados da pesquisa do levantamento bibliográfico e documental e o artigo científico que foi publicado com o levantamento dos padrões de metadados utilizados pelos repositórios de dados de pesquisa que disponibilizam conjuntos de dados em COVID-19 em acesso aberto e as diretrizes internacionais para padrões de metadados de dados de pesquisa.

Como informado, na seção da metodologia após o levantamento bibliográfico e documental foi realizado um mapeamento, no Re3Data, dos repositórios de dados com conjuntos de dados de pesquisa em COVID-19, como resultado foi produzido o artigo “Metadados para representação de dados em COVID-19: um estudo exploratório”.

Na seção a seguir será apresentada uma síntese do artigo desenvolvido durante esta etapa da pesquisa.

4.1 ARTIGO - METADADOS PARA REPRESENTAÇÃO DE DADOS EM COVID-19: UM ESTUDO EXPLORATÓRIO.

O artigo “Metadados para representação de dados em COVID-19: um estudo exploratório” foi submetido e apresentado como Comunicação Oral na 13ª Conferência Lusófona de Ciência Aberta (ConfOA) em outubro de 2022. Este trabalho foi desenvolvido por mim, com apoio de Isabella Henrique Lima Pereira e Mylena Cristhina Araujo de Oliveira e sob a supervisão da orientadora Viviane Veiga, como parte do Projeto VODAN BR. O texto completo será publicado na Revista Científica da Universidade Eduardo Mondlane - (RC-UEM).

O artigo apresenta o resultado do levantamento dos padrões de metadados utilizados pelos repositórios de dados de pesquisa que disponibilizam conjuntos de dados em COVID-19 em acesso aberto (Apêndice 1), a seguir apresenta-se uma síntese do artigo.

Por meio da revisão e da análise da literatura foram identificados vários padrões ou esquemas de metadados usados para a descrição de recursos em distintos domínios. Os padrões de metadados são elaborados para uma extensa diversidade de usos, porém, os esquemas são delimitados por seus próprios conjuntos de elementos de metadados, particularidades e domínios de utilização.

Uma maneira de auxiliar os países a melhorarem o fluxo de atendimento à população e de diagnósticos se dá pela partilha, integração, e interoperabilidade de dados clínicos. Para que isso ocorra, faz-se necessário a adoção de metadados e de padrões de metadados consolidados e metodologicamente construídos, almejando a construção e a modelagem de ambientes digitais bem estruturados e com alto grau de padronização na descrição dos dados, para a recuperação efetiva e de qualidade da informação.

Ao analisar a 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 na PubMed. Na Web of Science o metadado foi encontrado no artigo “COVID-19 pandemic reveals the peril of ignoring metadata standards”.

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.

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 de diferentes disciplinas acadêmicas. É mantido financeiramente pela Fundação Alemã de Pesquisa e coordenado por instituições científicas e acadêmicas na Alemanha.

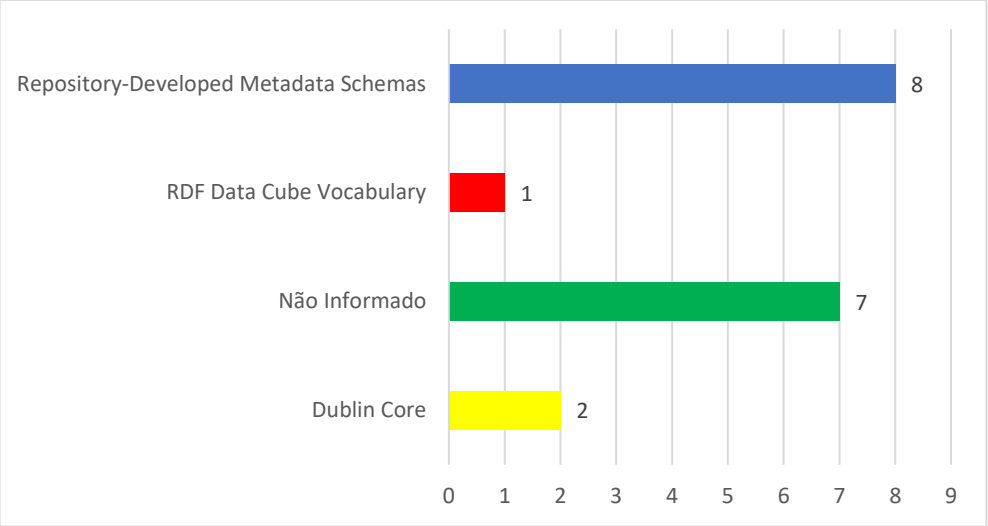
O levantamento ocorreu no dia 30 de março de 2022.

Nesta fonte identificou-se os repositórios de dados com conjuntos de dados em COVID-19, que nos deu o resultado de 843 repositórios. Em seguida, o campo das palavras-chave (COVID-19) o resultado foi de 18 repositórios.

O idioma adotado pelos repositórios é majoritariamente inglês, a maioria dos repositórios pertence aos Estados Unidos. O Brasil aparece com um repositório. 8 (oito) repositórios indicaram que utilizam esquemas de metadados desenvolvidos pelo próprio repositório, 2 (dois) o Dublin Core 1 (um) informou o RDF Data Cube

Vocabulary. 7 (sete) repositórios não informaram o padrão de metadados adotado (Figura 1).

Figura 1 - Padrão de Metadados utilizados pelos repositórios



Fonte: Araujo, 2023.

Verificou-se que os repositórios do campo da saúde foram os que mais armazenaram dados de pesquisa em COVID-19.

Quanto ao padrão de metadados verificou-se que a maioria destes repositórios 8, possui um esquema de metadados próprio, o que pode significar que os esquemas atuais não estão atendendo às demandas de representação descritiva e temática dos dados de pesquisa.

Em nenhum repositório foi encontrado menção do padrão que é recomendado pelo W3C, uma comunidade internacional que desenvolve padrões abertos para garantir o crescimento da Web a longo prazo (https://www.w3.org/), que é o DCAT (DATA CATALOG VOCABULARY), padrão próprio para descrever repositório de dados. O DCAT é um vocabulário RDF projetado para facilitar a interoperabilidade entre catálogos de dados publicados na Web (https://www.w3.org/TR/vocab-dcat-1/).

4.2 DIRETRIZES INTERNACIONAIS PARA PADRÕES DE METADADOS PARA DADOS DE PESQUISA.

Para cumprir essa etapa foi realizado um mapeamento das diretrizes internacionais para padrão de metadados para dados de pesquisa nas iniciativas

internacionais Fairsharing, Digital Curation Centre (DCC) e Research Data Alliance (RDA).

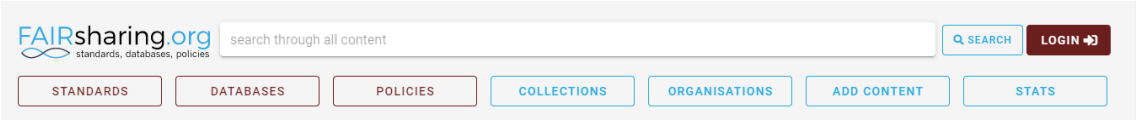
4.2.1 Fairsharing

O Fairsharing é um recurso com curadoria, informativo e educacional sobre padrões de dados e metadados, inter-relacionados a bancos de dados e políticas de dados (https://fairsharing.org/communities#governance). Segundo o "catalogue & Marketplace"[14] do European Open Science Cloud, os recursos referenciados no Fairsharing são interoperáveis com outros recursos compatíveis com os princípios FAIR.

O Fairsharing traz orientações para os usuários descobrirem, selecionarem e usarem esses recursos com confiança, e os produtores, a tornar seus recursos encontráveis, mais amplamente adotados e citados. É um portal pesquisável baseado na Web contendo descrições internas e coletivas de padrões, bancos de dados e políticas de dados.

As buscas podem ser feitas segundo as categorias na Figura 2:

Figura 2 – Campo de pesquisa Fairsharig



Fonte: Fairsharing.org

Standards: registro de artefatos de terminologia, modelos/formatos, diretrizes de relatórios e esquemas de identificadores.

Databases: registro de bases de conhecimento e repositórios de dados e outros ativos digitais.

Policies: registro de políticas de preservação, gerenciamento e compartilhamento de dados de agências internacionais de financiamento, reguladores, periódicos e outras organizações.

Collections: agrupam um ou mais tipos de recursos (padrão, banco de dados ou política) por domínio, projeto ou organização.

[14] https://marketplace.eosc-portal.eu/datasources/eosc.oxford_e-research_centre.21697de1a5b10b8eb5fad857edecf5c9, acesso em: 21/01/2023

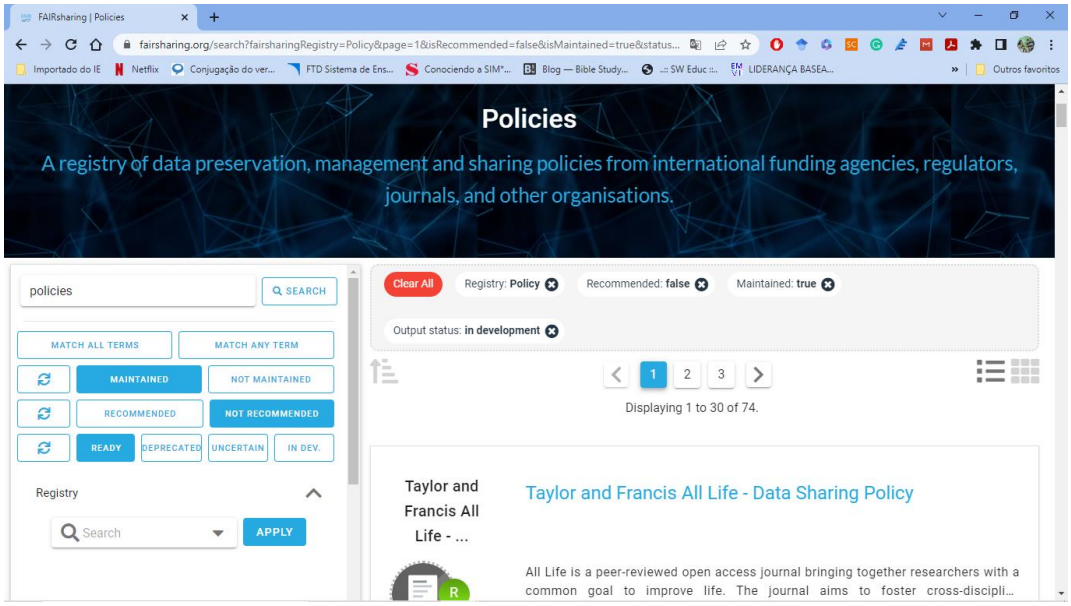
Organisations: lista as organizações por ordem alfabética e possui também um campo de busca.

Add Content: permite adicionar conteúdo no FAIRsharing. Local onde o pesquisador descreve seu padrão, banco de dados ou política de dados no FAIRsharing para torná-lo mais detectável, mais amplamente adotado e citado.

Stats: apresenta gráficos que descrevem os padrões, bancos de dados e políticas de dados armazenados nos registros FAIRsharing (Fairsharing.org).

Ao acionar o link “POLICIES” existe um menu onde se pode restringir os recursos sobre “POLICIES” segundo as seguintes categorias (Figura 3).

Figura 3 – Policies Fairsharing



Fonte: Fairsharing.org

Opções de filtro seriam “Match all terms” (Buscar todos os termos) “Match any terms” (Buscar quaisquer termos).

Podemos usar um filtro para restringir a busca pelos recursos cujos metadados são mantidos (ou não) diretamente por representante dos próprios recursos.

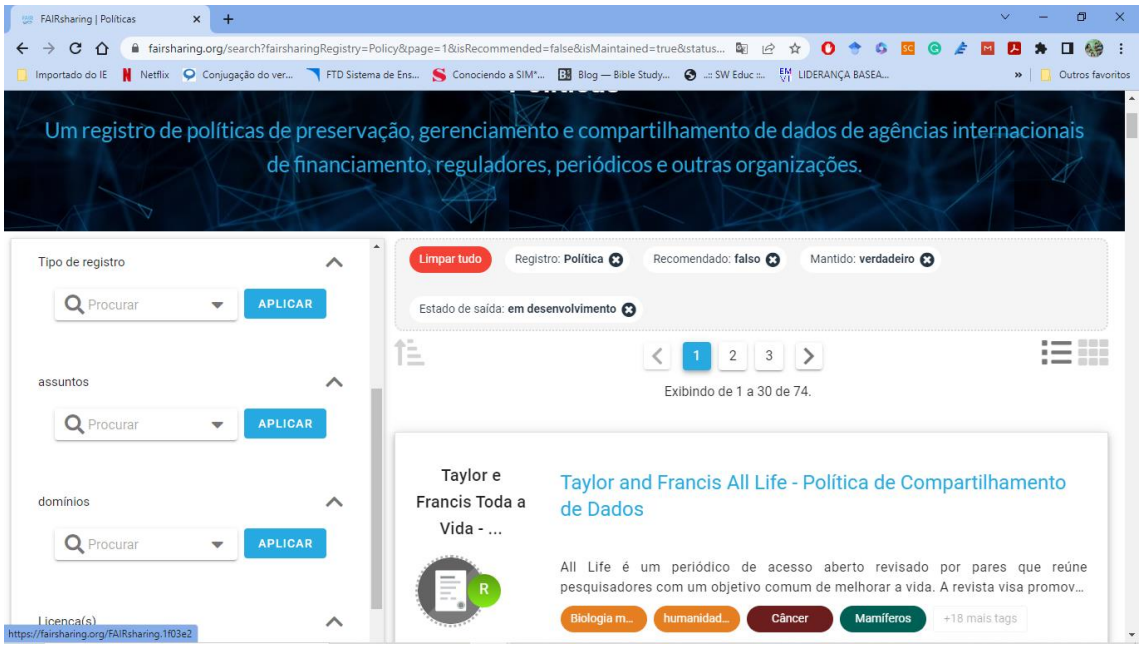
Podemos selecionar também um filtro para que a busca possa ser (ou não) por um recurso recomendado por uma política de dados de um periódico, editor de periódico ou financiador.

Podemos selecionar também um filtro para que a busca possa mostrar todos os registros independentemente de seu status: mostrando apenas os recursos que estão ativos e prontos para uso; mostrando apenas os recursos que foram obsoletos;

mostrando apenas os recursos que não temos certeza de seu status; mostrando apenas os recursos que estão atualmente em desenvolvimento e não estão prontos para uso.

Também encontramos filtros por: registro; tipo de registro; assuntos; domínios; licenças; organizações; países e espécies (Figura 4).

Figura 4 – Filtros Fairsharing



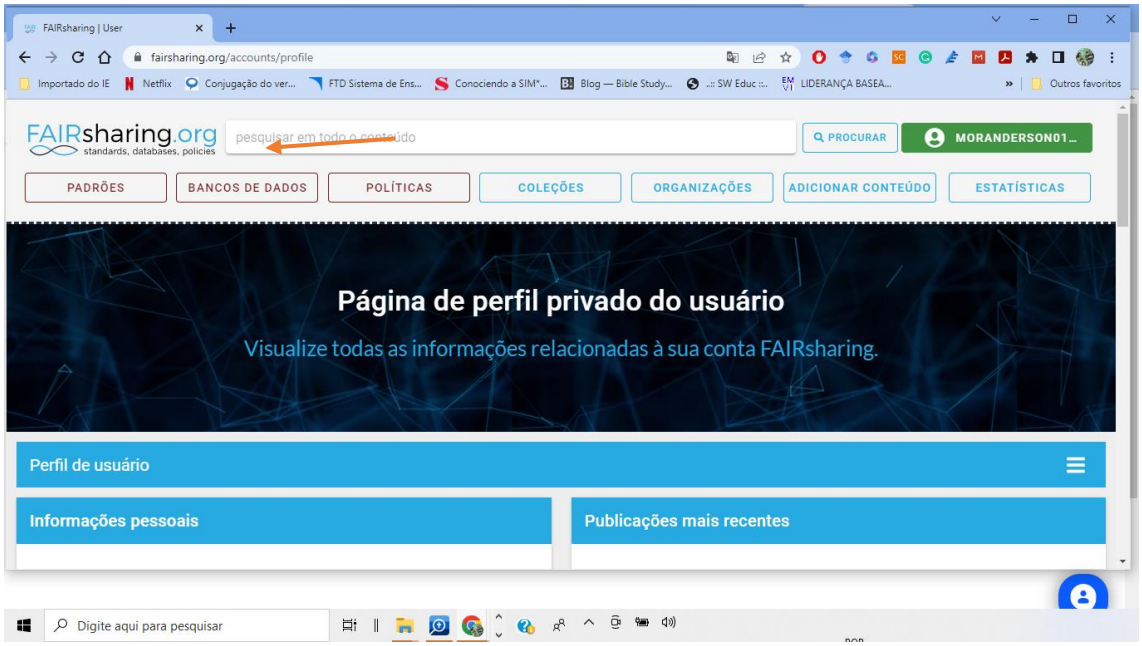
Fonte: Fairsharing.org

O Fairsharing também permite a extração dos dados quando o usuário está logado no formato Comma Separated Values / Valores Separados por Vírgula (CSV). Um formato de arquivo com alto potencial de interoperabilidade, compatível, por exemplo, com planilhas Excel.

Ao buscarmos as diretrizes nas categorias citadas acima constatamos que os resultados não foram satisfatórios, pois eles não permitem cruzar os resultados com outras palavras-chave.

Porém, o Fairsharing também permite a pesquisa sem os filtros pré-determinados acessando o campo de pesquisa geral, o que permite o cruzamento de mais de um termo de busca, como apresentado na imagem a seguir (Figura 5).

Figura 5 – Pesquisa sem filtro Fairsharing



Fonte: Fairsharing.org

Para alcançar o objetivo de levantar diretrizes, políticas ou *guidelines* para padrões de metadados de pesquisa, criamos a seguinte estratégia de busca: “metadata standard” AND “research data” policy.

Ao aplicarmos a estratégia no campo de busca, para mapearmos as diretrizes em relação aos metadados de pesquisa, encontramos dois resultados, conforme quadro a seguir.

Quadro 2– Resultado Fairsharing

NOME	DESCRIÇÃO	Padrões
Genomic Science Program (GSP) Information and Data Sharing Policy, (https://genomicscience.energy.gov/datasharing/)	O objetivo geral de pesquisa do Programa de Ciência Genômica (GSP) é fornecer a ciência fundamental necessária para entender, prever, manipular e projetar sistemas biológicos que sustentam inovações para produção de bioenergia e bioprodutos e para aprimorar nossa compreensão de processos ambientais naturais relevantes para o DOE. A GSP faz parte da Divisão de Ciência de Sistemas Biológicos (BSSD) do Programa de Pesquisa Biológica e Ambiental (BER) do Departamento de Energia dos Estados Unidos.	O Programa de Pesquisa Biológica e Ambiental (BER) exige que todos os dados, metadados e softwares publicáveis resultantes de pesquisas financiadas pelo programa Genomic Science (GSP) estejam em conformidade com os formatos padrão reconhecidos pela comunidade quando existirem, sejam claramente atribuíveis e sejam depositados dentro de um banco de dados público reconhecido pela comunidade apropriado para a pesquisa. “metadata schema” https://fairsharing.org/FAIRsharing.19ne3m

RDA COVID-19 WG Resources (https://www.rd-alliance.org/groups/rda-covid19)	É uma coleção de recursos mantida pelo RDA COVID-19 WG, com diversos tipos de recursos, entre eles políticas e diretrizes Os objetivos gerais deste Grupo de Trabalho (GT) são: Definir claramente diretrizes detalhadas sobre o compartilhamento de dados nas atuais circunstâncias do COVID-19 para ajudar as partes interessadas a seguir as melhores práticas para maximizar a eficiência de seu trabalho, Desenvolver diretrizes para formuladores de políticas para maximizar o compartilhamento oportuno de dados e respostas apropriadas em tais emergências de saúde, e Para atender aos interesses de pesquisadores, formuladores de políticas, financiadores, editores e fornecedores de infraestruturas de compartilhamento de dados.	Atualmente, existem quatro padrões genéricos de metadados que são amplamente utilizados, Dublin Core (DC), Vocabulário de Catálogo de Dados (DCAT), DataCite e Schema.org. Este último tem uma especialização chamado Bioschemas, que fornece uma maneira de adicionar marcação semântica a páginas da Web para melhorar a capacidade de localização de dados nas ciências da vida e atualmente está atualizando perfis para ajudar na descoberta de dados do COVID-19.

Fonte: Elaboração própria

Fundamental ressaltar que usando outros critérios pode se chegar a outros resultados no Fairsharing.

4.2.2 DCC

O Digital Curation Center (DCC) é um centro líder mundial de especialização em curadoria de informações digitais. Tem como foco a construção de capacidade e habilidades para gerenciamento de dados de pesquisa.

Com isto, o DCC tem como função apoiar os curadores de dados na escolha do padrão mais apropriado a cada disciplina (Digital Curation Centre 2022a).

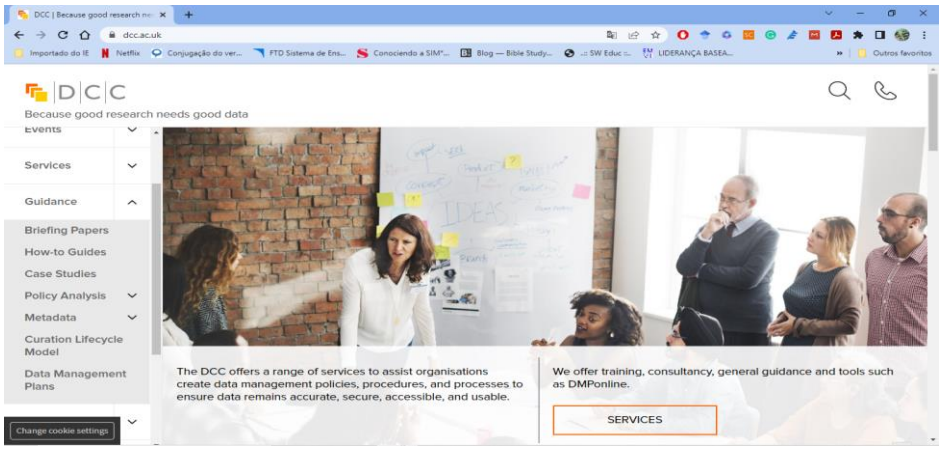
Ao usar a pesquisa geral do DCC usando a estratégia de busca: “metadata standard” “research data” policy, encontramos o resultado de 108 notícias, 63 eventos e 64 blogs. Constatamos que essa busca não fornecia os resultados que buscávamos.

Optei então por explorar a página navegando pelos seus “links”.

No DCC navegando pelos links, seguimos o caminho: Home » Guidance » Metadata » Disciplinary Metadata (Digital Curation Centre 2022b)

Na página inicial temos a aba Guidance, que ao clicar nela abre vários links e ao selecionarmos o link Metadata (Figura 6).

Figura 6 - Guidance DCC



Fonte: dcc.ac.uk

Abre uma página com links para metadados disciplinares (Figura 7) e para Estrutura de padrões DCC DIFFUSE (O DCC DIFFUSE Standards Frameworks foi desenvolvido em parceria com várias organizações com o objetivo de apresentar estruturas pesquisáveis de padrões relevantes para a curadoria e preservação digital).

Figura 7 - Metadados disciplinares DCC

Fonte: dcc.ac.uk

Optamos por navegar na página de metadados disciplinares (Figura 8).

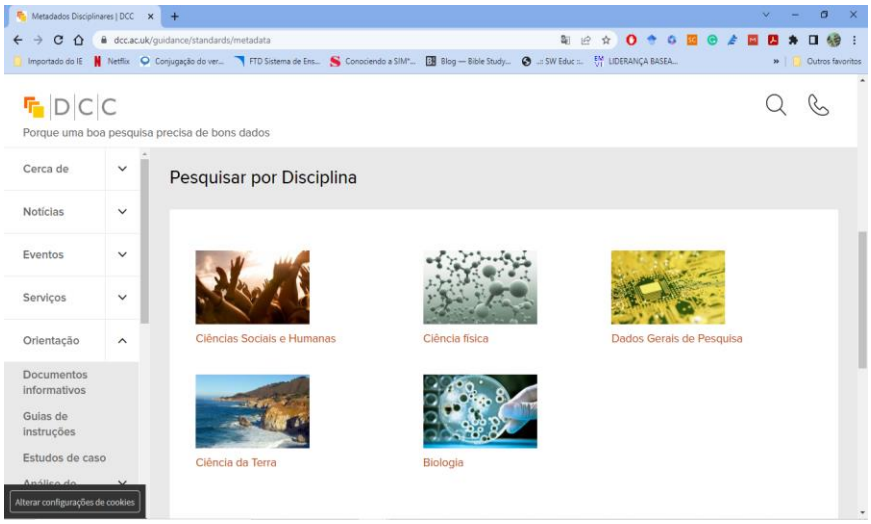
Figura 8 - Metadados Disciplinares/Disciplinas DCC

Fonte: dcc.ac.uk

Ao clicarmos no link para “metadados disciplinares” encontramos uma área com ícones para as seguintes disciplinas:

Ciências Sociais e humanas; Ciência física; Dados Gerais de Pesquisa; Ciência da Terra e Biologia (Figura 9).

Figura 9 – Pesquisa por Disciplina DCC



Fonte: dcc.ac.uk

Ao clicarmos no link Dados Gerais de Pesquisa (Figura 10) temos acesso a uma lista de disciplinas que são: Antropologia; Bioquímica; Cartografia; Química; Gestão de Dados; Ecologia; Ciência Ambiental; Genética Geral; Genômica; Geologia; História; Hidrogeologia; Hidrologia; Gestão da Informação; Meteorologia; Biologia Molecular Multidisciplinar; Administração Pública; Sensoriamento Remoto; Política Social; Ciência do Solo; Estatística; Teologia e Estudos Religiosos e Topografia.

Figura 10 – Dados gerais de pesquisa

Fonte: dcc.ac.uk

Ao clicarmos em cada link das disciplinas encontramos informações sobre os padrões de Metadados usados em sistemas de cada uma destas áreas. A seguir apresenta-se o quadro com a lista de disciplinas e seus respectivos padrões de metadados indicados.

Quadro 3 – Padrões de metadados/Disciplinas DCC

DISCIPLINA	PADRÕES DE METADADOS
ADMINISTRAÇÃO PÚBLICA	DCAT - Data Catalog Vocabulary
ANTROPOLOGIA	OAI-ORE - Open Archives Initiative Object Reuse and Exchange

BIOLOGIA MOLECULAR	PDBx/mmCIF – Protein Data Bank Exchange Dictionary and the Macromolecular Crystallographic Information Framework e Repository-Developed Metadata Schemas
BIOQUÍMICA	Repository-Developed Metadata Schemas
CARTOGRAFIA	Repository-Developed Metadata Schemas
CIÊNCIA AMBIENTAL	MIBBI - Minimum Information for Biological and Biomedical Investigations e Repository-Developed Metadata Schemas
CIÊNCIA DO SOLO	Repository-Developed Metadata Schemas
DISCIPLINA GERAL	PEMIS; PROV e QuDEx - Qualitative Data Exchange Format
ECOLOGIA	Repository-Developed Metadata Schemas

ESTATÍSTICAS	RDF Data Cube Vocabulary
GENÉTICA	Genome Metadata
GENÔMICA	Genome Metadata e Repository-Developed Metadata Schemas
GEOLOGIA	ABCD - Access to Biological Collection Data e Observations and Measurements
GESTÃO DA INFORMAÇÃO	DCAT - Data Catalog Vocabulary
GESTÃO DE DADOS	DataCite Metadata Schema e DCAT - Data Catalog Vocabulary
HIDROGEOLOGIA	ISO 19115 e Repository-Developed Metadata Schemas
HIDROLOGIA	DIF - Directory Interchange Format e Repository-Developed Metadata Schemas
HISTÓRIA	OAI-ORE - Open Archives Initiative Object Reuse and Exchange
METEOROLOGIA	Observations and Measurements

MULTIDISCIPLINAR	Data Packaged Core Datasets e OpenAIRE
POLÍTICA SOCIAL	SDMX - Statistical Data and Metadata Exchange
QUÍMICA	NeXus; Observations and Measurements e PDBx/mmCIF – Protein Data Bank Exchange Dictionary and the Macromolecular Crystallographic Information Framework
SENSORIAMENTO REMOTO	Observations and Measurements
TEOLOGIA E ESTUDOS RELIGIOSOS	BAV - Biblioteca Apostolica Vaticana
TOPOGRAFIA	Repository-Developed Metadata Schemas

Fonte: Elaboração própria

No que se refere a metadados disciplinares, o DCC reuniu um conjunto de disciplinas com links para informação sobre padrões de metadados disciplinares,

incluindo perfis, ferramentas para implementar os padrões e casos de uso de repositórios de dados.

Nos diferentes padrões disciplinares encontramos um conjunto de lista de Padrões de metadados gerais, que são os seguintes:

Quadro 4 – Padrão de metadados gerais/DCC

NOME	DESCRIÇÃO
CERIF - Common European Research Information Format. (https://eurocris.org/services/main-features-cerif)	CERIF é o padrão que a UE recomenda aos seus estados membros para registrar informações sobre atividades de pesquisa.
Data Package, (https://frictionlessdata.io/)	A especificação Data Package é um formato wrapper genérico (é um programa que extrai o conteúdo de uma fonte de informação específica e o converte em um formato relacional) para troca de dados, consistindo em uma pasta contendo arquivos de dados e um arquivo descritor.
DataCite Metadata Schema. (https://datacite.org/)	Um conjunto de metadados obrigatórios que devem ser registrados no DataCite Metadata Store ao criar um identificador persistente DOI para um conjunto de dados.

DCAT - Data Catalog Vocabulary. (https://www.w3.org/TR/vocab-dcat-2/)	DCAT é um vocabulário RDF projetado para facilitar a interoperabilidade entre catálogos de dados publicados na Web. Que serve para descrever catálogos e conjuntos de dados – datasets.
Dublin Core. (https://www.dublincore.org/ https://pt.wikipedia.org/wiki/Dublin_Core)	O Dublin Core é um padrão de metadados, composto por 15 elementos, planejado para facilitar a descrição de recursos eletrônicos, foi feito incialmente para descrever documentos digitais. É um dos padrões de metadados mais conhecidos e mais amplamente usados.
OAI-ORE - Open Archives Initiative Object Reuse and Exchange. (http://www.openarchives.org/ore/)	A Open Archives Initiative desenvolve e promove padrões que permitem a distribuição de conteúdo. A Iniciativa surgiu das necessidades dos domínios de acesso aberto e repositório institucional, mas agora promove o trabalho para fornecer acesso a recursos digitais para bolsa de estudos, pesquisa e aprendizado.
Observations and Measurements. (https://www.ogc.org/standards/om)	É um padrão internacional que define uma codificação de esquema conceitual para observações e para recursos envolvidos na amostragem ao fazer observações. O padrão O&M foi desenvolvido no contexto de sistemas de informações geográficas.

PREMIS. (http://www.loc.gov/standards/premis/)	O dicionário de dados PREMIS (Preservation Metadata: Implementation Strategies) define um conjunto de metadados que a maioria dos repositórios de objetos digitais precisaria registrar e usar para preservar esses objetos a longo prazo. Tem suas raízes no modelo de referência do sistema de informação arquivística aberta, mas foi fortemente influenciado pela experiência prática de tais repositórios.
PROV. (http://www.w3.org/2001/sw/wiki/PROV)	A Família de Documentos PROV define um modelo, serializações correspondentes e outras definições de suporte para permitir o intercâmbio interoperável de informações de proveniência em ambientes heterogêneos como a Web.
RDF Data Cube Vocabulary. (https://www.w3.org/TR/vocab-data-cube/)	O vocabulário Data Cube é uma base fundamental que suporta vocabulários de extensão para permitir a publicação de outros aspectos de fluxos de dados estatísticos ou outros conjuntos de dados multidimensionais.

Fonte: Elaboração própria

Alguns repositórios decidiram que os padrões atuais não atendem às suas necessidades de metadados e, portanto, criaram seus próprios requisitos. São denominados de “Repository-Developed Metadata Schemas”.

O DCC alinhou três esforços de descrição de políticas de dados, tornando mais fácil criar políticas de dados alinhadas ao FAIR e tornar as descrições de políticas mais acessíveis para humanos e máquinas.

O registro da política de dados FAIRsharing, a lista de verificação da política de dados FAIRsFAIR e os recursos da política de periódicos da RDA foram todos alinhados e integrados ao modelo de dados FAIRsharing.

Como resultado dessa colaboração, todos os campos da lista de verificação e os recursos da política endossados pela RDA no escopo do FAIRsharing estão disponíveis nos registros da política FAIRsharing. (https://blog.fairsharing.org/?p=451)

Isso cria um “fluxo de trabalho” da política de dados FAIR que consta das seguintes etapas:

a) verificar se o documento de política está na Lista de verificação da política de dados FAIR;

b) depósito do documento de política simultaneamente no Zenodo e no repositório institucional da instituição e atribuição de DOI a cada um destes documentos depositados;

c) registrar o documento desta política no FAIRsharing e informar os DOI tanto do Zenodo quanto do repositório institucional.

Esse processo ajuda a criar políticas de dados alinhadas ao FAIR, pois os metadados do FAIR estarão acessíveis tanto para humanos (através da publicação de políticas e da criação do registro FAIRsharing) quanto para máquinas (através da API FAIRsharing). Os resultados serão políticas de dados mais fáceis de encontrar, acessíveis e reutilizáveis.

4.2.3 RDA

A Research Data Alliance (RDA) através dos seus grupos de trabalho e grupos de interesse, divulga os seus resultados com o intuito de incentivar à gestão de dados. A RDA é uma organização internacional, que engloba participantes de todas as áreas e de diferentes nacionalidades.

A RDA foi lançada como uma iniciativa comunitária em 2013 pela Comissão Europeia, pela Fundação Nacional de Ciência do Governo dos Estados Unidos e pelo Instituto Nacional de Padrões e Tecnologia e pelo Departamento de Inovação do governo australiano com o objetivo de construir uma infraestrutura social e técnica para permitir a partilha aberta e a reutilização de dados (https://www.rd-alliance.org/about-rda).

Ao usar a pesquisa geral da RDA usando a estratégia de busca: “metadata standard” “research data” policy, encontramos 16.764 resultados, trazendo notícias, “post” de blogs, eventos, documentos, biografia etc. Resultado muito vasto que não especifica diretrizes e nem políticas de metadados.

No entanto, a RDA apresenta uma lista significativa de recomendações fornecidas pelos grupos de trabalho ou grupos de interesse da RDA. Optamos por seguir essa lista. As recomendações são os resultados oficiais e endossados da RDA e considerados os seus principais resultados (https://www.rd-alliance.org/recommendations-outputs/standards).

Na RDA são quatro grupos que foram criados para padrões de metadados. São eles: MIG (Metadata Interest Group), MSDWG (Metadata Standards Directory WG); DICIG (Data in Context IG) e RDPIG (Research Data Provenance IG).

Quadro 5 – Grupos padrões de metadados RDA

NOME do Grupo de Interesse	**DESCRIÇÃO do Grupo de Interesse**	**PADRÃO proposto pelo Grupo de Interesse**
MIG (Metadata Interest Group – Grupo de interesse em metadados) (https://www.rd-alliance.org/mig-metadata-interest-group.html)/ Atualizado para MASDIR WG (Metadata Directory Working Group), (https://www.rd-alliance.org/masdir-metadata-standards-directory-wg-update.html)	Grupo criado para discutir metadados e trabalhar com Grupos de Trabalho (GTs), realizando tarefas específicas. Como descrever padrões de metadados, dados de pesquisa e política ou diretrizes. O grupo foi atualizado para MASDIR WG.	Metadata Standards Directory WG
MSDWG (https://www.rd-alliance.org/groups/metadata-standards-directory-working-group.html)	Grupo desenvolve um diretório de padrões de metadados para que um usuário possa procurar criar padrões apropriados para sua finalidade e/ou domínio de pesquisa.	Metadata Standards Directory WG

DICIG (https://www.rd-alliance.org/groups/data-context-ig.html)	Grupo Desenvolvendo através de casos de uso os requisitos dentro e entre domínios de pesquisa para metadados contextuais.	Metadata Standards Directory WG
RDPIG (https://www.rd-alliance.org/groups/research-data-provenance.html)	Concentrando-se em fornecer informações de proveniência para conjuntos de dados. Esses grupos surgiram espontaneamente "bottom-up", mas agora estão se coordenando entre si para formar uma forte presença de metadados em RDA.	Metadata Standards Directory WG

Fonte: Elaboração própria

MIG (Metadata Interest Group), MSDWG (Metadata Standards Directory WG); DICIG (Data in Context IG) e RDPIG (Research Data Provenance IG) se uniram com a 'tarefa' de produzir um diretório de metadados, o Metadata Standards Directory WG.

Os grupos de metadados RDA concentram-se em todos os aspectos de metadados para dados de pesquisa, incluindo descoberta de dados, contextualização, validação, processamento analítico e interoperação.

Os metadados não são importantes apenas para documentar dados (incluindo direitos, proveniência, bem como as informações descritivas usuais) e avaliar conjuntos de dados quanto à relevância e qualidade (que inclui conjunto de dados e citação de publicação) para o reuso em questão; também é necessário para a Interoperabilidade dos dados de pesquisa.

O Catálogo de Metadados Disciplinares desenvolvido pelo Centro de Curadoria Digital do Reino Unido (DCC) foi lançado em janeiro de 2013, mais ou menos quando o Grupo de Trabalho do RDA estava sendo planejado.

Foi concebido como um recurso que os curadores de dados institucionais poderiam consultar ao aconselhar os pesquisadores sobre como eles deveriam documentar seus dados. O pensamento era que tais curadores gostariam, primeiro, de saber quais padrões estão em uso dentro da disciplina em questão. Se não houvesse nenhum ou muito poucos, eles poderiam querer saber sobre padrões mais amplos que poderiam ser adaptados.

O Grupo de Trabalho RDA e o DCC entraram em colaboração para usar o catálogo como ponto de partida para o Diretório de Padrões de Metadados (BOLA, 2016).

Foi criado o Metadata Standards Catalog (https://rdamsc.bath.ac.uk/). O Catálogo de Padrões de Metadados RDA é um diretório aberto e colaborativo de padrões de metadados aplicáveis a dados de pesquisa. É oferecido à comunidade acadêmica internacional para ajudar a enfrentar os desafios de infraestrutura.

Fairsharing, DCC e RDA trabalham em conjunto para a criação de padrões de metadados alinhados ao FAIR.

Com a criação do Metadata Standards Catalog o pesquisador ou a instituição pode pesquisar qual o padrão da lista que aparece no catálogo se adequa melhor a sua área ou campo de conhecimento.

Quanto a diretrizes internacionais gerais fizemos uma pesquisa no buscador do Google usando a estratégia de busca: “metadata standard” “research data” policy. Constatamos, com isso, que realmente só temos padrões para determinada área ou campo de conhecimento e não diretrizes internacionais gerais.

Nos padrões visto nos quadros acima encontramos com bastante frequência o padrão para catálogo de Dados, o DCAT, que é recomendado pelo W3C, que incorpora vários metadados de proveniência.

O W3C- O World Wide Web Consortium é a principal organização de padronização da World Wide Web. Consiste em um consórcio internacional com 450 membros, agrega empresas, órgãos governamentais e organizações independentes com a finalidade de estabelecer padrões para a criação e a interpretação de conteúdo para a Web.

4.3 ANÁLISE DO QUESTIONÁRIO

Nessa seção vamos descrever como aplicamos a metodologia que foi descrita na seção 3 para a análise do questionário. Identificando a percepção dos usuários quanto aos dados de proveniência, caracterizando-a do ponto de vista dos objetivos, procedimentos técnicos, o ambiente e amostra, a coleta de dados, instrumento de coleta e os procedimentos para análise dos dados.

A pesquisa é de natureza aplicada, com abordagem mista em relação ao problema de pesquisa. Conforme Gil (2012, p. 147), a abordagem mista "também pode ser utilizada quando o pesquisador deseja definir grupos de acordo com os resultados quantitativos e fazer o seu acompanhamento mediante pesquisa qualitativa".

A considerar que a pesquisa objetiva buscar conhecer os sentimentos e pontos de vista que podem alterar a maneira como os pesquisadores se comportam quanto aos dados de proveniência essenciais para o reuso de dados de pesquisa em COVID-19, a abordagem qualitativa na análise dos resultados mostra-se adequada, pois a "pesquisa qualitativa é usada para obter insights sobre os sentimentos e pensamentos das pessoas, o que pode fornecer a base para um futuro estudo qualitativo independente ou pode ajudar os pesquisadores a mapear instrumentos de pesquisa para uso em um estudo quantitativo" (SUTTON; AUSTIN, 2015, p. 226).

Do ponto de vista dos objetivos é do tipo exploratória e descritiva. Exploratória durante a coleta e tratamento de dados, no sentido de ampliar as percepções sobre impacto de pesquisa, e descritiva no propósito de descrever os fatores de impactos percebidos pelos pesquisadores quanto aos dados de proveniência essenciais para o reuso de dados de pesquisa em COVID-19 (GIL, 2012).

Perguntas fechadas foram utilizadas para identificar e posteriormente analisar as respostas recebidas.

4.3.1 Coleta de dados e análise dos dados

O corpus da pesquisa escolhido foi um conjunto pesquisadores da Fiocruz que tiveram seus projetos aprovados em editais específicos para trabalhar com COVID-19.

Selecionamos dois editais. O primeiro edital foi do Programa Inova Fiocruz Ideias e Produtos Inovadores – COVID-19 Encomendas Estratégicas (https://portal.fiocruz.br/edital-ideias-e-produtos-inovadores-COVID-19-encomendas-estrategicas). O objetivo do edital foi apoiar propostas voltadas para a pandemia da COVID-19 que pudessem trazer ações, decisões e respostas rápidas.

O segundo edital foi do Programa Inova Fiocruz Geração de Conhecimento - Enfrentamento da Pandemia e Pós-Pandemia COVID-19 Encomendas Estratégicas (https://portal.fiocruz.br/edital-geracao-de-conhecimento-enfrentamento-da-pandemia-e-pos-pandemia-COVID-19-encomendas). O edital selecionou projetos interdisciplinares que abordaram lacunas importantes para a compreensão de questões relevantes para o combate à pandemia no Brasil e no mundo, gerando conhecimento de forma original com potencial para transformar decisivamente o entendimento, a forma ou a conduta em relação à pandemia bem como período pós pandêmico.

Com esses dados foi elaborada uma planilha Excel com os títulos dos projetos, os nomes dos pesquisadores coordenadores, unidade da Fiocruz e endereço de e-mail.

A etapa seguinte consistiu na extração da planilha os dados referentes ao nome pessoal e e-mail dos pesquisadores. Com a identificação dos pesquisadores através do levantamento chegou ao quantitativo de 135 coordenadores dos projetos aprovados.

O questionário eletrônico foi o instrumento de coleta de dados selecionado para a obtenção do objetivo, pois o uso deste possibilita longo alcance geográfico, e atinge um grande número de pessoas, proporciona flexibilidade temporal para os respondentes, bem como a padronização de respostas (GIL, 2012; VIEIRA, 2009).

Uma versão preliminar do questionário foi apresentada para um grupo de 2 pesquisadores.

O pré-teste obteve como resultado 1 resposta, nas quais não foram relatados constrangimentos na abordagem ou outras dificuldades, mas houve sugestões que foram acatadas buscando a melhoria do instrumento da coleta de dados. Sugestões em relação ao problema do *link* que leva ao formulário do google, a melhoria na questão do termo de consentimento e na questão da idade sugestão em colocar faixas etárias.

O questionário foi aplicado através do envio por plataforma de correio eletrônico para os 135 coordenadores identificados no levantamento.

O primeiro envio foi realizado no dia 03 de janeiro de 2023 e apresentou retorno automático de 5 e-mails que “não puderam ser entregues porque o endereço não foi encontrado ou não pode receber mensagens”. Os e-mails retornados foram verificados para checagem de erros de grafia ou digitação. Após esse procedimento, foram novamente encaminhados para os pesquisadores, sendo que 3 e-mails retornaram como indisponíveis.

No dia 12 de janeiro foi feito novo envio de e-mails eliminando os e-mails que responderam na primeira vez, 3 pesquisadores, com o prazo de resposta até o dia 19 de janeiro. Com o resultado de mais 5 pesquisadores respondendo o questionário somando então 8 questionários respondidos.

No dia 01 de fevereiro foi feita mais uma tentativa com o prazo de resposta até o dia 08 de fevereiro. Para Freitas, Janissek-Muniz e Mascarola (2004, p. 7), "um dos problemas frequentes [com questionários on-line] é que diversos endereços eletrônicos atribuídos não estão mais ativos”. Outra questão foi que o envio do questionário se deu em período de início de ano, quando muitos pesquisadores estão de férias.

O questionário aceitou respostas até 10 de fevereiro de 2023, totalizando 18 respostas, cerca de 23,4%.

O questionário constava de 23 perguntas fechadas e abertas e foi dividido em 5 partes com objetivos distintos de levantar e identificar: 1. Informações sobre o Repositório FAPESP para descobrir a familiaridade dos pesquisadores sobre Repositório de dados de pesquisa; 2. Informações sobre os dados que os pesquisadores consideram relevantes; 3 informações sobre o Produtor de dados como perfil acadêmico, instituição etc. 4 Informações sobre o repositório onde os dados estão disponibilizados como garantia de preservação e *links* de identificador persistente e 5 Informações sobre o perfil do pesquisador que respondeu o questionário com perguntas sobre instituição, idade, escolaridade etc.

As respostas do questionário foram recebidas em planilha do Google Forms e em seguida foram exportadas para o Excel para organização, tabulação e codificação dos dados. Às respostas foram aplicadas análises estatísticas para atestar a confiabilidade e a conformidade da distribuição dos dados.

Os resultados foram agrupados e analisados de acordo com a divisão acima explicitada e foram descritos nas subseções abaixo. O perfil dos pesquisadores em relação a formação acadêmica revelou o resultado da maioria dos pesquisadores possuírem o pós-doutorado 50% seguido por 44,6% com doutorado.

Com o objetivo de caracterizar a amostra, foi levantado o perfil do pesquisador, quanto ao sexo, formação acadêmica, carreira e faixa etária.

Tabela 1 – Perfil dos respondentes

Pergunta	Perfil	Percentual
Qual o seu sexo?	Feminino	44,4%
	Masculino	55,6%
Qual a sua faixa etária?	31 – 40	27,8%
	41 – 50	33,3%
	51 – 60	22,2%
	61 – 70	11,1%
	Acima de 70	5,6%
Qual a sua formação acadêmica?	Mestrado	5,6%
	Doutorado	44,4%
	Pós-doutorado	50,0%
Qual a categoria que ocupa na carreira docente ou de pesquisa?	Pesquisador	45,3%
	Pesquisador/Professor permanente	38,8%
	Pesquisador/Professor colaborador	5,3%
	Bibliotecário	5,3%
	Pesquisador com bolsa de produtividade	5,3%

Fonte: Dados obtidos das respostas às questões 8, 9, 10 e 11 do questionário

O resultado apontou que a maioria dos pesquisadores que respondeu ao questionário é do sexo masculino, correspondendo a 55,6%. Os pesquisadores do sexo feminino correspondem a 44,4%.

Esse resultado está de acordo com os pesquisadores levantados na etapa da busca dos editais em COVID-19. A planilha geral trouxe o resultado de 130 coordenadores.

Destes coordenadores 53% são do sexo masculino e 47% do sexo feminino.

Quanto a faixa etária dos pesquisadores que responderam ao questionário tivemos o resultado de 33,3% na faixa etária dos 41-50 anos e de 27,8% na faixa etária de 31-40 anos. Um pesquisador apareceu na faixa etária acima dos 70 anos correspondendo a 5,6% na análise do questionário.

O perfil dos pesquisadores em relação a formação acadêmica revelou o resultado da maioria dos pesquisadores possuírem o pós-doutorado 50% seguido por 44,6% com doutorado.

Quanto a carreira docente ou de pesquisa a maioria 45,3% informou serem pesquisadores, com 38,8% informando serem professores permanentes e pesquisadores; tanto pesquisador e professor colaborador, bibliotecário e pesquisador com bolsa de produtividade apareceram com 5,3% para cada.

Quanto ao nome do laboratório, departamento ou unidade de pesquisa onde os pesquisadores atuam tivemos o seguinte resultado demostrado no quadro a seguir:

Tabela 2 – Laboratório, Departamento ou unidade

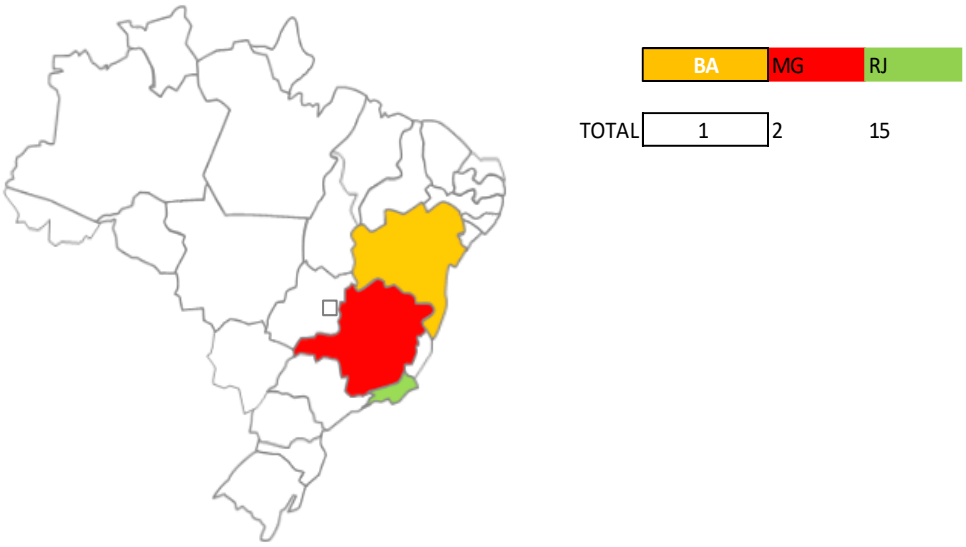
NOME LABORATÓRIO, DEPARTAMENTO, UNIDADE	TOTAL
CDTS / Centro de Desenvolvimento Tecnológico em Saúde	1
Gestão do Conhecimento e Prospecção em Saúde/ Instituto de Tecnologia em Fármacos/Farmaguinhos	1
Grupo integrado de Pesquisas em Biomarcadores/ FIOCRUZ MINAS	1
INI/ Instituto Nacional de Infectologia	1
Instituto de Tecnologia em imunológicos/ Bio-Manguinhos/Fiocruz	1
Instituto Oswaldo Cruz	1
Instituto René Rachou/ FIOCRUZ MINAS	1
Laboratório Avançado de Saúde Pública / Fiocruz-BA	1
Laboratório de Comunicação e Saúde/ LACES/ICICT	1
Laboratório de Informação em Saúde/ ICICT	5
Laboratório de produtos naturais para saúde/ Farmanguinhos/ Fiocruz.	1
Plataforma de Nível de Biossegurança 3 /	1
Instituto Oswaldo Cruz	
Síntese de peptideos e Oligonucleotideos/ IOC	1
Departamento de Direitos humanos, Saúde e Diversidade	1
Cultural -- DIHS-ENSP/FIOCRUZ	
Total Geral	**18**

Fonte: Dados obtidos da questão 12 do questionário

A unidade da Fiocruz com maior adesão ao questionário foi o ICICT com 41,6% dos respondentes.

Além dos pesquisadores lotados no campus sede da Fiocruz no Rio de Janeiro, obtivemos resposta de pesquisadores das Unidades Fiocruz Minas e Fiocruz Bahia.

Gráfico 1 – Brasil/Estado/Laboratórios Fiocruz



Fonte: Araujo, 2023.

Verificamos se os pesquisadores utilizaram dados de pesquisa de outros repositórios, principalmente do repositório nacional criado na temática, o Repositório COVID-19 DataSharing/BR. Este repositório é uma iniciativa da FAPESP em parceria com a Universidade de São Paulo, Grupo Fleury, Hospital Sírio- Libanês, Hospital das Clínicas da FMUSP e da Beneficência Portuguesa de São Paulo e Hospital Israelita Albert Einstein. Disponibiliza dados relacionados à COVID-19 e conta com dados de quase meio milhão de pacientes, e dezenas de milhões de registros de exames clínicos e desfechos (https://repositoriodatasharingfapesp.uspdigital.usp.br/).

Para exemplificar o tipo de dados que encontramos no repositório descreveremos o conjunto de dados do repositório *datasharing* da FAPESP com o título de: Dados COVID-19 beneficência Portuguesa de São Paulo. Em uma pasta de arquivos encontramos dados anonimizados de pacientes que fizeram teste para COVID-19 a partir de 11/01/2019, compactada em formato .zip.

O conjunto de dados possui 3 arquivos em formato csv e um arquivo em formato xlsx: (1) Planilha com dados anonimizados sobre pacientes que fizeram teste para o COVID-19 (sorologia ou PCR), incluindo: identificador anonimizado do paciente, gênero, país, estado , município e região de residência; (2) Respectivos resultados de

exames laboratoriais, incluindo entre outros o identificador anonimizado do paciente e um identificador de atendimento; (3) Desfechos - cada registro inclui entre outros o identificador anonimizado do paciente e um identificador de atendimento; descreve um atendimento de um paciente, e o resultado correspondente, quando aplicável; e (4) Dicionário de dados: planilha em que cada aba descreve respectivamente todos os campos das planilhas de Pacientes, Exames e Desfechos.

As planilhas Pacientes, Exames e Desfechos são interligadas pelo identificador anonimizado do paciente. As planilhas Exames e Desfechos são interligadas pelo par (identificador do paciente, identificador do atendimento). (https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/101?show=full)

Foi verificado o uso desta plataforma, repositório FAPESP, pelos pesquisadores da Fiocruz. Dentre os respondentes nenhum dos pesquisadores utilizou o repositório.

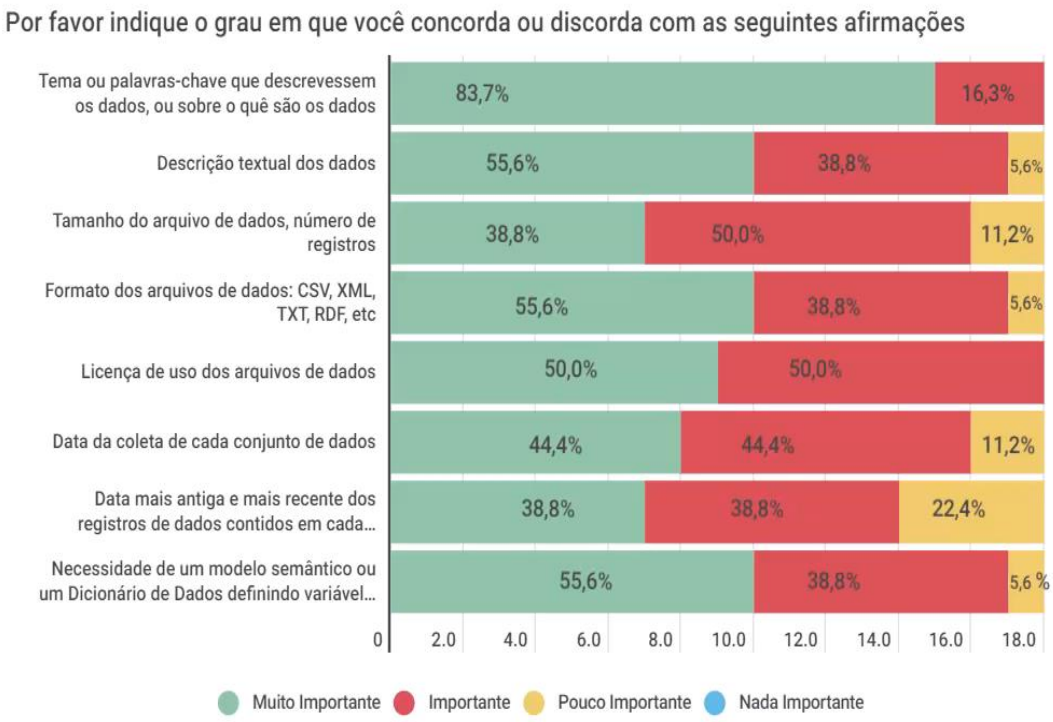
Uma limitação da pesquisa é que ela não se aprofundou para verificar os motivos do não uso da plataforma, porém, pode se inferir que provavelmente haja um desconhecimento da plataforma pelos pesquisadores consultados. Outra possibilidade é que as características das pesquisas dos respondentes, atuantes no campo da saúde são das perspectivas das Ciências Sociais e humanas.

Outras pesquisas futuras podem investigar com pesquisadores que trabalham diretamente com dados clínicos a importância e o uso dos dados do Repositório FAPESP.

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 8 questões para identificar os quesitos que os respondentes consideram como mais importante.

Gráfico 2 – Informações sobre os dados



Fonte: Dados obtidos da questão 4 do questionário

Verificou-se que 83,7% dos respondentes indicaram ser muito importante a descrição do tema ou palavras-chave que descrevessem os dados ou sobre o quê são os dados. Consideram como importante 16,3% dos respondentes.

Nenhum respondente indicou que fosse pouco importante, demonstrando a importância de tal descrição. Esta posição demonstra alinhamento com os princípios FAIR, pois, para que os dados sejam localizáveis (princípio *Findable*) o tema ou palavras-chave precisam estar bem descrito.

Em relação à importância da descrição textual dos dados de pesquisa a resposta foi de 55,6% considerando como muito importante, 38,8% como importante e 5,6% como pouco importante. Cabe ressaltar que a descrição textual dos dados é um fator facilitador para a localização dos dados.

Quanto à importância do tamanho do arquivo de dados e número de registros foi constatado que 50% consideram importante, seguido por 38,8% que consideram muito importante e 12,2% que consideram pouco importante.

Quanto à importância dos formatos dos arquivos de dados: em CSV, XML, TXT, RDF etc. 55,6% dos respondentes consideram muito importante, 38,8% importante e 5,6% pouco importante.

Ter os conjuntos de dados disponíveis em um formato acessível após o término do seu projeto de pesquisa aumentará as possibilidades de reutilização deles, o que está relacionado a um dos princípios FAIR (princípio *Reusable*).

Quanto à importância da licença de uso dos arquivos de dados 50,0% considera muito importante e 50,0% como importante.

A importância da licença de uso dos arquivos de dados é um fator que está em consenso com os princípios FAIR. A abertura dos dados está presente no movimento da ciência aberta e os princípios FAIR tratam estes dados como inteligentemente abertos. Isto significa que estarão livremente disponíveis, desde que a sua privacidade e reutilização sejam preservadas e estejam condicionadas às licenças específicas, além dos créditos reconhecidos e das citações devidamente adotadas (SCIENTIFIC ELECTRONIC LIBRARY ONLINE, 2016).

Os direitos de uso ligados aos dados é um fator importante, pois os direitos poderão limitar severamente a reutilização dos dados, o que tem a ver com o princípio de reuso (princípio Reusable), especialmente em função da necessidade de cumprimento das restrições de licenciamento.

Quanto à importância da data coleta de cada conjunto de dados 44,4% dos respondentes consideram como muito importante, o mesmo número dos que consideram como importante e 12,2% consideram como pouco importante.

A Informação referente a data da coleta de cada conjunto de dados foi vista em sua maioria como importante, porém alguns respondentes não consideram esse um fator importante para o reuso dos dados de pesquisa.

Quanto a importância da informação da data mais antiga e mais recente dos registros de dados contidos em cada arquivo 38,8% dos respondentes consideram como muito importante, o mesmo número dos que consideram como importante e 22,4 consideram como pouco importante.

O que se aproxima da questão anterior também referente a data dos dados.

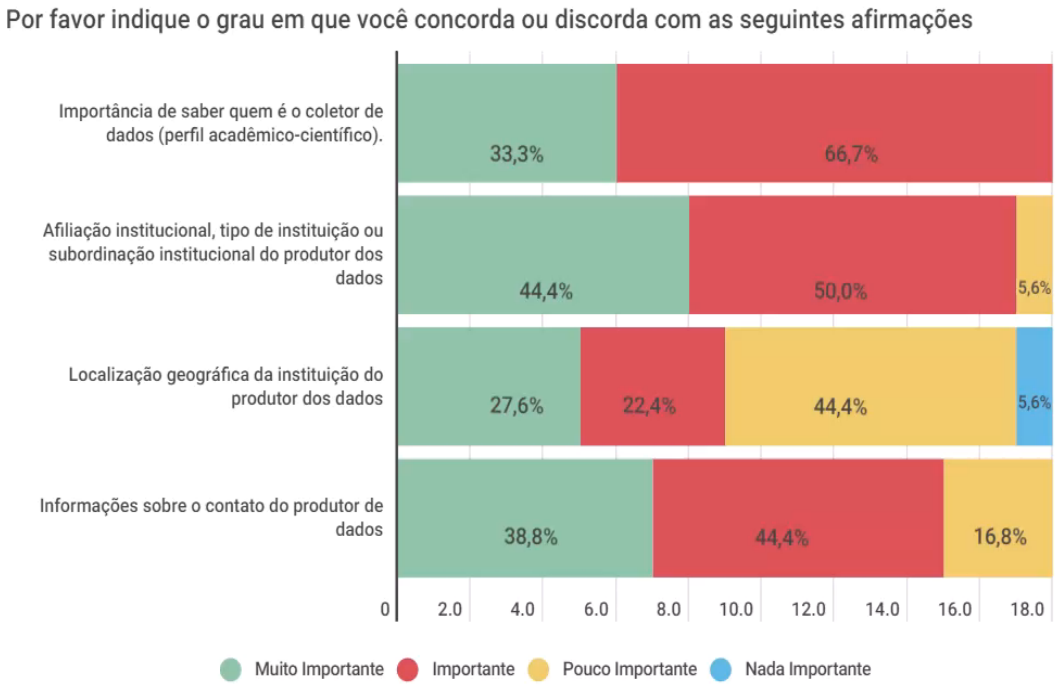
Quanto a importância da necessidade de um modelo semântico ou um Dicionário de Dados definindo variável, descrição, formato e conteúdo esperado 55,6% dos respondentes consideram como muito importante, 38,8% como importante e 5,6% como pouco importante.

Para garantir a interoperabilidade dos conjuntos de dados, um dos princípios FAIR (princípio *interoperable*), é fundamental a utilização de vocabulários controlados,

ontologias e tesauros, o que foi percebido como importante pela maioria dos respondentes.

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.

Gráfico 3 – Informação sobre o produtor de Dados



Fonte: Dados obtidos da questão 5 do questionário.

Quanto à importância de saber quem é o coletor de dados (perfil acadêmico-científico) 66,7% dos respondentes consideram como importante e 33,3% consideram como muito importante.

Segundo os princípios FAIR os dados precisam ser reusáveis (princípio *Reusable*), para isso saber a proveniência é fundamental como demonstrado nas respostas.

Quanto à importância de saber sobre a afiliação institucional ou subordinação institucional do produtor de dados 50% dos respondentes consideram importante, 44,4% consideram com muito importante e somente 5,6% consideram pouco importante.

A questão da afiliação institucional se enquadra também no princípio de reuso e pode reforçar de acordo com o status da instituição o reuso dos dados de pesquisa.

Quanto à importância de saber a localização geográfica da instituição do produtor dos dados 44,4% dos respondentes consideram como pouco importante, 27,6% consideram como muito importante, 22,4% consideram importante e 5.6% consideram nada 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.

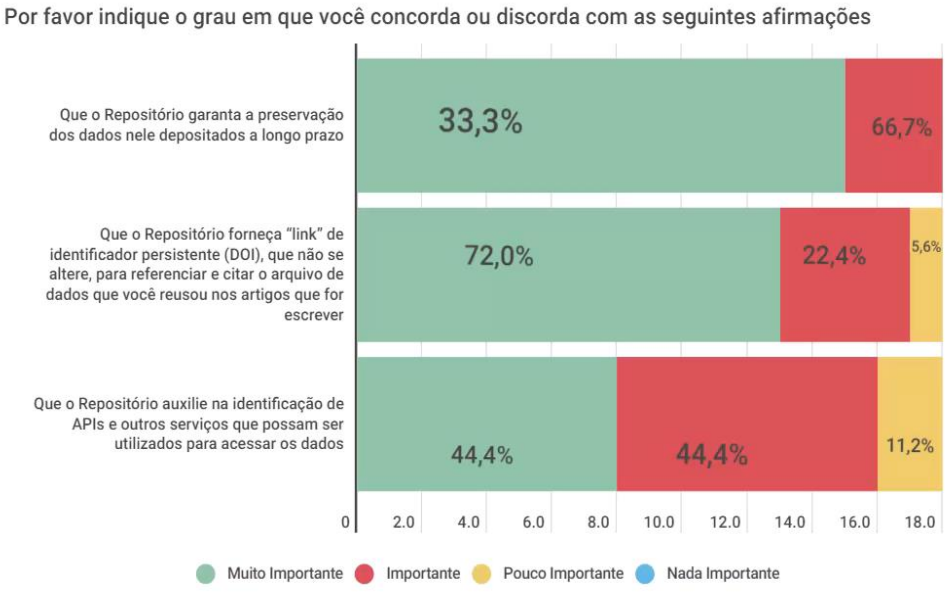
A questão foi a única que recebeu a resposta de nada importante por parte de um respondente.

Quanto à importância de saber informações sobre o contato do produtor de dados 44,4% dos respondentes consideram como importante, 38,8% como muito importante e 16,8% como pouco importante.

Essa questão proporciona no reuso dos dados uma averiguação direta ao produtor dos dados em caso de dúvida e necessidade de mais informação.

A informação sobre os repositórios de dados é importante para uso e reuso de dados e precisa ser otimizada consonante os princípios FAIR (WILKINSON, 2016).

Gráfico 4 - Informações sobre o Repositório onde os dados estão disponibilizados



Fonte: Dados obtidos da questão 6 do questionário

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% consideram como importante. O princípio FAIR da acessibilidade (princípio A*cessible*) requer que os dados estejam disponíveis por um longo tempo. Os conjuntos de dados tendem a degradar-se ou a desaparecer completamente tornado 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% consideram importante e 5,6% consideram como pouco importante. Para atender o princípio FAIR de dados localizáveis (princípio Findable) deve ser atribuído aos dados, dentre outros quesitos, um identificador globalmente exclusivo e persistente.

Para o princípio FAIR de interoperabilidade (princípio *Interoperable*) o vocabulário controlado usado para descrever conjuntos de dados também precisa ser documentado, usando identificadores globais únicos e persistentes.

Essa questão demonstra que para a maioria dos respondentes os dados que são recuperáveis pelo seu identificador persistente (DOI) é de grande valor para o reuso dos dados de pesquisa.

Quanto à questão que avalia a importância de que o repositório auxilie na identificação de APIs e outros serviços para identificar os dados 44,4% dos respondentes consideram como muito importante, também 44,4% consideram como importante e 11,2% como pouco importante.

API é a sigla em inglês para Application Programming Interface, ou interface de programação de aplicações. O uso de APIs permite a comunicação com outros dados facilitando o princípio FAIR da interoperabilidade.

Foi formulada também uma questão aberta sobre outras informações relevantes para a reutilização dos dados, onde obtivemos as seguintes afirmações:

Quadro 6 – Informações Relevantes

Quais outras informações seriam relevantes para sua decisão em reutilizar um arquivo de dados?
Descrição clara sobre a obtenção e processamento dos dados.
Citação e Referenciamento dos dados, dos registros e do sistema de dados em pesquisas
Objetivo original da coleta dos dados
Nada a acrescentar

Os dados sejam FAIR
A existência de um suporte técnico de recurso imediato, como um chat, porque nem todo pesquisador é experiente com a prática de fazer recurso a repositórios, principalmente com os procedimentos de busca de cada repositório.
Ser de fácil manuseio, estar em um servidor seguro, que os dados possam ser baixados ser interrupções, a formatação das colunas da planilha seja de fácil entendimento, disponibilidade das informações respectivas à amostragem original, principalmente data da coleta e data do início dos sintomas (para os sintomáticos).
Facilidade de acesso aos dados; disponibilidade de ferramentas básicas de análise.
Confiabilidade e rastreabilidade
Informar se houve alteração ou atualização do referido banco e quando isso ocorreu.

Fonte: Dados obtidos da questão 7 do questionário.

Destaco a consideração por parte de um respondente sobre a possibilidade da "existência de um suporte técnico de recurso imediato, como um chat, porque nem todo pesquisador é experiente com a prática de fazer recurso a repositórios, principalmente com os procedimentos de busca de cada repositório" (respondente 6). Algo que implementado ajudaria bastante aos pesquisadores.

Outra possibilidade, a meu ver, viria da capacitação dos pesquisadores através de cursos e atualizações para o uso dos repositórios.

Outro pesquisador pontuou a importância do repositório "ser de fácil manuseio, estar em um servidor seguro, que os dados possam ser baixados ser interrupções, a formatação das colunas da planilha seja de fácil entendimento, disponibilidade das informações respectivas à amostragem original, principalmente data da coleta e data do início dos sintomas (para os sintomáticos)" (respondente 7). Isto certamente facilitaria ao pesquisador no tratamento e reuso dos dados.

Essas considerações práticas explanadas pelos pesquisadores visam facilitar o reuso dos dados de pesquisa alinhadas aos princípios FAIR.

Como sugestão para estimular a utilização de repositórios de dados de pesquisa, seria interessante uma maior divulgação de ferramentas como tutoriais, cursos, atualizações etc.

Com a análise das respostas dos pesquisadores em relação ao reuso dos dados científicos pelos princípios FAIR verificamos que a informação sobre "Tema ou palavras-chave que descrevessem os dados, ou sobre o quê são os dados" foi a que teve o maior percentual de respostas "Muito importante". Nenhum respondente considerou esse fator como sendo pouco importante ou nada importante.

As questões sobre *informação sobre licença de uso*, sobre o *coletor de dados* e sobre o *repositório garantir a preservação* dos dados nele depositado a longo prazo

tiveram a avaliação por parte dos respondentes como sendo muito importante e importante.

No quadro seguinte temos uma síntese dessas questões comparando com os princípios FAIR correspondentes. Esta comparação é importante porque os princípios FAIR destacam o reuso, ligada à nossa questão de pesquisa, dos metadados de proveniência.

Quadro 7– Síntese/Questões e princípios FAIR

Questões consideradas como muito importantes ou importante por todos os respondentes	**Princípios FAIR**
1. Tema ou palavra-chave	Apoia o Findable. os metadados são descritos fazendo uso de metadados enriquecidos
2. Licença de uso	Apoia o Reusable. Os metadados devem ser liberados com licenças de uso de dados claras e acessíveis.
3 Quem é o coletor de dados?	Apoia o Reusable. Os metadados devem estar associados à sua proveniência.
4. Que o repositório garanta a preservação dos dados nele depositado a longo prazo	Apoia o Acessible. Os metadados devem ser acessíveis, mesmo quando os dados não estão mais disponíveis

Fonte: Araujo, 2023

Percebe-se que há um consenso por parte dos respondentes nas questões como: tema ou palavra-chave; licença de uso, coletor de dados e preservação dos dados a longo prazo por parte dos repositórios, apesar de outras questões serem percebidas como importante para uma grande parte dos respondentes.

5 CONSIDERAÇÕES FINAIS

Esta pesquisa teve como objetivo identificar metadados de proveniência fundamentais para o reuso de dados no contexto da pesquisa em COVID-19 alinhados aos princípios FAIR.

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 e na pesquisa documental, o que pode comprometer o alinhamento destes dados aos princípios FAIR. A maioria destes repositórios possui um esquema de metadados próprio.

Em seguida, foi feito um mapeamento das diretrizes internacionais para padrões de metadados para dados de pesquisa. nas iniciativas internacionais Fairsharing (https://fairsharing.org/), Digital Curation Centre (DCC) (https://www.dcc.ac.uk/) e Research Data Alliance (RDA) (https://www.rd-alliance.org/).

Encontramos no Fairsharing as diretrizes Genomic Science Program e a RDA COVID-19 WG Resources.

No DCC nos diferentes padrões disciplinares encontramos um conjunto de lista de Padrões de metadados gerais como: CERIF; Data Package; DataCite Metadata Schema; DCAT, Dublin Core; OAI-ORE - Open Archives Initiative Object Reuse and Exchange; Observations and Measurements; PREMIS; PROV; RDF Data Cube Vocabulary e o “Repository-Developed Metadata Schemas”, criados por alguns repositórios que decidiram que os padrões atuais não atendiam às suas necessidades de metadados.

Na RDA encontramos quatro grupos que foram criados para padrões de metadados. São eles: MIG (Metadata Interest Group), MSDWG (Metadata Standards Directory WG); DICIG (Data in Context IG) e RDPIG (Research Data Provenance IG).

Com as respostas dos pesquisadores pode-se perceber que eles não estão familizarizados com o reuso dos dados. Sendo interessante o uso do questionário para aplicabilidade em área como epidemologia, ou área em que os pesquisadores estão mais habituados a utilizarem dados de pesquisa para verificação de divergências ou concordâncias com as percepções dos pesquisadores que responderam o questionário.

Por conta da limitação de tempo a pesquisa teve como foco os coordenadores de projetos na área de Covid 19 financiados pela Fiocruz, Essa limitação suscita a necessidade de investigar a questão no âmbito de outras instituições de financiamento brasileiras, o que ampliaria ainda mais a noção da percepção dos pesquisadores quanto ao reuso dos dados de pesquisa no campo da saúde

Ressalto a relevância do questionário, tanto pelo seu ineditismo, pelo menos no Brasil, quanto da sua replicabilidade para mais pesquisadores.

Os resultados encontrados, não só na pesquisa bibliográfica e documental, mas especialmente no questionário, embora parciais (poucos questionários, uma única instituição), demonstram como a questão da proveniência é importante para encorajar o reuso dos dados de pesquisa.

REFERÊNCIAS

ALVES, R. C. V. **Metadados como elementos do processo de catalogação**. Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 132 f. 2010. Disponível em: http://hdl.handle.net/11449/103361. Acesso em: 28 jul. 2022.

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.

ARAUJO, Anderson Silva de. **Dados de dissertação** [Data set]. 2023. Disponível em: https://doi.org/10.5281/zenodo.7643810. Acesso em: 17 fev. 2023.

BÍBLIA. Português. **Bíblia sagrada NVI**. São Paulo: Sociedade Bíblica Internacional, 2001.

BOLA, A. **Diretório de padrões de metadados**. RDA Europe Webinar, Finlândia, 2016. https://researchportal.bath.ac.uk/en/publications/metadata-standards-directory

BURLINGAME, N; NIELSEN, L. **A Simple Introduction to Data Science**. Wickford: New Street Communications, 2012.

CASTRO, F. F. de; 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**: Estudos, João Pessoa, v. 17, n. 2, p. 13-19, maio/ago. 2007.

CHAN, L. M.; ZENG, M. L. Metadata interoperability and standardization: a study of methodology part i: achieving interoperability at the schema level. **D-Lib Magazine**, v. 12, n. 6, jun. 2006. Disponível em: http://www.dlib.org/dlib/june06/chan/06chan.html . Acesso em: 28 jul. 2022.

CUNHA, Murilo Bastos da. A biblioteca universitária na encruzilhada. DataGramaZero: **Revista de Ciência da Informação**, v. 11, n. 6, dez. 2010. Disponível em: https://repositorio.unb.br/handle/10482/14869. Acesso em: 20 dez. 2022.

CURTY, R. **Para onde os dados devem ir afinal?** Grupo de Trabalho da Rede de Dados de Pesquisa Brasileira (RDP Brasil), 2018. Disponível em: https://dadosdepesquisa.rnp.br/?author=14. Acesso em: 30 jul. 2022.

COSTA, Micheli; BRAGA, Tiago. Repositórios de dados de pesquisa no mundo. **Cadernos BAD**, n. 2, p. 80-95, 2016. Disponível em: https://brapci.inf.br/index.php/res/v/82195. Acesso em: 23 dez. 2022.

DAGLIATI, A.; et al. Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview. **Brief Bioinform**., v. 22, n. 2, p. 812-822, Mar.

2021. DOI: 10.1093/bib/bbaa418. Disponível em: https://pubmed.ncbi.nlm.nih.gov/33454728/. Acesso em 18 jul. 2022.

DAVENPORT, T. H. **Big Data at Work**: Dispelling the Myths, Uncovering the Opportunities. Boston: Harvard Business Review Press, 2014.

DAVIDSON, S. B.; FREIRE, J. Provenance and scientific workflows: Challenges and opportunities. **ACM SIGMOD international conference on Management of data**, pages 1345–1350, 2008.

DAVIS, H. M.; VICKERY, J. N. Datasets, a shift in the currency of scholarly communication: Implications for library collections and acquisitions. **Serials Review**, Greenwich, v. 33, n. 1, p. 26-32, 2007.

DEMCHENKO, Y. et al. **Addressing big data issues in Scientific Data Infrastructure**. Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013, p. 48–55, 2013. Disponível em: https://www.academia.edu/4307681/Addressing_Big_Data_Issues_in_Scientific_Data_Infrastructure. Acesso em: 22 jul. 2022.

Digital Curation Centre. **Disciplinary Metadata**. 2022 a. Disponível em: http://www.dcc.ac.uk/resources/metadata-standards. Acesso em: 30 dez. 2022.

Digital Curation Centre. **Disciplinary Metadata**. 2022 b. Disponível em: https://www.dcc.ac.uk/resources/subject-areas/general-research-data. Acesso em: 30 dez. 2022.

DRUCKER, D.P. et al. Implantação da Rede Temática GO-FAIR Agro Brasil: Primeiros Passos. **Anais** do XIII Congresso Brasileiro de Agroinformatica. 2021. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/228163/1/PL-Implantacao-Rede-GO-FAIR-SBIAgro-2021.pdf. Acesso em: 03 fev. 2023.

DUDZIAK, E. A. **Dados de Pesquisa agora devem ser armazenados e citados**. 2016. Disponível em: http://www.sibi.usp.br/?p=6189. Acesso em: 20 dez. 2022.

DUDZIAK, E. A. **Gestão de dados de pesquisa**: o que precisamos saber hoje! 2018. Disponível em: http://www.sibi.usp.br/noticias/gestao-de-dados-de-pesquisa-o-que-precisamos-saber-hoje/. Acesso em: 28 jul. 2022.

FANIEL, I.; JACOBSEN, T. E. Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data, **Computer Supported Cooperative Work (CSCW)**, v. 19, n. 3/4, p. 355-375, 2010. Disponível em: https://doi.org/10.1007/s10606- 010-9117-8. Acesso em: 30 ago. 2022.

FEDERER, L. Research data management in the age of big data: Roles and opportunities for librarians. **Information Services & Use**, v. 36, n. 1-2, p. 35-43, 2016. Disponível em: http://content.iospress.com/articles/information-services-and-use/isu797. Acesso em: 20 dez 2022.

FORMENTON, D.; GRACIOSO, L. S.; CASTRO, F. F. Revisitando a preservação digital na perspectiva da ciência da informação: aproximações conceituais. **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, SP, v. 13, n. 1, p. 170-191, jan./abr. 2015. Disponível em: http://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/1587/1571. Acesso em: 21 jul. 2022.

FREUND, G. P.; SEMBAY, M. J.; DE MACEDO, D. D. J. Proveniência de Dados e Segurança da Informação: relações interdisciplinares no domínio da Ciência da Informação. **Revista Ibero-Americana de Ciência da Informação**, [S. l.], v. 12, n. 3, p. 807–825, 2019. DOI: 10.26512/rici.v12.n3.2019.21203. Disponível em: https://periodicos.unb.br/index.php/RICI/article/view/21203. Acesso em: 31 ago. 2022.

GIL, Carlos, A. **Como Elaborar Projetos de Pesquisa**. 6ª edição. São Paulo, Atlas, 2017.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Editora Atlas S.A, 2012.

GOLD, Anna. Cyberinfrastructure, data, and libraries, part 1: a cyberinfrastructure primer for librarians. **D-Lib Magazine**, v. 13, n. 9-10, 2007. Disponível em: http://www.dlib.org/dlib/september07/gold/09gold-pt1.html. Acesso em: 22 jul. 2022.

GREEN, Ann; MACDONALD, Stuart; RICE, Robin. **Policy-making for research data in repositories**: a guide. Maio 2009. Disponível em: https://www.coar-repositories.org/files/guide.pdf. Acesso em: 22 jul. 2022.

HENDERSON, M. **Data Management**: a practical guide for librarians. Lanham: Rowman & Littlefield, 2017.

HENNING, P. C.; RIBEIRO, C. J. S.; SALES, L. F.; MOREIRA, L. R.; SANTOS, L. O. B. S. 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, 2019. DOI: 10.22478/ufpb.1981-0695.2019v14n3.46969. Disponível em: https://brapci.inf.br/index.php/res/v/150613. Acesso em: 29 jul. 2022.

HEY, Tony; HEY, Jessie. E-science and its implications for the library community. **Library Hi Tech**, v. 24, n. 4, p. 515-528, 2006.

HJØRLAND, B. Data (with big data and database semantics). **Knowledge Organization**, Baden-Baden, v. 45, n. 8, p. 685-708, 2018. Disponível em: https://www.isko.org/cyclo/data#top. Acesso em: 08 jul. 2022.

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICES AND TECHNOLOGY (IASSIST). Defining data librarian - call for

comments, 2006. Disponível em: https://iassistdata.org/blog/2006/06/12/defining-data-librarian-call-comments/. Acesso em: 20 dez. 2022.

JENSEN, Howard E. 1950 'Editorial note', in **Through Values to Social Interpretation**: Essays on Social Contexts, Actions, Types, and Prospects by Howard Paul Becker. Durham, NC: Duke University Press, vii–xi. Disponível em: https://archive.org/stream/in.ernet.dli.2015.234024/2015.234024.Through-Values_djvu.txt. Acesso em: 20 jul. 2022.

KOLTAY, Tibor. Data literacy for researchers and data librarians. **Journal of Librarianship and Information Science**, v. 49, n. 1, 2017. Disponível em: http://journals.sagepub.com/doi/abs/10.1177/0961000615616450. Acesso em: 20 dez. 2022.

LEITE, F. C. L. **Como gerenciar a visibilidade da informação científica brasileira**: repositórios institucionais de acesso aberto. IBICT: Brasília. 2009. 120 p. il.

MARCONDES, C. H. **O papel dos vocabulários no *Big Data***: o desafio dos dados de pesquisa. Apresentação no IV Seminário do Grupo de Pesquisa MHTX – 9 nov. 2021, ECI/UFMG. Diponível em: https://drive.google.com/file/d/1e4kKipUFasLSFIARcCwuSi5vsVWYGKKK/view. Acesso em: 20 jul.2022.

MARCONDES, C. H.; RAMOS JUNIOR, M. A. C.; MARTINS, S. C. O papel dos vocabulários no acesso e reuso dos Big Data. **Informação & Informação**, Londrina, v. 26, n. 4, p.146-174. 2021.

MARCONDES, C.H.; SAYÃO, L.H. Documentos digitais e novas formas de cooperação entre sistemas de informação em C&T. **Ci Inf**, Brasília, 2002; v. 31, n. 3, p. 42-54. 2002.

MARTINS, H.C.; PERLIN, M.S. Call for papers data reuse: what new information can we learn from used data?. **Revista de Administração Contemporânea**. Zenodo, 29 jun. 2020. http://doi.org/10.5281/zenodo.3858031. Disponível em : https://zenodo.org/record/3858031#.Yuwq-3bMK00. Acesso 10 jul. 2022.

MEDEIROS, C.B. Gestão de Dados Científicos – da coleta à preservação [online]. **SciELO em Perspectiva**, 2018. Disponível em: https://blog.scielo.org/blog/2018/06/22/gestao-de-dados-cientificos-da-coleta-apreservacao/. Acesso em: 11 nov. 2019

MICHENER, W, K. Ten Simple Rules for Creating a Good Data Management Plan. **PLOS Computational Biology**. 2015. Disponível em: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004525. Acesso em: 26 jul. 2022.

NASSI-CALÒ, L. Promovendo e acelerando o compartilhamento de dados de pesquisa [online]. **SciELO em Perspectiva**, 2019. Disponível em: https://blog.scielo.org/blog/2019/06/13/promovendo-e-acelerando-o-compartilhamento-de-dados-de-pesquisa/. Acesso em: 11 mai. 2023.

NATIONAL INFORMATION STANDARDS ORGANIZATION (NISO). **Data Dictionary**: technical metadata for digital still images. Baltimore: NISO, c2011. 104 p. Disponível em: http://www.niso.org/apps/group_public/download.php/14698/z39_87_2006_r2011.pdf . Acesso em: 26 jul. 2022.

NATIONAL INSTITUTES OF HEALTH. **NIH Data Sharing Policy and Implementation Guidance**, 2003. Disponível em: https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm. Acesso em: 21 dez. 2022.

NATIONAL RESEARCH COUNCIL (NRC). **A question of balance**: Private rights and the public interest in scientific and technical databases, 1999. Disponível em: http://www.nap.edu/read/9692/chapter/1. Acesso em: 20 dez. 2022.

NATIONAL SCIENCE BOARD (NSB). **Digital Research Data Sharing and Management**, 2011. Disponível em: https://www.nsf.gov/nsb/publications/2011/nsb1124.pdf. Acesso em: 20 dez. 2022.

ORGANIZATION for Economic Co-operation and Development (OECD). **Declaration on Access to Research Data from Public Funding Organization for Economic Co-operation and Development** (OECD), 2004. Disponível em: https://www.oecd.org/sti/inno/38500813.pdf. Acesso em: 20 dez. 2022.

PAMPEL, H. et al. **Making research data repositories visible**: the re3data.org Registry. PLoS One, San Francisco, v. 8, n. 11, 2013. Disponível em: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3817176/. Acesso em: 21 dez. 2022.

PIWOWAR, H. A.; VISION, T. J. Data reuse and the open data citation advantage. **Peer J**, San Diego, v. 1, n.175, out. 2013. DOI: https://doi.org/10.7717/peerj.175. Disponível em: https://peerj.com/articles/175/. Acesso em: 26 jul. 2020.

PIORUN, M. E. **E-Science as a Catalyst for Transformational Change in University Research Libraries**. 2013. 369f. Tese (Doutorado em Filosofia) – Faculty of the Simmons College Graduate School of Library and Information Science. University of Massachusetts Medical School, 2013. Disponível em: http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1151&context=lib_articles. Acesso em: 20 dez. 2022.

RICE, R.; SOUTHALL, S. **The data librarian's handbook**. London: Facet Publishing, 2016.

QUERALT-ROSINACH, N. et al. Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic. **J Biomed Semant**, v.13, n.12. 2022. DOI: https://doi.org/10.1186/s13326-022-00263-7. Disponível em: https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-022-00263-7. Acesso em: 10 jul. 2022.

SAYÃO, L. F.; SALES, L. F. **Guia de Gestão de Dados de Pesquisa para Bibliotecários e Pesquisadores**. Rio de Janeiro: CNEN/IEN, 2015. Disponível em: http://www.icb.usp.br/~sbibicb/images/guia%20gestaoPDF/Guia%20de%20gestao%20dados%20de%20pesquisa.pdf. Acesso em: 22 jul. 2022.

SCIENTIFIC ELECTRONIC LIBRARY ONLINE. Princípios reitores FAIR publicados em periódico do Nature Publishing Group. SciELO em **Perspectiva**, [S.l], 2016.

SEMELER, A. R.; PINTO, A. L. Os diferentes conceitos de dados de pesquisa na abordagem da biblioteconomia de dados. **Ciência da Informação**, [S. l.], v. 48, n. 1, 2019. Disponível em: https://revista.ibict.br/ciinf/article/view/4461. Acesso em: 20 dez. 2022.

SEMIDÃO, R. A. M. **Dados, Informação e Conhecimento enquanto elementos de compreensão do universo conceitual da Ciência da Informação**: contribuições teóricas. Marília, 2014. 198 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Informação, Faculdade de Filosofia e Ciências, Universidade Estadual Paulista – UNESP, Marília, 2014. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/110783/000799485.pdf?sequence=1. Acesso em: ago. 2017.

SCIENTIFIC ELECTRONIC LIBRARY ONLINE. Princípios reitores FAIRpublicados em periódico do Nature Publishing Group. **SciELO em Perspectiva**, [S.l], 2016

VAN DER AALST, W. M. P. Data Scientist: The Engineer of the Future. In: MERTINS, K. et al. (Eds.). **Enterprise Interoperability VI**, 13 Proceedings of the I – ESA Conferences 7, 2014.

VEIGA, Viviane Santos de Oliveira et al. **Plano de Gestão de Dados de Pesquisa - PGD**: guia de elaboração. Rio de Janeiro: Fiocruz/Icict, ago. 2022. 32 p.

VEIGA, V. S. de Oliveira. **Percepção dos pesquisadores portugueses e brasileiros da área de Neurociências quanto ao compartilhamento de artigos científicos e dados de pesquisa no acesso aberto verde**: custos, benefícios e fatores contextuais. 2017. 294 f. Tese (Doutorado em Ciências) -Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fundação Oswaldo Cruz, Rio de Janeiro, 2017.

VEIGA, Viviane Santos de Oliveira; QUEIROZ, Claudete Fernandes de. **Rede GO FAIR Brasil Saúde**: uma rede de apoio à Gestão e Abertura de Dados de Pesquisa em Saúde no Brasil. In: ENCONTRO IBÉRICO EDICIC, 9., Barcelona, 2019.

VEIGA, V. S. de O.; HENNING, P.; DIB, S.; PENEDO, E.; LIMA, J. da C.; SILVA, L. O. B. da; PIRES, L. F. Plano de gestão de dados fair: uma proposta para a Fiocruz. **Liinc em Revista**, [S. l.], v. 15, n. 2, 2019. DOI: 10.18617/liinc.v15i2.5030. Disponível em: https://revista.ibict.br/liinc/article/view/5030. Acesso em: 22 jul. 2022.

U.S. NATIONAL COMMITTEE FOR CODATA; COMMITTEE FOR A PILOT STUDY ON DATABASE INTERFACES. **Bits of power**: issues in Global Access to Scientific

Data. Washington, D.C.: National Academy Press, 1997. Disponível em: https://nap.nationalacademies.org/catalog/5504/bits-of-power-issues-in-global-access-to-scientific-data. Acesso em: 24 jul. 2022.

WELLCOME TRUST. **Sharing research data to improve public health: full joint statement by funders of health research**. 2012. Disponível em: https://wellcome.org/whatwe-do/our-work/sharing-research-data-improve-public-health-full-joint-statement-fundershealth. Acesso em: 20 jul. 2022.

WEITZEL, S. da R. **Os Repositórios de e-prints como nova forma de organização da produção científica**: o caso área das ciências da comunicação no Brasil. 2006. 360 f. Tese (Doutorado em Ciência da Informação). Universidade de São Paulo, São Paulo, 2006. Disponível em: http://www.teses.usp.br/teses/disponiveis/27/27151/tde-14052009-133509/publico/3787212.pdf. Acesso em: 27 jul. 2022.

WILKINSON, M. et al. The FAIR Guiding Principles for scientific data management and stewardship. **Sci Data 3**. 2016. DOI: https://doi.org/10.1038/sdata.2016.18. Disponível em: https://www.nature.com/articles/sdata201618. Acesso em: 15 jul. 2022.

WIKI. **Gerenciar Metadados**. IBICT. Disponível em: http://wiki.ibict.br/index.php/Gerenciar_metadados. Acesso em: 28 jul. 2022.

WHO. **Coronavirus disease (COVID-19) pandemic**. Disponível em: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. Acesso em: 20 jul. 2022.

WHO. **Global COVID-19 Clinical Platform Case Report Form (CRF) for Post COVID condition (Post COVID-19 CRF)**. Disponível em: https://iris.paho.org/handle/10665.2/54126. Acesso em: 22 jul. 2022.

XIA, J.; WANG, M. Competencies and Responsibilities of Social Science Data Librarians: An Analysis of Job Descriptions. **College & Research Libraries**, v. 75, n. 3. p. 362-388, 2014. Disponível em: http://crl.acrl.org/content/75/3/362.abstract. Acesso em: 4 ago. 2015.

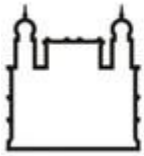
YIN, R. K. **Estudo de caso**: planejamento e métodos. 2.ed. Porto Alegre: Bookman, 2001.

YOON, Ayoung; LEE, Yoo Young. Factors of trust in data reuse. **Online Information Review**, 2019. Disponível em: https://scholarworks.iupui.edu/bitstream/handle/1805/23359/Yoon_2019_factors.pdf?sequence=1. Acesso em: 23 jun. 2022.

ZENG, M. L.; QIN, J. **Metadata**. New York: Neal-Schuman Publishers, 365 p. 2008.

ZHU, Y.; XIONG, Y. Towards Data Science. **Data Science Journal**, v. 14, p. 8, 2015. Disponível em: http://doi.org/10.5334/dsj-2015-008. Acesso em: 21 dez. 2022.

APÊNDICE A – Plano de Gestão de Dados



Ministério da Saúde

FIOCRUZ

Fundação Oswaldo Cruz

PLANO DE GESTÃO DE DADOS

Padrões de metadados de proveniência para reuso de dados de pesquisa em Covid 19 alinhados aos princípios FAIR

última atualização: 24/02/2023 - 14:23

ID: 328

Dados do Pesquisador Principal

Nome do pesquisador principal:

Anderson Silva de Araujo

E-mail do pesquisador principal:

anderson.araujo@icict.fiocruz.br

Link do currículo Lattes:

http://lattes.cnpq.br/3170428254350334

Link do identificador persistente:

https://orcid.org/0000-0001-5905-8213

Afiliação:

Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT)

Etapa do Plano de Gestão de Dados

Durante a execução do projeto

Resumo do projeto

A Organização Mundial da Saúde em 30 de janeiro de 2020, declarou o surto do coronavírus ou novo coronavírus, como emergência de saúde pública de interesse internacional e em 11 de março de 2020, como pandemia (WHO, 2020). Agências de saúde de todo o mundo produzem diagnósticos e tratamentos dos casos, com isso gerando dados importantes para o estudo da doença. Como Esses dados podem ser usados, compartilhados e reusados? Como Esses dados podem capacitar os pesquisadores a formular novos tipos de indagações, hipóteses no estudo de questões cruciais para a ciência e para a sociedade? Se os dados forem compartilhados de forma adequada tem um grande potencial para responder essas questões científicas e proporcionarem uma resposta adequada para a sociedade. A realização de pesquisas em saúde pública é demorada e cara. Garantir que os dados de pesquisa, juntamente com suas descobertas publicadas, sejam amplamente disponibilizados para a comunidade de pesquisa leva a mais descobertas e maior eficiência. Esta situação suscita as seguintes questões. Que informação é importante para que o pesquisador reutilize os dados? O que faz o pesquisador considerar se esses dados estão bem descritos ou não para reuso? Qual a importância de se conhecer quais os metadados de proveniência que são fundamentais para garantir que os pesquisadores venham a reutilizar esses dados em pesquisa sobre COVID-19? Responder essas perguntas são fundamentais. Diante do exposto esse projeto busca responder a seguinte questão: Quais os metadados de proveniência necessário para garantir o reuso dos dados de pesquisa?

Palavras-chave:

- Metadados
- Princípios FAIR
- Reuso de dados de pesquisa
- COVID-19
- Metadados de proveniência

Data de início do projeto:

21/01/2022

O projeto tem financiamento?

Sim

- CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

Descrição dos dados coletados ou reuso de dados existentes

Seus dados serão coletados e/ou Produzidos:

Sim

Explique quais metodologias ou softwares serão utilizados se novos dados forem coletados ou produzidos:

O estudo se caracteriza como pesquisa exploratória, descritiva, de caráter qualitativo. As pesquisas exploratórias pretendem observar e compreender os mais variados aspectos relativos ao fenômeno estudado pelo pesquisador. As pesquisas exploratórias mais comuns são os levantamentos bibliográficos. A pesquisa descritiva descreve uma realidade, visa descrever um determinado evento, realidade ou situação (GIL, 2012). Pesquisa qualitativa examina evidências baseadas em dados verbais e visuais para entender um fenômeno em profundidade. Portanto, seus resultados surgem de dados empíricos, coletados de forma sistemática. Nas pesquisas qualitativas, você pode utilizar entrevistas, grupos focais ou observações (GIL, 2012). Para alcançar os objetivos propostos foram realizados os seguintes passos metodológicos: • Para identificar os padrões de metadados utilizados para dados de pesquisa e mapear suas diretrizes internacionais (objetivos 1 e 2) foi realizado levantamento bibliográfico e documental, em duas etapas: Etapa 1 - Levantamento bibliográfico No contexto da revisão bibliográfica, segundo Lakatos e Marconi (2001, p. 183) a pesquisa bibliográfica “[...] abrange toda bibliografia já tornada pública em relação ao tema estudado, desde publicações avulsas, boletins, jornais, revistas, livros, pesquisas, monografias, teses, materiais cartográficos, etc.

[...]", proporcionando ao pesquisador contato direito com contribuições de diversos autores sobre determinado assunto. Para a pesquisa bibliográfica foram selecionadas fontes de informações internacionais e nacionais como: SCOPUS (via portal Capes), Medical Literature Analysis and Retrieval System Online (Medline) via PubMED, Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI) e a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD). O Google Acadêmico (Google Scholar) também será utilizado como fonte de informação complementar. Etapa 2 - Levantamento documental Esse procedimento distancia-se do caráter bibliográfico no que tange à natureza das fontes, pois "vale-se de materiais que não receberam ainda um tratamento analítico, ou que ainda podem ser reelaborados de acordo com os objetivos da pesquisa" (GIL, 2012, p. 51), como políticas, editais, relatórios, procedimentos, diretrizes, entre outros documentos dessa natureza. Para isso foi realizado um mapeamento das diretrizes internacionais para padrão de metadados para dados de pesquisa nas principais iniciativas internacionais como o Fairsharing (https://fairsharing.org/), o Digital Curation Centre (DCC) (https://www.dcc.ac.uk/) e o Research Data Alliance (RDA) (https://www.rd-alliance.org/). O FAIRSHARING é um recurso de suporte ao FAIR que fornece um registro informativo e educacional sobre padrões de dados, bancos de dados, repositórios e políticas, juntamente com ferramentas e serviços de pesquisa e visualização que interoperam com outros recursos que estão de acordo com os princípios FAIR. O FAIRsharing orienta os consumidores a descobrir, selecionar e usar padrões, bancos de dados, repositórios e políticas com confiança, e os produtores a tornar seus recursos mais detectáveis, mais amplamente adotados e citados. O GRUPO DO DCC para estudo dos padrões de metadados é um centro de especialização reconhecido internacionalmente em curadoria digital, com foco na criação de capacidades e habilidades para o gerenciamento de dados de pesquisa. A Research Data Alliance (RDA) é uma iniciativa para construir conexões técnicas e sociais para viabilizar o compartilhamento aberto de dados de pesquisa. A visão da RDA é tornar viável que pesquisadores possam compartilhar abertamente seus dados entre diferentes tecnologias, disciplinas e países, de forma a endereçar os grandes desafios da sociedade em escala global. • Para verificar quais os padrões de metadados de proveniência apoiam os pesquisadores no reuso dos dados em COVID-19 (Objetivo 3) foi elaborado instrumento de coleta de dados (questionário online). Etapa 3 – Pesquisa empírica. A pesquisa empírica serve para ancorar e comprovar no

plano da experiência aquilo apresentado conceitualmente, investiga um fenômeno contemporâneo dentro do seu contexto da vida real, especialmente quando os limites entre o fenômeno e o contexto não estão claramente definidos (YIN, 2001 p. 33). A observação e experimentação empíricas oferecem dados para sistematizar a teoria. Nesta etapa foi dada voz também ao ator principal que é o pesquisador, para que informe o que o predisporia a reusar dados de pesquisa. O corpus da pesquisa foi formado por pesquisadores que receberam financiamento, via edital da Fiocruz, para investigação em COVID-19, no período de 2020 a 2021. A Fiocruz é uma destacada instituição de ciência e tecnologia em saúde da América Latina reconhecida como importante na pesquisa no campo da COVID-19. É referência internacional em pesquisa no campo da saúde pública, presente fisicamente em todas as regiões do Brasil, e lidera no país os esforços mundiais contra o novo coronavírus. Também foram utilizados como fontes de informação para esta pesquisa os insumos e sugestões do Grupo GOFAIR Brasil Saúde e do projeto VODAN-Br, GT-Metadados e proveniência.

Seus dados serão reutilizados de outras fontes:

Não

Quais os tipos de dados que serão coletados, produzidos ou reutilizados?

- Imagem
- Numéricos - Textuais

Quais os formatos de dados que serão coletados, produzidos ou reutilizados?

- CSV - Comma-Separated Values
- PDF - Portable Document Format
- TXT - Text Format

Qual o volume aproximado dos dados coletado, produzidos ou reutilizados: Menor que 1 GB

Documentação e Qualidade dos Dados

Indique os metadados específicos, se houver:

Os metadados são: Padrão de metadados; COVID-19; Dados de pesquisa; reuso de dados de pesquisa; princípios FAIR; Metadados; FAIR.

Armazenamento e Backup durante o processo de pesquisa

Onde os dados serão armazenados:

- Dispositivos de Armazenamento Externos como pendrive
- Laptops
- Repositório de dados institucional da Fiocruz
- Serviços de Nuvem Privado

Requisitos Legais e éticos

A equipe tem ciência das normas relacionadas a avaliação ética na pesquisa:

Sim

A pesquisa envolve tratamento de dados em que se exija avaliação ética: Sim

O projeto de pesquisa está aprovado pelo órgão de avaliação ética: Sim

Indique o órgão responsável:

- CONEP/CEP - Comissão Nacional de ética em Pesquisa/Comitê de ética em Pesquisa

A equipe está familiarizada com a legislação sobre o tratamento e proteção dos dados pessoais:

Sim

A pesquisa envolve tratamento de dados pessoais:

Sim

O tratamento de dados pessoais está em conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD) - Lei nº 13.709/2018:

Sim

Há previsão de período de embargo relacionado ao direito autoral: Sim

A pesquisa envolve outras categorias de dados que exijam período de embargo para sua disponibilização: Não

Informe a licença a ser aplicada a base de dados:

CC-BY - Creative Commons Attribution 4.0 International

Compartilhamento de Dados e Preservação a longo prazo

A pesquisa prevê compartilhamento de dados:

Sim

Como e quando os dados serão compartilhados:

O compartilhamento dos dados foi realizado no repositório Zenodo, durante o processo da pesquisa.

Existem possíveis motivações para submeter os dados à períodos de embargo no compartilhamento de dados: Sim

Qual motivo para embargo:

- Outro

Outro motivo para embargo:

O conjunto de dados depositado no repositório Zenodo estará em acesso fechado até a aprovação da dissertação para validação da pesquisa e publicação de artigo científico para a Comunicação dos resultados da pesquisa.

Qual o tempo estimado de embargo:

6 meses

Plano de Gestão de Dados (PGD)

APÊNDICE B – ARTIGO

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

Anderson Silva de Araujo.

Mestrando Programa de Pós-Graduação em Informação e Comunicação em Saúde/FIOCRUZ.

ORCID: https://orcid.org/0000-0001-5905-8213.

E-mail: moranderson0182@gmail.com.

Viviane Santos de Oliveira Veiga

Programa de Pós-Graduação em Informação e Comunicação em Saúde, ICICT//Fundação Oswaldo Cruz (Fiocruz)

ORCID: https://orcid.org/0000-0001-8318-7912

E-mail: viviane.veiga@icict.fiocruz.br

Isabella Henrique Lima Pereira

Bolsista de iniciação científica na Fundação Oswaldo Cruz (Fiocruz). Mestranda do Programa de Pós-graduação em Ciência da Informação PPGCI/UFF.

ORCID: https://orcid.org/0000-0002-8463-0629.

E-mail: isabellalima@id.uff.br.

Mylena Cristhina Araujo de Oliveira.

Mestranda do Programa de Pós-graduação em Ciência da Informação, convênio IBICT/UFRJ.

ORCID: https://orcid.org/0000-0002-0637-9053

E-mail: mycristh@gmail.com

Resumo da proposta

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. Apresenta os padrões de metadados utilizados para representar dados em COVID-19 encontrados na literatura e os padrões e 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. A maioria destes repositórios possui um esquema de metadados próprio, é necessário um estudo mais aprofundado sobre os esquemas de metadados criados e seu alinhamento com os princípios FAIR e as necessidades da área.

Tipo de proposta

- Comunicação

Tema em que se enquadra a proposta

Repositórios digitais – institucionais, temáticos, de dados de investigação ou de património cultural; Gestão de Dados de Investigação e dados FAIR; Modelos e padrões de metadados

Palavras-chave

Padrões de metadados, COVID-19, dados de pesquisa, repositórios digitais.

Audiência

Bibliotecários, gestores de dados de investigação, profissionais de comunicação de ciência, gestores de tecnologias de informação (programadores, administradores de sistemas e gestores de tecnologias de informação).

CONTEXTUALIZAÇÃO

Com o surto do novo coronavírus várias organizações uniram esforços para impedir o seu avanço 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) e no Brasil, a Rede VODAN BR.

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 reutilização e a interoperabilidade de dados de pesquisa. Para que conjuntos de dados sejam FAIR, dados e metadados precisam estar alinhados a estes princípios.

Os metadados e os 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 (definições e usos dos metadados.

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, 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 para todos, e principalmente outros pesquisadores.

PADRÕES DE METADADOS PARA DADOS DE PESQUISA EM COVID-19

Para conhecer os metadados utilizados para representar dados para 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*, utilizando os seguintes termos: covid; interoperability; COVID-19;

interoperable; sars vírus; metadata standards; sars-cov-2; metadata; sars; medical records; coronavírus; coronavirus disease.
Destacou-se nesta literatura a presença de 1 padrão de metadados na Web of Science e 4 padrões de metadados na PubMed utilizados na pesquisa em COVID-19.

Na *Web of Science* apareceu 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* foi encontrado o Outbreak.info que 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 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.

Recommendations and Guidelines on data sharing são uma visão geral completa e abrangente de como compartilhar dados (e software de pesquisa) de várias disciplinas para informar a resposta a uma pandemia, juntamente com diretrizes e recomendações sobre o compartilhamento de dados nas atuais circunstâncias do COVID-19. É um documento longo (mais de 140 páginas), mas muito completo e bem estruturado.

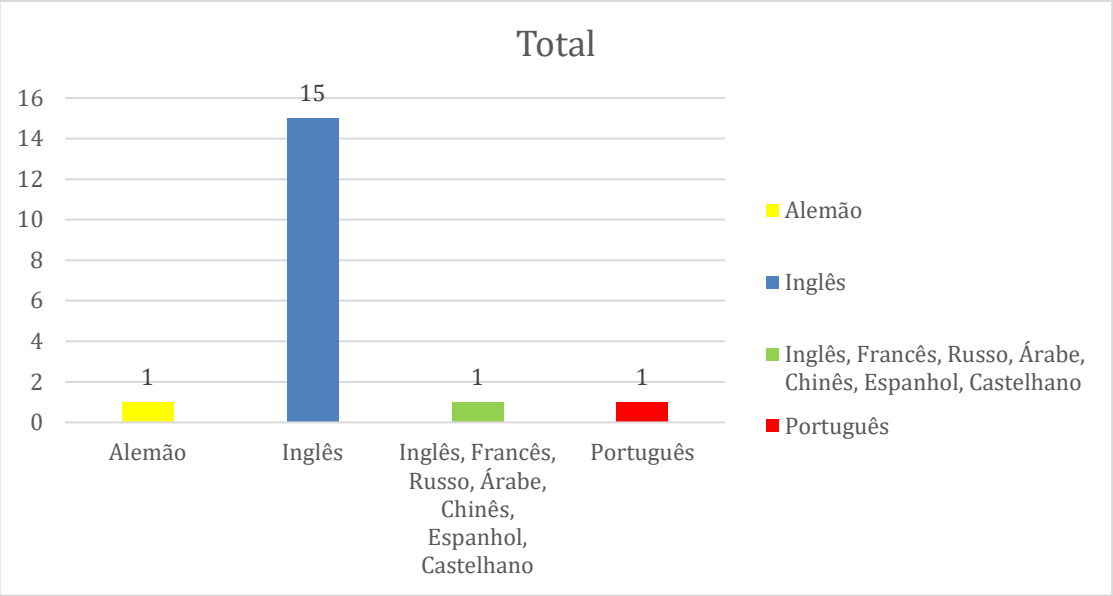
E o GISAID que tem como objetivo facilitar o compartilhamento de sequências do genoma viral e metadados clínicos e epidemiológicos relacionados para ajudar os pesquisadores a entender como os vírus evoluem e se espalham durante epidemias e pandemias.

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
Idioma dos repositórios

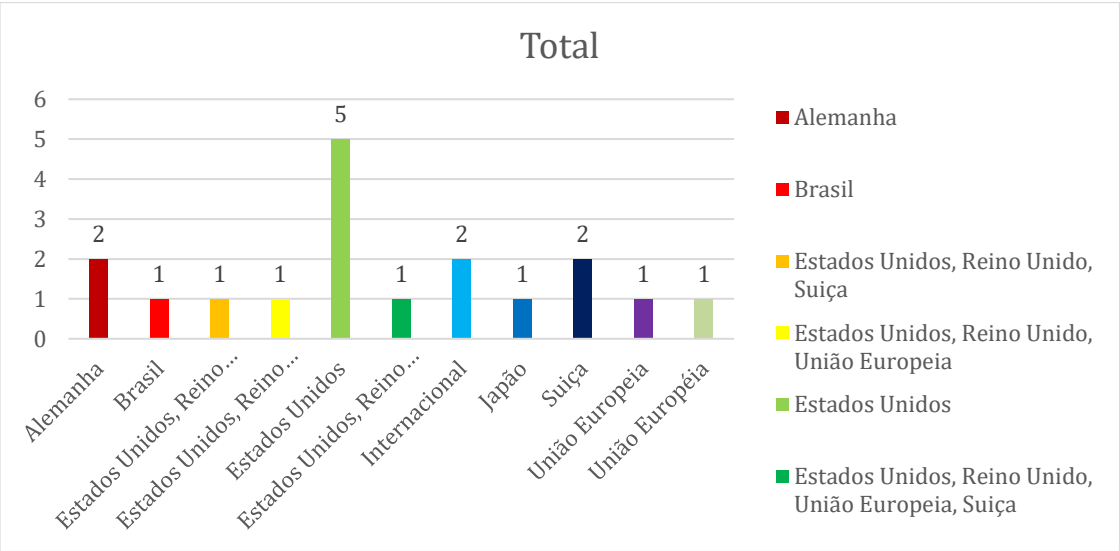


Fonte: Os autores, 2022.

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



Fonte: Os autores, 2022.

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

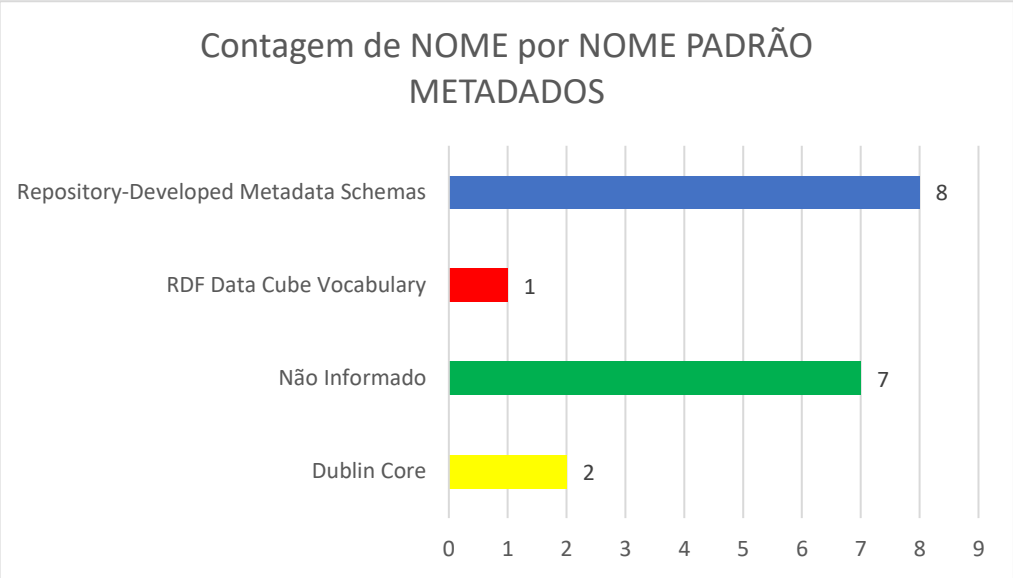
Nuvem de palavras

Fonte: Os autores, 2022.

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

Figura 4

Padrão de Metadados



Fonte: Os autores, 2022.

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 EM ANDAMENTO

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 se constatou que os repositórios analisados não adotaram os padrões de metadados específicos do campo da saúde, como os descritos no levantamento documental, 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 bibliográficas

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*, n. 23, v. 12, 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.

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

APÊNDICE C – Questionário

Questionário Roteiro para a entrevista de metadados de proveniência confiáveis.

REGISTRO DE CONSENTIMENTO LIVRE E ESCLARECIDO (RCLE)

Você está sendo convidado(a) como voluntário(a) a participar da pesquisa intitulada “Padrões de metadados de proveniência para reuso de dados de pesquisa em Covid-19”, a qual é realizada no âmbito do programa de pós-graduação em Informação e Comunicação em Saúde ICICT/FIOCRUZ, sob a responsabilidade de Anderson Silva de Araujo, sob orientação da prof. Viviane Veiga (PPGIS/Icict/Fiocruz) e Carlos Henrique Marcondes de Almeida (PPGCI/UFF). O convite se deve ao fato de você ter recebido financiamento da Fiocruz para realizar pesquisa sobre Covid- 19.

O estudo tem como objetivo geral identificar as percepções dos pesquisadores quanto ao reuso de dados de pesquisa em COVID-19, mais particularmente, quanto aos metadados de proveniência necessários para alcançar tal reuso.

Você está sendo convidado a participar porque seu e-mail institucional que consta na lista de projetos em COVID-19 aprovados nos editais da Fiocruz. Caso concorde em participar, solicito que dê seu aceite no RCLE e, em seguida, clique no link do Google Form quando terá acesso a um questionário com questões relacionadas a sua prática de registro dos dados de sua pesquisa. Responder ao questionário deverá lhe tomar cerca de 15 minutos. Não será solicitada sua identificação pessoal. É solicitado que você mantenha uma via do RCLE em seu poder.

Os riscos da pesquisa são considerados mínimos, relacionados a qualquer desconforto ao responder alguma pergunta. Você poderá optar por não responder a alguma pergunta específica, ou mesmo desistir de participar, a qualquer momento, sem qualquer prejuízo. Suas repostas serão anonimizadas e, a despeito dos riscos inerentes às pesquisas em ambientes virtuais, serão tomadas todas as precauções para manter sua confidencialidade. Todos os dados serão baixados para um ambiente seguro, e a eles só terão acesso o pesquisador e sua orientadora.

Os benefícios da pesquisa são, principalmente, indiretos, provendo conhecimento sobre a prática dos pesquisadores em relação ao uso de metadados que podem ser reutilizados em outras pesquisas, com o potencial de promover o avanço da ciência.

A qualquer momento, você poderá solicitar informação sobre a pesquisa, e poderá ter acesso aos resultados da mesma. Igualmente, você poderá ser ressarcido de qualquer dano decorrente da pesquisa, nos termos da lei.

Em caso de dúvidas sobre a pesquisa ou para relatar algum problema, você poderá contatar o pesquisador Anderson Silva de Araujo nos telefones: (21) 968335510, ou por e-mail: moranderson0182@gmail.com. Você também pode contatar o Comitê de Ética em Pesquisa da Escola Politécnica de Saúde Joaquim Venâncio/Fiocruz, no e-

mail **cep.epsjv@fiocruz.br**, no telefone 3865-9809, ou pessoalmente, no endereço Avenida Brasil, 4365 Térreo – Manguinhos. Rio de Janeiro, CEP 21040-360. O Comitê de Ética é a instância responsável por examinar os aspectos éticos das pesquisas que envolvem seres humanos, zelando pela proteção à dignidade, autonomia e direitos dos participantes.

Estamos realizando uma pesquisa sobre a reutilização/reaproveitamento de dados sobre COVID-19. Um exemplo dos dados sobre os quais estamos falando é o conjunto de dados do repositório datasharing da FAPESP com o título de: Dados COVID-19 beneficência Portuguesa de São Paulo https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/101?show=full

Se você dispusesse de um repositório (como o repositório de dados sobre COVID-19 da FAPESP, https://repositoriodatasharingfapesp.uspdigital.usp.br/) com diversos arquivos com dados sobre a sua especialidade de pesquisa ou seu interesse, prontos e disponíveis para "download", que informações você considera importantes estarem associadas a cada arquivo de dados para que você reutilizasse estes dados em suas pesquisas?

Responda as questões abaixo.

1. Você já trabalhou com os dados da FAPESP?

☐ Sim

☐ Não

Dados da FAPESP

2. Quais os dados da FAPESP você trabalhou?

_____

_____

_____

_____

_____

1. Informações sobre os DADOS:

	Nada Importante	Pouco Importante	Importante	Muito importante
Tema ou palavras-chave que descrevessem os dados, ou sobre o quê são os dados	□	□	□	□
Descrição textual dos dados	□	□	□	□
Tamanho do arquivo de dados, número de registros	□	□	□	□
Formato dos arquivos de dados: CSV, XML, TXT, RDF, etc	□	□	□	□
Licença de uso dos arquivos de dados	□	□	□	□
Data da coleta de cada conjunto de dados	□	□	□	□
Data mais antiga e mais recente dos registros de dados contidos em cada arquivo	□	□	□	□

2. Informações sobre o PRODUTOR DE DADOS, tais como:

	Nada Importante	Pouco Importante	Importante	Muito importante
Quem é o coletor de dados (perfil acadêmico-científico)	□	□	□	□
Afiliação institucional, tipo de instituição ou subordinação institucional do produtor dos dados	□	□	□	□
Localização geográfica da instituição do produtor dos dados	□	□	□	□

3. Informações sobre o REPOSITÓRIO onde os dados estão disponibilizados

	Nada Importante	Pouco Importante	Importante	Muito importante

que o Repositório garanta a preservação dos dados nele depositados a longo prazo	□	□	□	□
Que o Repositório forneça “link” de identificador persistente (DOI), que não se altere, para referenciar e citar o arquivo de dados que você reusou nos artigos que for escrever	□	□	□	□

4. Quais outras informações seriam relevantes para sua decisão em reutilizar um arquivo de dados?

_____

_____

_____

_____

_____

PERFIL

5. Qual o seu sexo?

- Masculino
- Feminimo

6. Qual a sua faixa etária?

- [ ] 21 - 30
- [ ] 31 - 40
- [ ] 41 - 50
- [ ] 51 - 60
- [ ] 61 - 70
- [ ] Acima de 70

7. Formação académica

Assinale, por favor, apenas o nível mais elevado

- [ ] Graduação
- [ ] Curso de especialização
- [ ] Mestrado
- [ ] Doutorado
- [ ] Pós-doutorado

8. Indique, por favor, a categoria que ocupa na carreira docente ou de pesquisa *

Selecione até duas opções

Marque todas que se aplicam.

- [ ] Professor permanente
- [ ] Professor visitante
- [ ] Professor Auxiliar
- [ ] Pesquisador com bolsa de produtividadePesquisador
- [ ] Assistente de PesquisaEstagiário de Pesquisa
- [ ] Outro: _____

9. Nome do seu laboratório, departamento ou unidade de pesquisa

_____

_____

_____

_____

_____

AGRADECIMENTO

Obrigado por contribuir para a realização dessa pesquisa a sua participação foi muito útil!