Datasets Curados e Enriquecidos com Proveniência da Campanha Nacional de Vacinação Contra COVID-19

Marcus Vinicius Ferreira Gonçalves[1,2], Jamile Santos dos Santos[1], Caio Zava Ferreira[1], Jorge Zavaleta[1], Sérgio Manuel Serra da Cruz[1,3], Jonice Oliveira Sampaio[1]

[1]Programa de Pós - Graduação em Informática (PPGI)
Universidade Federal do Rio de Janeiro (UFRJ) - Rio de Janeiro – RJ – Brasil

[2]Escola Nacional de Saúde Pública Sergio Arouca – Fundação Oswaldo Cruz (Fiocruz) Rio de Janeiro, RJ – Brasil

[3]Programa de Pós-graduação em Humanidades Digitais
Universidade Federal Rural do Rio de Janeiro (UFRRJ) – Seropédica – RJ – Brasil

{marcus.goncalves, jamile.santos, caio.zava, jorge.zavaleta, serra}@ppgi.ufrj.br

Abstract. *The COVID-19 pandemic is a global threat. If, on the one hand, we account for many losses, on the other hand, the generation of datasets and urgent analytical demands has accelerated. Among the combat strategies, vaccination and data-centered epidemiological investigations stand out. This dataset paper presents the process of building cured and annotated datasets with provenance metadata. The main dataset is based on the registration data of the Vaccination Campaign against COVID-19 in Brazil. The dataset contains thousands of records processed up to March 2021. The data were analyzed, investigated, treated and cross-checked with other sources, in order to correct and complement them, resulting in cured datasets and aligned to the FAIR principles.*

Resumo. *A pandemia da COVID-19 é uma ameaça global. Se por um lado contabilizamos muitas mortes, por outro lado, tem-se acelerado a geração de datasets e demandas analíticas. Dentre as estratégias de combate, destacam-se a vacinação e as investigações epidemiológicas centradas em dados. Este artigo apresenta o processo de construção de datasets curados e anotados com metadados de proveniência, tendo como base os dados de registro da Campanha de Vacinação contra COVID-19 no Brasil. O dataset contém milhares de registros tratados até março de 2021. Os dados foram analisados, investigados, tratados e cruzados com outras fontes, de modo a corrigi-los e complementá-los, resultando em datasets curados e alinhados aos princípios FAIR.*

1. Introdução

A doença Corona Virus Disease-19 (COVID-19) é a maior pandemia da história recente da humanidade. Ela é causada pelo novo coronavírus, designado como Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2). Segundo as informações disponíveis no site do Ministério da Saúde[1] (MS), o COVID-19, é uma infecção respiratória aguda potencialmente grave, de distribuição global e elevada transmissibilidade. Ela é disseminada através de gotículas, aerossóis ou através de superfícies contaminadas.

[1]Sobre a doença - site do Ministério da Saúde: https://coronavirus.saude.gov.br/sobre-a-doenca

Até março de 2021, quando se iniciou este trabalho, o Painel do Coronavírus (COVID-19) da OMS[2], contabilizava 114.451.590 casos confirmados no mundo, acréscimo de 0,31% de novos casos ao dia e 2.548.082 mortes pela doença, com uma taxa de aumento de 0,31% de mortes diárias. O Brasil ocupava o segundo lugar no ranking, sendo responsável por 10.587.001 casos, equivalente a 9,25% de casos no mundo, com aumento de 0,34% novos casos por dia - taxa de crescimento superior a mundial - e segundo lugar em mortes com 255.720, equivalente a 10,04% mortes no mundo, com um aumento de 0,31% na taxa de óbitos diários.

Entre o primeiro caso de COVID-19 relatado em fevereiro de 2020 e a primeira pessoa vacinada em janeiro de 2021, o Brasil formou parcerias para pesquisa e desenvolvimento de vacinas que incluem transferência de tecnologia por intermédio da Fundação Oswaldo Cruz (FIOCRUZ) e do Instituto Butantan [Martins et al. 2021] para as vacinas Covshield da AstraZenaca-Oxford-FIOCRUZ e Coronavac da Sinovac-Butantan. Além delas, foram também certificadas para uso emergencial pela Agência Nacional de Vigilância Sanitária (Anvisa), as vacinas da Pfizer e BioNTec e da Janssen-Cilag, que já haviam sido utilizadas antes para fins experimentais no Brasil.

A Portaria nº 69 [Ministério da Saúde - Brasil 2021], de 14 de janeiro de 2021 do MS instituiu a obrigatoriedade do registro de aplicação de vacinas contra o COVID-19 nos Sistemas de Informação em Saúde (SIS), valendo para as instituições públicas ou privadas. A responsabilidade de armazenar essas informações, a necessidade de planejamento e execução de respostas ao enfrentamento da COVID-19 fica com a cargo do MS, que por questões de transparência publica-os soba a forma de dados abertos em saúde.

Segundo o Informe Técnico do Conselho Nacional de Secretarias Municipais de Saúde (CONASEMS), na Campanha Nacional de Vacinação contra a COVID-19 é observada a necessidade de acompanhar e monitorar os vacinados, a movimentação de imunobiológico para facilitar a rastreabilidade e controle dos itens distribuídos, facilitando o planejamento e o acompanhamento dos vacinados em situações de Eventos Adversos Pós Vacinação (EAPV).

Segundo o Our World in Data[3] [Mathieu et al. 2021] até 11 de março de 2021, o mundo contabilizava 206.24 milhões de pessoas vacinadas, ou seja 2.6% da população mundia, onde 75.8 milhões estavam totalmente vacinadas, equivalente a 0,97% da população e 130.43 milhões com a primeira dose, equivalente a 1,7% da população. O Brasil tinha 8.09 milhões de pessoas vacinadas, 3,8% dos brasileiros, onde 2.65 milhões estavam totalmente vacinadas, ou seja 1,3% e 5.44 milhões com pelo menos a primeira dose, equivalente a 2,6% dos brasileiros, ocupando a 4^a posição de vacinação mundial e 11^a posição entre o percentual de vacinados em relação ao tamanho da população.

Neste contexto, o objetivo deste *dataset paper* é disponibilizar para a sociedade um conjunto de dados (*datasets*) curados, anotados e enriquecidos com metadados de proveniência e aderente aos princípios FAIR sobre a Campanha Nacional de Vacinação contra a COVID-19. O *dataset* foi desenvolvido através da construção de um *pipeline* de processamento, fundamentado nas etapas do processo de Ciência de Dados. Adicionalmente, são compartilhados *datasets* complementares, não disponíveis no site do OpenDataSUS

[2]Painel do Coronavirus (COVID-19) da Organização Mundial de Saúde (OMS): https://covid19.who.int/
[3]Coronavirus (COVID-19) Vaccinations: https://ourworldindata.org/covid-vaccinations

e que são necessários para complementação e entendimento do *dataset* original. A partir desses *datasets* é possível utilizar os dados em ferramentas de análise e visualização. Isto não poderia ser feito a partir dos dados brutos sem toda análise, tratamento e enriquecimento realizados.

Este documento foi organizado da seguinte forma: a seção 2 apresenta os trabalhos relacionados, a seção 3 apresenta os métodos utilizados na obtenção do *dataset* original, os processos básicos de pré processamento e limpeza usados em Ciência de Dados, a seção 4 apresenta os *datasets* resultantes, seus dicionários de dados com informações sobre Proveniência e enquadramento aos Princípios FAIR. Por fim, a seção 5 apresenta as considerações finais.

2. Trabalhos relacionados

A literatura científica relacionada a metodologia de geração de *datasets* curados, enriquecidos com proveniência e alinhados aos princípios FAIR sobre a vacinação contra a COVID-19 é insuficiente. Até o momento não foi encontrada referência nacional de disponibilização de *datasets* deste tema. [Oliveira et al. 2021] apresenta uma análise técnica de recorte temporal, baseado nos microdados de COVID-19 referentes ao estado de Mato Grosso. Estes compõem apenas uma parte dos dados utilizados neste trabalho. Ainda neste tema, existem painéis sumarizados da OMS, Our World in Data [Mathieu et al. 2021] e também *datasets* sumarizados de alguns países, por exemplo a Índia. Possivelmente, a dificuldade de localização de trabalhos neste tema, se dá pelo grande volume dos dados e arquivos e a necessidade de alto processamento para análise.

[Clarindo et al. 2020] apresenta o QualiSUS, um *dataset* construído a partir de dados oriundos de bases públicas de saúde para apoiar pesquisadores e gestores. Outros trabalhos apresentam abordagens, análises e experiências com *datasets*. [Barbosa Pina et al. 2020] apresenta uma abordagem para coleta e análise utilizando a bilioteca Keras, centrada em dados de proveniência, usando uma aplicação real com rede neural. [Rocha et al. 2021] apresenta uma análise de *datasets* usando inteligência artificial espacial para planejamento das ações do Plano Nacional de Vacinação contra a COVID-19, utilizando seis fontes de dados públicas. Embora o tema seja a vacinação, ele não utiliza o mesmo *dataset* deste trabalho.

3. Materiais e Métodos

Segundo [Squire 2015], projetos em Ciência de Dados são caraterizados por seis etapas principais, a saber: definição do problema, coleta e armazenamento, limpeza, análise e métodos de processamento, representação e visualização, resolução do problema e comunicação dos resultados. A autora afirma que embora sejam etapas que ocorram de forma interativa, algumas etapas podem ser revisitadas novamente.

A Figura 1 ilustra conceitualmente o *pipeline* desenvolvido especificamente para este caso e está constituído por 4 etapas: aquisição de dados; análise e investigação; ajuste; limpeza, enriquecimento e disponibilização dos *datasets* curados. Estas etapas são detalhadas nos próximos subitens e na seção 4.

3.1. *Dataset* bruto

A Rede Nacional de Dados em Saúde (RNDS), criada pelo Departamento de Informática do SUS (DATASUS) é uma plataforma nacional de interoperabilidade de dados em

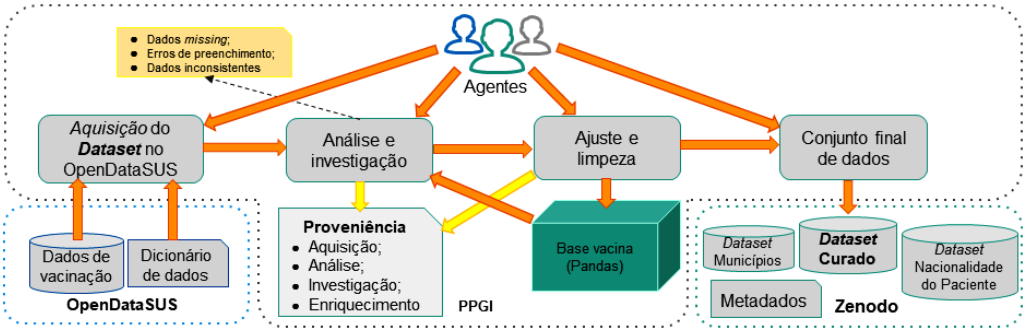

Figura 1. Pipeline do método desenvolvido.

saúde. Ela promove o compartilhamento de informações entre a Rede de Atenção à Saúde nos setores público e privado. O RNDS concentra os dados de vacinação contra o COVID-19 advindos dos diversos SIS dos municípios. O OpenDataSUS concentra os *datasets* brutos e não curados de vacinação contra a COVID-19 no site: https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao.

O *dataset* bruto mantém registros individualizados e anonimizados dos pacientes, respeitando o disposto na Lei Geral de Proteção de Dados Pessoais (LGPD) Lei nº 13.709, de 14 de agosto de 2018. Os dados estão publicados no formato de dados abertos utilizando *Comma Separeted Values* (CSV) ou *Application Programming Interface* (API). A obtenção desses dados pode ser feita via site, selecionando o documento e clicando no botão de download ou via API do *Comprehensive Knowledge Archive Network* (CKAN).

3.2. Aquisição dos dados brutos

A amostra coletada em março de 2021 em formato CSV. Continha 7.908.949 registros de vacinação (doses: única, primeira ou segunda) até 11 de março de 2021, composto por 33 campos e tamanho total de 4.0 GB. A primeira versão do *dataset* bruto foi criada em janeiro de 2021 e é atualizada diariamente. Segundo informações do OpenDataSUS, a série temporal iniciou-se partir de 18 de janeiro de 2021, porém encontraram-se registros anteriores, correspondentes aos estudos preliminares à campanha. O *dataset* possui granularidade geográfica em escala municipal, faz referência à Tabela de Códigos de Municípios (IBGE) e ao Cadastro Nacional de Estabelecimentos de Saúde. No entanto, identificamos outras referências que não estão citadas explicitamente e serão discutidas mais à frente.

3.3. Análises de dados e investigações no *Dataset* bruto

De posse do *dataset* bruto e do dicionário de dados, utilizamos a linguagem Python v.3.8.8 e suas bibliotecas, com destaque para o Pandas v.1.2.4 para a construção do *pipeline* de análises e investigações de dados, na seguinte ordem:

1. Comparar os campos e seus tipos de dados com o dicionário disponibilizado e seus conteúdos;
2. Verificar a distribuição de cada campo de acordo com o conteúdo para entender os possíveis valores e quantidades de ocorrências;
3. Identificar os dados que diferenciam drasticamente dos demais ou fogem da normalidade (*outliers*);
4. Identificar as relações entre os campos ou com dados externos;

5. Identificar as inconsistências e erros existentes.

A análise preliminar do *dataset* permitiu inferir as seguintes investigações: verificar a idade do paciente na data de vacinação através de sua data de nascimento, relacionar o código da raça e cor com sua descrição e relacionar nome do município, unidade federativa e código do país do endereço com o código do IBGE do município do endereço.

Outras investigações permitiram avaliar a consistência dos dados no *dataset*, por exemplo, se a idade do paciente informada no dia da vacinação corresponde com a idade calculada através da data de nascimento; se o paciente possui identificador único; se um paciente foi vacinado mais de uma vez com a mesma dose da vacina ou se foi vacinado mais de duas vezes com registros diferentes; se o paciente foi vacinado durante o segundo semestre de 2020 até a data de aquisição dos dados; se o código do IBGE do município corresponde ao nome do município e com o código do país na base do IBGE e se o código, nome e fabricante da vacina correspondem com os dados do MS.

3.4. Problemas detectados no *dataset* bruto

Não raro, existem falhas nos dados brutos abertos disponibilizados pelo governo, o que de fato encontramos neste *dataset*. O problema inicial foi de ordem técnica, pois processar um grande volume de dados, necessitou recursos computacionais robustos para realizar as análises. Já em relação ao *dataset*, foi verificado que existiam problemas de dados nulos e faltantes (*missing data*), registros incorretos e com inconsistências na maioria dos campos. Também foi identificada a ausência de campos, algumas descrições e itens de categorias no dicionário de dados, além da falta de padrões de metadados e relacionamentos com outros *datasets*. Estes problemas justificaram a revisão e elaboração de um novo dicionário que é apresentado na seção 4.1.

Os dados nulos e faltantes foram identificados através das ferramentas de análise estatística e representaram 2,02% dos registros. Os registros incorretos e inconsistências representaram 19,18% e foram encontrados através da exploração manual da base com ferramentas de sumarização, análise da distribuição dos dados, consultas pontuais, contagens e agrupamentos. Para estas tarefas foram utilizadas as funções da biblioteca Pandas.

Durante a exploração buscou-se compreender os padrões dos dados, além de detectar os erros, foram constatadas ausência de padronização entre os campos e de informações de relacionamentos com dados externos. Outros problemas foram identificados, como espaçamentos extras nas extremidades dos valores dos campos de tipo alfanumérico, erros de digitação, duplicação indevida do registro de vacinação e valores inválidos em alguns campos, como por exemplo: nomes e e-mails no campo de lote da vacina e CEP que não existem na base dos Correios.

Um problema intrínseco à natureza dos dados utilizados no SIS, trata de um erro sobre os preenchimentos dos campos de código e nome do endereço do país do paciente (paciente_endereco_copais e paciente_endereco_nmpais), pois eram preenchidos com o código e nome da nacionalidade do paciente, segundo a tabela de dados do SIS-BR.

Adicionalmente, foram encontrados pacientes com idades abaixo de 18 anos (2897 pacientes) e acima de 115 anos (4340 pacientes); pacientes sem identificadores únicos; pacientes que tomaram mais de uma vez a mesma dose ou acima de três doses e pacientes que foram vacinados fora do intervalo possível do período da campanha ou em datas futuras. Essas constatações configuram erros e inconsistências graves nos registros. Também

foram detectadas divergências entre o código do IBGE, nome do município e código e nome do país. Esses erros correspondem a 0.55%. Os dados relativos ao CEP do paciente e ao lote da vacina não foram ratificados, devido ao grande volume de dados faltantes, errados, inconsistentes e sem uma tabela de associação para validação destes campos.

3.5. Limpeza e ajustes dos dados

A etapa de limpeza e ajustes de dados também foi processada em lotes. Foi necessária a utilização do *dataset* adicional do IBGE contendo o código do município, nome, e unidade federativa, disponível no Sistema IBGE de Recuperação Automática (SIDRA)[4]. Também foi utilizado um dataset de nacionalidade dos pacientes dos SIS.

A primeira tarefa foi a padronização dos campos segundo seu tipo, seguido da conversão de tipos numéricos e temporais e remoção de espaçamentos. Os erros digitação mais simples foram corrigidos e a duplicação de registros foi avaliada caso a caso, com remoção para os casos de registros semelhantes ou correção quando possível, mantendo registros que configuraram mais de duas doses com informações diferentes.

A seguir, foi utilizado o registro de valores fixos sem correspondência com a realidade (*highvalue*), para funcionarem como marcadores de substituição de valores nulos, não identificados ou incorretos e uma padronização nos campos possíveis, conforme a composição dos dados. Em casos mais específicos, foram adicionados rótulos para representar dados faltantes, de forma a constarem na base para futuras análises. Todos os marcadores e rótulos estão descritos no novo dicionário de dados.

Um exemplo que merece destaque, foi a identificação de registros que continham dados conflitantes nos campos do paciente: nome do município e do país do endereço do paciente. O problema foi detectado quando o município preenchido era brasileiro, mas o país era outro que não o Brasil. Esses registros foram considerados e, consequentemente, tratados como erros de preenchimento no formulário, visto que a frequência de ocorrências era muito baixa em relação aos demais registros do *dataset*. Por isso, optou-se por substituir o nome do país, pelos nomes que faziam sentido, considerando a informação contida no campo do município.

Adicionalmente, foi verificado que os campos: código do CNES, código do grupo de atendimento do paciente, descrição da dose, código da vacina e nome da vacina não possuíam dados nulos, logo foram mais facilmente tratados e, alguns serviram de suporte para a reconstrução de dados faltantes através de comparação.

4. *Datasets* curados e enriquecidos com metadados de proveniência

Nesta seção, apresenta-se o produto resultante do método adotado. Ele é constituído por um *dataset* principal, por um dicionário de dados - construído após o enriquecimento - e mais dois *datasets* complementares. Também explicitamos a proveniência retrospectiva dos dados e uma discussão sobre o processo de enquadramento dos princípios FAIR. Os *datasets* e seus arquivos complementares estão disponíveis em um repositório no endereço: ***https://doi.org/10.5281/zenodo.5193920***[5].

[4]Sistema IBGE de Recuperação Automática (SIDRA): https://sidra.ibge.gov.br/home/pmc/brasil

[5]Dataset: `https://doi.org/10.5281/zenodo.5193920`

4.1. Dicionário de dados

O novo dicionário de dados está apresentado na Tabela 1. Foi reconstruído em função das necessidades que emergiram ao longo da aplicação do método. Durante as atividades recorremos ao dicionário de dados original para entendermos o conteúdo dos campos, porém por vezes não existia a informação, ou não estava completa, ou não condizia com os dados encontrados. Os valores nulos representaram um problema nesta pesquisa, suas análises não apareceriam nas representações e não indicariam resultados confiáveis, isso levou a incluí-los e categorizá-los no dicionário de dados, criando novos rótulos e categorias para valores sem informação nos campos.

Dicionário de dados da Base de Vacinação[6]				
Sobre	**Nome dos campos**	**Descrição**	**Tipo**	**Categoria**
Registro	document_id	Identificador único anonimizado do registro de vacinação	padronizado	
Paciente	paciente_id	Identificador único anonimizado do paciente	padronizado	
	paciente_idade	Idade do paciente na data de vacinação	número inteiro	
	paciente_dataNascimento	data de nascimento do paciente	data e hora	
	paciente_enumSexoBiologico	sexo biológico do paciente	categoria	'F': 'FEMININO'; 'M': 'MASCULINO', 'I': 'NAO INFORMADO'
	paciente_racaCor_codigo	código de raça e cor do paciente	número inteiro	1: 'BRANCA'; 2: 'PRETA'; 3: 'PARDA'; 4: 'AMARELA'; 5: 'INDIGENA'; 99: 'SEM INFORMACAO'
	paciente_racaCor_valor	Raça do Vacinado (Branca, Preta, Parda, Amarela, Indígena e Sem informação)	categoria	1: 'BRANCA'; 2: 'PRETA'; 3: 'PARDA'; 4: 'AMARELA'; 5: 'INDIGENA'; 99: 'SEM INFORMACAO'
	paciente_endereco_coIbgeMunicipio	código do IBGE do município do endereço do paciente	número inteiro	referencia municipio.csv e não informado: 999999
	paciente_endereco_coPais	código do país da nacionalidade do endereço do paciente	número inteiro	referencia nacionalidadepaciente.csv e não informado: 0
	paciente_endereco_nmMunicipio	nome do município do endereço do paciente	alfanumérico	referencia municipio.csv e não informado 'SEM INFORMACAO'
	paciente_endereco_nmPais	nome do país da nacionalidade do endereço do paciente	alfanumérico	referencia nacionalidadepaciente.csv e não informado: 'SEM INFORMACAO'
	paciente_endereco_uf	unidade federativa do endereço do paciente	categoria	referencia municipio.csv e não informado: 'XX'
	paciente_endereco_cep	código do grupo de atendimento do paciente	número inteiro	não informado: 0
	paciente_nacionalidade_enumNacionalidade	lista enumerada da nacionalidade do Paciente	categoria	'B': 'BRASILEIRO'; 'E': 'ESTRANGEIRO', 'I': 'SEM INFORMAÇÃO'
Estabelecimento	estabelecimento_valor	código do CNES	número inteiro	referencia a tabela do CNES
	estabelecimento_razaoSocial	razão social do estabelecimento	alfanumérico	
	estalecimento_noFantasia	nome fantasia do estabelecimento	alfanumérico	
	estabelecimento_municipio_codigo	código do IBGE do município do estabelecimento	número inteiro	referencia municipio.csv e não informado: 999999
	estabelecimento_municipio_nome	nome do município do estabelecimento	alfanumérico	referencia municipio.csv e não informado 'SEM INFORMACAO'
	estabelecimento_uf	unidade federativa do estabelecimento	categoria	referencia municipio.csv e não informado: 'XX'
Vacina	vacina_grupoAtendimento_codigo	código do grupo de atendimento do paciente	número inteiro	
	vacina_grupoAtendimento_nome	nome do grupo de atendimento do paciente	alfanumérico	
	vacina_categoria_codigo	código da categoria do paciente	número inteiro	
	vacina_categoria_nome	nome da categoria do paciente	alfanumérico	
	vacina_lote	lote da vacina	alfanumérico	
	vacina_fabricante_nome	nome do fabricante da vacina	alfanumérico	
	vacina_fabricante_referencia	referência do fabricante da vacina alfanumérico	alfanumérico	
	vacina_dataAplicacao	data de aplicação da vacina	data e hora	
	vacina_descricao_dose	descrição da dose aplicada	categoria	PRIMEIRA DOSE'; 'SEGUNDA DOSE'; 'DOSE ÚNICA'
	vacina_codigo	código da vacina	número inteiro	81: 'Inválido'; 85: 'Vacina Covid-19-Covishield'; 86: 'Covid-19-Coronovac-Sinovac/Butantan';
	vacina_nome	descrição da dose aplicada	categoria	87: 'Vacina Covid-19-BNT162b2-BioNTech/Fosun Pharma/Pfizer'; 88: 'Vacina Covid-19-Ad26.COV2.S-Janssen-Cilag'
Registro	sistema_origem	sistema de saúde de origem do registro de vacinação	alfanumérico	
	data_importacao_rnds	data de importação pela RNDS	data e hora	

Tabela 1. Representação do novo dicionário de dados da base de Vacinação.

[6]Melhor visualizado em: *https://doi.org/10.5281/zenodo.5193920*.

4.2. *Datasets* complementares

Durante a elaboração do novo dicionário de dados e do encaminhamento da solução dos problemas encontrados na aplicação do método, foi identificada a necessidade de agregação de *datasets* adicionais sobre os municípios brasileiros e outro sobre as nacionalidades dos pacientes. Estes *datasets* auxiliaram as análises e investigações sobre o *dataset* principal e foram criados dicionários de dados para ambos.

O *dataset* dos municípios foi criado a partir do SIDRA/IBGE, contendo o código do IBGE do município, o nome e a unidade federativa correspondente. Foram utilizados para referenciar o código do IBGE do município do endereço do paciente e do estabelecimento de saúde, o nome do município do endereço do paciente e do estabelecimento de saúde e das respectivas unidades federativas, conforme apresentado na Tabela 2.

Conteúdo de Dados do Município a partir do SIDRA/IBGE			
Nome do Campo	**Descrição**	**Tipo**	**Categoria**
coIbgeMunicipio	código do IBGE do município do endereço do paciente	número inteiro	
nmMunicipio	nome do município do endereço do paciente	alfanumérico	
uf	sigla da unidade federativa	categoria	sigla da UF

Tabela 2. Dicionário de dados do município.

O *dataset* de nacionalidades de pacientes foi construído a partir de informações em sistemas hospitalares. Estes dados foram utilizados para referenciar o código e nome do país do endereço do paciente, conforme apresentado na Tabela 3.

Conteúdo de Dados em relação a nacionalidade do paciente		
Nome do Campo	**Descrição**	**Tipo**
Codigo	código do país	número inteiro
Nacionalidade	nome do país	alfanumérico

Tabela 3. Dicionário da nacionalidade do paciente.

4.3. Proveniência do *dataset*

O conceito de proveniência inicialmente se referia à explicitação das origens ou procedência de objetos de arte e recentemente foi incorporado pela e-Ciência, sendo paulatinamente disseminado e utilizado nas áreas de Ciência de Dados e Aprendizados de Máquina. Conforme a definição clássica de [Buneman et al. 2001], proveniência são metadados complementares de um dado ou processo, contendo informações de como, quando, onde e por que os dados foram obtidos e quem os produziu.

Segundo o W3C Provenance Working Group [Missier et al. 2013], a proveniência é representada como um grafo cujos registros, explicitam como agentes, processos, entidades e atividades estão relacionadas. Atualmente, os metadados de proveniência são um dos indicativos de qualidade dos dados, porém ainda estão ausentes em muitos trabalhos de Ciência de dados [Sikos and Philp 2020].

Neste trabalho foi adotado o modelo PROV [Missier et al. 2013]. O modelo é capaz de expressar a proveniência de dados através de entidades, atividades e agentes, envolvidos em produção, entrega ou enunciado de um recurso de dado na Web. Durante a produção dos *datasets* curados foram envolvidos três agentes distintos, que executaram

sete tarefas de processamento de dados brutos, que foram representadas em seis entidades. Foi realizado o histórico de registro de execução das rotinas sob a forma de grafo de proveniência, usando o modelo supracitado. Os registros de proveniência aumentaram a qualidade do trabalho, ajudando a descrever os processos e insumos utilizados durante o método. O grafo de proveniência foi disponibilizado no repositório junto com os *datasets*.

4.4. Enquadramento dos princípios FAIR

Os princípios FAIR visam tornar os *datasets* mais localizáveis (*Findable*), acessíveis (*Accessible*), interoperáveis (*Interoperable*) e reutilizáveis (*Reusable*) [Wilkinson et al. 2016]. Os princípios indicam características aplicáveis tanto aos recursos humanos e tecnológicos, quanto a ferramentas de software, vocabulários e infraestrutura de dados de forma a contribuir com a facilidade da descoberta e reutilização de dados e metadados. Tais princípios são compostos por subprincípios e estão voltados principalmente para tornar os dados de pesquisa legíveis por máquina e humanos.

Os princípios FAIR vem alcançando crescente aceitação na área de Ciência de Dados. Buscando conformidade com tais princípios, neste trabalho foram observados os seguintes pontos: o *dataset* curado, os *datasets* complementares e seus arquivos descritos por metadados suficientemente ricos, registrados e indexados em um recurso pesquisável e acessível aos usuários (humanos ou máquinas) em potencial. Foi utilizado o Zenodo[7], cujos metadados estão em conformidade com o DataCite's Metadata Schema[8]. Nele foi atribuído um *Digital Object Identifier* (DOI) para o conjunto de dados. O identificador (DOI) permite que ligações persistentes sejam estabelecidas entre os dados, metadados e outros materiais relacionados, de modo a auxiliar no registro e indexação em ferramentas de pesquisa, provendo conformidade com o princípio *Findable*.

Desta forma, humanos e máquinas podem acessar mais facilmente os *datasets* e arquivos complementares através do Zenodo, mediante autorização apropriada e por meio do protocolo REST API. Inclusive mantendo os metadados mesmo quando os dados não estiverem mais disponíveis, buscando conformidade com o princípio *Accessible*. Os *datasets* e seus metadados foram descritos utilizando *JSON Schema* para a representação do conhecimento em um vocabulário padrão conhecido. Os arquivos estão em um formato aberto CSV e TXT, facilitando o reuso, buscando conformidade com o princípio *Interoperable*.

A reutilização implica que os dados sejam liberados com uma licença de uso clara e acessível, neste caso utilizou-se a *Creative Commons CC-BY*, que permite distribuir, alterar ou reinventar o *dataset*, desde que se cite a fonte inicial. Esta licença foi utilizada para respeitar a licença original do *dataset* bruto. Para possibilitar reutilização foram disponibilizados metadados ricos e documentação complementar, permitindo versionamento. Estes itens foram providos no Zenodo, que atende aos padrões relevantes da comunidade de interesse, além de fornecer informações sobre a proveniência, buscando conformidade com o princípio *Reusable*.

[7]Zenodo: https://zenodo.org/

[8]DataCite's Metadata Schema: https://schema.datacite.org/

5. Considerações Finais

O Brasil é um país que abriga grandes desigualdades, em especial nas áreas da saúde, educacional e tecnológica. Isso se refletiu nos problemas encontrados nas entradas de dados nos SIS utilizados pela campanha de vacinação. A elevada imprecisão do *dataset* bruto serviu de motivação inicial para desenvolvimento deste trabalho que envolveu análise, investigação, limpeza e correção de dados, resultando em *datasets* curados e abertos. Os dados resultantes representam uma oportunidade para que a sociedade, principalmente pesquisadores, compreendam a dinâmica e os problemas da campanha de vacinação e do enfrentamento e combate à COVID-19 no Brasil, em termos gerenciais, de aplicação das vacinas, temporais e geográficos.

Nossas contribuições estão no desenvolvimento do *pipeline* de processamento e principalmente na disponibilização do *dataset* com os dados curados, enriquecidos com proveniência e alinhados aos princípios FAIR sobre a vacinação contra a COVID-19 no Brasil. Inclui também os *datasets* complementares, importantes para a compreensão e estudos futuros. Os novos dicionários de dados são outros itens imprescindíveis neste trabalho. A disponibilização destes *datasets* a longo prazo, proporcionam a reusabilidade e o uso constante para outros fins, com destaque para a utilização direta destes dados em ferramentas de visualização e análise.

Destacamos que a proveniência de dados e os princípios FAIR desempenharam um papel importante neste trabalho, tornando-se mais relevante se for usado por grupos colaborativos, permitindo criar uma linha de base para futuras pesquisas e gerar análises, visualizações, contribuições para investigações e buscas de melhorias contínuas sobre os dados, que estão sob a licença *Creative Commons CC-BY*. As atividades realizadas foram desenvolvidas de forma iterativa envolvendo várias execuções sob a fonte de dados em uma infraestrutura computacional de servidores, permitindo a verificabilidade das atividades, expondo a capacidade de reproduzir os resultados de maneira consistente, autêntica e transparente. De forma indireta, é possível potencializar discussões sobre as ações de enfrentamento da doença e seus impactos, fiscalização do processo de vacinação e a elaboração de evidências e *insights*, auxiliando principalmente nas tomadas de decisão.

Compreender esse cenário em rápido movimento é particularmente desafiador e relevante, visto que grande parte desses *datasets* ainda não foram investigados através de um processo de revisão por pares ou mesmo adequadamente tratados e processados por computador. Espera-se que estes *datasets* auxiliem nas tarefas de análise, pesquisa e acompanhamento do avanço da vacinação contra COVID-19 no Brasil, ajudando a compreender o processo e a apoiar os formuladores de políticas de saúde. Caso sejam encontradas imprecisões significativas, contamos com a colaboração para melhorarmos o método e atualizamos continuamente os *datasets*.

Como trabalhos futuros, planejamos incluir novos *datasets* com registros de vacinação desde o último registrado neste trabalho até os mais recentes, aumentando o espaço temporal, permitindo abranger toda a campanha de vacinação, o que necessita de um trabalho contínuo e periódico. Pretendemos também fazer recortes pontuais por regiões ou estados, incluindo dados geográficos para expandir as possibilidades de análises e visualizações. A indicação é que este trabalho torne-se uma ferramenta, que poderá ser ampliada à medida que forem produzidos novos dados sobre a campanha.

Agradecimentos

Este estudo foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-Tecnodigital) - Código Financeiro: 88887.514128/2020-0 e parcialmente patrocinado pelo Fundo Nacional de Desenvolvimento da Educação (FNDE), Programa de Tutoria Educacional (PET-SI / UFRRJ), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - Bolsa DT -II (315399 / 2018-0), e Fundação de Pesquisa Carlos Chagas Filho (FAPERJ) – Código E-26/210.192/2020.

O ambiente computacional utilizado foi cedido pela Escola Nacional de Saúde Pública Sergio Arouca (ENSP) da Fundação Oswaldo Cruz (Fiocruz) e o suporte de rede lógica provido pela equipe de rede lógica do Serviço de Gestão de TI (SGTI/ENSP/Fiocruz).

Referências

Barbosa Pina, D., Kunstmann, L., de Oliveira, D., Valduriez, P., and Mattoso, M. (2020). Uma abordagem para coleta e análise de dados de configurações em redes neurais profundas. In Proceedings of 2nd SBBD DSW, pages 187–192.

Buneman, P., Khanna, S., and Wang-Chiew, T. (2001). Why and where: A characterization of data provenance. In International conference on database theory, pages 316–330. Springer.

Clarindo, J. P., Fontes, W., and Coutinho, F. (2020). Qualisus: um dataset sobre dados da saúde pública no brasil. In Proceedings of 2nd SBBD DSW, pages 418–428.

Martins, W. A., de Oliveira, G. M. M., Brandão, A. A., Mourilhe-Rocha, R., Mesquita, E. T., Saraiva, J. F. K., Bacal, F., and Lopes, M. A. C. Q. (2021). Vacinação do Cardiopata contra COVID-19: As Razões da Prioridade. Arquivos Brasileiros de Cardiologia, 116:213 – 218.

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., and Rodés-Guirao, L. (2021). A global database of covid-19 vaccinations. Nature human behaviour, pages 1–7.

Ministério da Saúde - Brasil (2021). Portaria nº 69, de 14 de janeiro de 2021. institui a obrigatoriedade de registro de aplicação de vacinas contra a covid-19 nos sistemas de informação do ministério da saúde. [Acessado em 13 abr. 2021].

Missier, P., Belhajjame, K., and Cheney, J. (2013). The w3c prov family of specifications for modelling provenance metadata. In Proceedings of the 16th International Conference on Extending Database Technology, pages 773–776.

Oliveira, L. A., Muraro, R., Cristina, A. P., Andrade, A., Cecconello, S., and Lalucci, M. M. (2021). Vacinação contra a covid-19 em mato grosso: primeiros resultados. Nota Técnica - Universidade Federal de Mato Grosso.

Rocha, T. A. H., Boitrago, G. M., Mônica, R. B., Almeida, D. G. d., Silva, N. C. d., Silva, D. M., Terabe, S. H., Staton, C., Facchini, L. A., and Vissoci, J. R. N. (2021). Plano nacional de vacinação contra a covid-19: uso de inteligência artificial espacial para superação de desafios. Ciência & Saúde Coletiva, 26:1885–1898.

Sikos, L. F. and Philp, D. (2020). Provenance-aware knowledge representation: A survey of data models and contextualized knowledge graphs. Data Science and Engineering, 5:293–316.

Squire, M. (2015). Clean Data: Save time by discovering effortless strategies for cleaning, organizing, and manipulating your data. Birmingham, Packt Publishing Ltd.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. Scientific data, 3(1):1–9.