Curating, Enriching and FAIRifying Datasets from the Brazilian COVID-19 Vaccination

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,
jonice@dcc.ufrj.br

Abstract. As the world struggles to face the challenges of vaccination against COVID-19, more attention needs to be paid to the issues related to the lack of transparency and accessibility of curated vaccination datasets. Among the strategies to combat COVID-19, vaccination and data-centered epidemiological investigations are the best ones. This paper presents the process of building cured and annotated datasets with provenance metadata. The primary 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, treated, cross-checked, and linked with other sources to correct and complement them, resulting in cured datasets and aligned to the FAIR Data principles.

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: Data Science, COVID-19, Data Provenance, FAIR Pipelines, Data paper

1. INTRODUCTION

Corona Virus Disease-19 (COVID-19) is one of the most significant pandemics in recent human history. It brought attention to the challenges in securing access to vaccines on a global scale. The disease is caused by a new coronavirus designated as Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2). According to information available on the website of the Brazilian Ministry of Health (BMH)[1], COVID-19 is a potentially severe acute respiratory infection with global distribution, high transmissibility. Several variants of the SARS-CoV-2 virus cause it. BMH also states that the disease spreads through aerosols, droplets, or across contaminated surfaces.

Up to October 2021, when this version of the work was written, the WHO Coronavirus Panel (COVID-19)[2], reported 246.297,757 million confirmed cases worldwide. A slight upward trend (3% increase) in new cases was observed weekly, with just over 3 million new cases reported. The number of new deaths also increased slightly (5%), with over 49,000 new deaths reported. Brazil has been one of the hotspots for COVID-19 in 2021. The country occupied seventh place in the ranking, being

[1]About the disease - Ministry of Health website: https://coronavirus.saude.gov.br/sobre-a-doenca
[2]World Health Organization (WHO) Coronavirus Panel (COVID-19): https://covid19.who.int

responsible for 101.841.44 of the cases in the world and with an increase of 2.03% new cases per day. This growth rate was higher than the world's average rate. Brazil also had second place in deaths with 611.48 of the world's deaths and a 2.8% increase in the daily death rate.

The Brazilian vaccination campaign against COVID-19 started in mid-January 2021, with only two imported vaccines: Coronavac (Sinovac, from China) and AZD1222 (Oxford-AstraZeneca, from the UK). The national vaccination campaign has been initially targeted at four priority groups: the health workers, the elderly (starting with those aged 85 years or more, and gradually vaccinating younger citizens), the institutionalized individuals, and the indigenous populations [Victora et al. 2021].

Between the first case of COVID-19 reported in late February 2020 and the first person being vaccinated, Brazil has formed partnerships for vaccine research and development that include technology transfer. These partnerships were formed through the Oswaldo Cruz Foundation (FIOCRUZ) for Covshield vaccines AstraZeneca-Oxford-FIOCRUZ, and the Butantan Institute [Martins et al. 2021] for Coronavac vaccines Sinovac-Butantan. A few months later, the Pfizer and BioNTech, and Janssen-Cilag vaccines, which had already been used for experimental purposes in Brazil in 2020, were also certified for official use by the National Health Surveillance Agency (ANVISA).

The vaccination campaigns around the world have been associated with reductions in hospital admissions, and mortality among targeted population groups [Vasileiou et al. 2021], [Bernal et al. 2021]. Thus, to mitigate COVID-19's spread and maintain the Brazilian vaccination records, the ordinance No. 69 [Ministério da Saúde – Brasil 2021], of January 14, 2021, of the BMH instituted the mandatory registration of the application of vaccines against COVID-19 in the Brazilian Health Information Systems (HIS). Both public and private institutions should register these records. The BMH is responsible for storing this data and for planning and executing activities to confront COVID-20. For reasons of transparency, the BMH publishes the data in disaggregated, user-friendly, and open-source formats.

The National Council of Municipal Health Secretariats (CONASEMS)[3] issued a Technical Report which affirms that the Brazilian COVID-19 Vaccination Campaign needs to monitor not only who was vaccinated but also the distribution of immunobiological supplies. Such recommendation facilitates the traceability and control of distributed items, facilitating the monitoring of those vaccinated in case of Post Vaccination Adverse Events (AEPV).

Historically, populations in the Global South have been more likely to suffer from inadequate or delayed supplies of vaccines and other essential medical products. According to Our World in Data[4] [Mathieu et al. 2021] as of October 31, 2021, about 3.91 billion people were vaccinated. Of those vaccinated, 3.05 billion were fully vaccinated (38.78% of the population), and 852.89 million have been given the first dose, equivalent to 10.83% of the population. Brazil had 159.20 million people vaccinated, 74.39% of Brazilians, where 118.28 million were fully vaccinated, i.e., 55.27% and 40.92 million had at least the first dose, equivalent to 19.12% of Brazilians. Brazil occupies the 4^{th} position of worldwide vaccination and 14^{th} position between the percentage of vaccinated about the population size.

The objective of this dataset paper is to offer to researchers and society the curated datasets that were cured, annotated, and enriched with metadata provenance and adhering to the FAIR principles on the Brazilian Vaccination Campaign against COVID-19. The primary dataset was developed by constructing a reproducible pipeline that follows the steps of the Data Science cycle. Additionally, complementary datasets are shared, which are not previously available on the OpenDataSUS website and are necessary for complementing and understanding the original dataset. With these datasets, it is possible to use analysis and visualization tools, and this could not be done from the raw data without analysis, treatments, and enrichment.

[3]About the Council website: https://www.conasems.org.br/
[4]Coronavirus (COVID-19) Vaccinations: https://ourworldindata.org/covid-vaccinations

This paper is an extended version of a work presented at Dataset Showcase Workshop at the Brazilian Symposium of Databases [Gonçalves et al. 2021]. We highlight that this paper was written before COVID-19 vaccine boosters or additional vaccine doses were authorized by the Brazilian Ministry of Health. The paper is organized as follows: section 2 presents the revised related works, section 3 presents the materials and methods used to produce the curated datasets and also presents the rationale of the design of reproducible pipelines used in Data Science analytical cycles, section 4 presents the structure of curated datasets, their data dictionaries with information on retrospective provenance and compliance with the FAIR Principles. Section 5 presents results and discussion. Finally, section 6 presents the final remarks and future works.

2. RELATED WORKS

Vaccination across the world has progressed slowly due to several reasons, from vaccine hesitancy and distrust in scientific expertise to unfair distribution of supplies [Fridman et al. 2021]. While high-income countries learn how to vaccinate their entire populations amidst the COVID-19 pandemic, the low and middle-income countries have been relying on the COVID-19 Vaccines Global Access (COVAX) Facility to obtain an insufficient amount of supplies [Tagoe et al. 2021].

Despite the dissimilarities, these countries have a standard feature. The majority do not offer transparent, findable, or interchangeable methods to aid researchers in analyzing or reusing vaccinations datasets. Besides, few countries proactively release timely and curated data regarding vaccination strategies and accomplishments in disaggregated, well-annotated, and open-source formats. To circumvent such issues, some researchers are proposing an integrated, centralized, and a global collection of COVID-19 patient and citizen big data [Alimadadi et al. 2020], [Depoux et al. 2020].

We do not share this opinion about a global centralized repository. Like, [Doyle and Conboy 2020], and [Ienca and Vayena 2020], we believe that i) the massive, centralized, big data patterns are inevitably undermined by the continuous, mutable flow of the COVID-19 data itself; ii) the input errors can result in magnified research inaccuracies and implementations; iii) the lack of precise semantics and provenance annotation would cause problems; iv) the national jurisdictional differences must be considered to avoid the creation of anemic and untrustworthy data scenarios. Furthermore, we foresee that national distributed data networks focused on making the SARS CoV-2 virus data FAIR are more feasible than a centralized approach. It means that wherever the origins of the data, the datasets must follow specific patterns to be Findable, Accessible, Interoperable, and Reusable by both humans and machines.

To track COVID-19 vaccine distribution, perform administration activities, collect the data, and maintain open FAIRified and curated datasets. Collaboration between policymakers, national health organizations, public information systems, and well-trained data scientists is necessary.

The scientific literature related to the methodology for generating cured datasets, enriched with data provenance [Cruz et al. 2009] and aligned with the FAIR principles on vaccination against COVID-19, is insufficient or still immature. So far, no Brazilian reference has been found for the availability of datasets on this topic. [Oliveira et al. 2021] presents a technical analysis of time frame, based on COVID-19 microdata of Mato Grosso state. On the other hand, there are summarized WHO panels and *Our World in Data* [Mathieu et al. 2021] and, also, reduced datasets from some countries, for example, Australia and India. The difficulty of locating papers on this topic is possibly due to the large volume of papers disconnected from datasets and files or the need for high processing for quantitative analysis.

[Clarindo et al. 2020] presents QualiSUS, a dataset built from data from public health databases to support researchers and managers. Other works present approaches, analyzes, and experiments with datasets. [Barbosa Pina et al. 2020] presents an approach for the collection and analysis using the Keras library. The paper is centered on provenance metadata and uses an actual application with a

neural network. [Rocha et al. 2021] presents an analysis of datasets using artificial intelligence to plan the actions of the National Vaccination Plan against COVID-19, using six public data sources. We stress that their work does not offer the raw datasets supporting their results.

3. MATERIALS AND METHODS

The COVID-19 pandemic highlighted how poor research practices, the lack of data provenance, untrustful and inaccessible repositories, and timely data could cause uncertainty in decision-making and foster mistrust in the population or reduce the reproducibility of the experiments of health researchers. Thus, ensuring the availability of timely and curated open data on critical issues, such as the number of people vaccinated, the number and types of doses administered, geographical coverage, and the number of people experiencing adverse reactions, will facilitate scientific data analysis and increase the effectiveness of government vaccination strategies.

According to [Squire 2015], The data science projects are characterized by six main stages, which start with the problem definition, data collection and storage, data cleaning, data analysis and processing, representation, and data visualization, problem-solving, and communication of results. Although these steps occur interactively, the author states that some steps can be revisited and refined.

Figure 1 conceptually illustrates the reproducible pipeline tasks developed for this work, which consists of four steps: data acquisition tasks; analysis and investigation tasks; linking, wrangling, cleaning, enrichment tasks and finally making available the cured datasets. The orange arrows represent the flow of the developed process. To represent the flow of provenance, we use the green arrows. These steps are detailed in the sub-items of the following section.

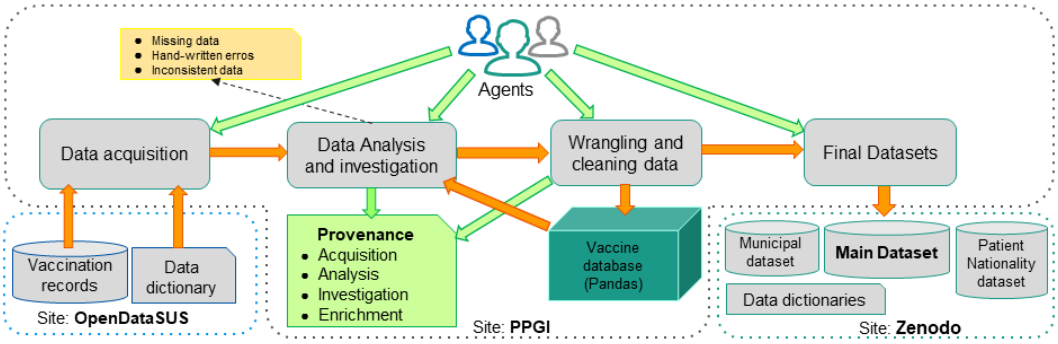

Fig. 1. Conceptual overview of the Data Science pipeline.

3.1 Raw Data

The National Health Data Network (NHDN), created by the SUS IT Department (DATASUS), is a national platform for health data interoperability. It promotes information sharing between the Health Care Network in the public and private sectors. The NHDN concentrates all the data on vaccination against COVID-19 produced by the different HIS and municipalities. OpenDataSUS provides information used to support analysis of public health, evidence-based decision-making, and the development of novel health programs. It concentrates the raw and uncurated datasets of vaccination against COVID-19 at the site: `https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao`.

The raw dataset maintains individualized and anonymized records of patients, in compliance with the provisions of the General Law for the Protection of Personal Data (LGPD), Brazilian Law No. 13.709, of August 14, 2018. The data are published in open data format using Comma Separated

Values (CSV) or Application Programming Interface (API). These data can be obtained via the website or the Comprehensive Knowledge Archive Network (CKAN) API[5].

3.2 Data Acquisition

The sample used in this work was collected in March 2021 in CSV format. It contains 7,908,949 vaccination records (single, first, or second vaccines shots) as of March 11, 2021, consisting of 33 fields and a total size of 4.0 GB. The first version of the raw dataset was created in January 2021 and is updated daily by the OpenDataSUS. According to OpenDataSUS, the time series began on January 18, 2021, but previous records were found. These records correspond to preliminary studies for the campaign. The dataset has granularity at a municipal scale and references the Table of Municipal Codes of the Geographic and Statistics Brazilian Institute (IBGE) and the National Register of Health Establishments (RNDS).

3.3 Data Pipeline

Once we identified the raw datasets and the data dictionary, we designed and built the reproducible pipeline considering the FAIR data principles and data provenance requirements. The pipeline uses Python v.3.8.8 language and libraries, emphasizing Pandas v.1.2.4 and MatPlotLib in the Jupyter Notebook environment. We considered the following steps:

(1) Compare fields and their data types with the available dictionary and its contents;
(2) Check the distribution of each field according to the content to understand the possible values and quantities of occurrences;
(3) Identify and treat data outliers;
(4) Identify the relationships between fields or with external data;
(5) Identify existing data inconsistencies, nulls, and errors.

The initial analysis of the dataset allowed us to perceive several data flaws and deduce rules to be implemented in the reproducible pipeline to treat the data. For instance, verify the patient's age on the date of vaccination through his birth date, relating the code of race and color with its description.

Furthermore, we coded other rules regarding the consistency of the dataset. For example, verify the correctness of the patient's age informed on the day of vaccination corresponds with the age calculated through the date of birth. We noticed duplicated records in the dataset. Thus, checking and ensuring the uniqueness of patients' identification and removal of copies was utterly necessary.

Other consistency checks regarding patient data were executed, such as Testing if he has been vaccinated more than once with the same dose of vaccine. Check if he has been vaccinated more than twice with different registries. Besides, we also check if the patient was vaccinated during the second half of 2020 during the vaccines validation pre-tests.

We also verify geo-referenced data to check if the IBGE code of the municipality corresponds to the name of the municipality in the patient's records or validate it with the country code in the IBGE database. Finally, verify if the code, name, and manufacturer of the vaccine correspond with the data from the BMH.

3.4 Problems Detected in the Raw Datasets

One technical problem faced was that processing a large volume of data requires powerful computational resources to perform the analysis. Other than that, there are often gaps in the raw open data

[5]Comprehensive Knowledge Archive Network (CKAN): https://ckan.org/

provided by the Brazilian government, which were present in this dataset. Regarding the dataset, it was verified that there were issues with null and missing values, incorrect records, and inconsistencies in most fields. In addition, it was also identified the absence of fields, some descriptions and category items in the data dictionary, and a lack of metadata patterns and relationships with other datasets. These problems justified the revision and elaboration of a new data dictionary that is presented in the 4.1 section.

Null and missing values were identified using statistical analysis tools and represented 2.02% of the records. Incorrect records and inconsistencies represented 19.18% and were found through semi-automatic exploration of the database with summarization tools, data distribution analysis, queries, counts, and groupings. For these tasks, we used Panda's library functions in Python.

During the data exploration, we attempted to understand the data patterns. We found a lack of standardization between fields. Other problems were identified in the fields: extra spacing at the ends, record duplication, and invalid values, like names and emails typed in the vaccine batch field, and invalid ZIP code according to the Brazilian Postal Office database.

Another problem is intrinsic to the nature of the data used in the HIS. It concerns an error filling in the code and name fields of the patient's country address (patient_address_copais and patient_address_nmpais). According to the HIS data table, they were often filled in with the code and name of the patient's nationality instead of the patient's address.

Additionally, some data that should not be possible were found: Patients aged below 18 years (2897 patients) and above 115 years (4340 patients); patients without unique identifiers; patients who took the same dose more than once or above three doses and patients who were vaccinated outside the possible time range of the campaign. These findings constitute serious errors and inconsistencies in the records. Discrepancies were also detected between the IBGE code, city name, and country code, and name. These errors correspond to 0.55%. Data related to the patient's CEP and vaccine batch were not ratified due to the lack of an association table to validate these fields, which had a large volume of missing, inaccurate, inconsistent data.

3.5 Data Wrangling, Cleaning and Record Linking

A cleaning, wrangling, and record linkage process were necessary to get a clean and enriched dataset. [Christen 2012] proposed steps (also known as data linkage, entity resolution, object identification, or field matching) which were adopted in this work. The dataset was processed in batches because it was necessary to combine records for the same individual in two distinct databases, and this would be computationally expensive if done on the whole dataset at once. To enrich the data, it was necessary to use an additional IBGE dataset containing the city code, name, and federative unit, available in the IBGE Automatic Recovery System (SIDRA)[6]. A dataset of the nationality of the HIS patients was also created and used for enrichment.

The first task was to standardize the fields according to their type, then convert numerical and temporal data types and remove spacing or nulls. Simple typing errors were corrected, and the duplication of records was evaluated on a case-by-case basis, with removal for cases of similar records or correction when possible, keeping records that configured more than two doses of vaccines with different information.

Then, the registration of fixed values with no correspondence with real values replaced null, unidentified, or incorrect values. It was also applied a standardization when possible. In more specific cases, labels were added to represent missing data to be included in the base for future analysis. All markers and labels are described in the new data dictionary.

[6]IBGE Automatic Recovery System (SIDRA): `https://sidra.ibge.gov.br/home/pmc/brasil`

An example that deserves to be highlighted was identifying records that contained conflicting data in the patient's fields: name of the municipality and the country of the patient's address. The problem was detected when the filled municipality was Brazilian, but the country was other than Brazil. These records were considered filling errors in the form and treated as such since the frequency of occurrences was very low compared to other records in the dataset. Therefore, it was decided to replace the country's name with the one following the information contained in the field of the municipality.

Additionally, it was verified that the fields: CNES code, patient care group code, dose description, vaccine code, and vaccine name did not have null data, so they were more easily treated, and some served as support for the reconstruction of missing data by comparison.

4. CURATED DATASETS ENRICHED WITH PROVENANCE METADATA

In this section, the product resulting from the adopted method is presented. It consists of the main dataset, a data dictionary - built after the enrichment - and two additional datasets. We also explain the aggregation of retrospective provenance metadata and discuss the first steps of the process towards compliance with the FAIR data principles. The datasets and their complementary files are available in a repository at: `https://doi.org/10.5281/zenodo.5193920`[7].

4.1 Data dictionary

The novel data dictionary is shown in Table I. It was rebuilt according to the needs that emerged during the application of the methodology. During the activities, we checked the original data dictionary to understand the content of the fields, sometimes the information did not exist, and in others, it was incomplete. The existence of Null values represented a problem in this research. Analysis using them would provide unreliable results. With that in mind, we included and categorized these Null values in the data dictionary, creating new labels and categories for values without information in the fields.

4.2 Complementary Datasets

During the development of the new data dictionary and the treatment of the raw dataset, the authors identified that it would be advantageous to aggregate data from additional datasets. These datasets would aid in analyzing and investigating the primary dataset on fields related to nationalities and Brazilian municipalities. These datasets were included in this work, as well as new data dictionaries for them.

The dataset of municipalities was created from SIDRA/IBGE, containing the IBGE code, the name, and the corresponding federative unit of the municipality. The dataset was used to correct the IBGE code of the municipality and the federative unit of the patient's and the healthcare unit's addresses, as shown in Table II.

The nationality of patients dataset was constructed from information in hospital systems. These data were used to reference the patient's country code and name, as shown in Table III.

4.3 Applying FAIR and Provenance on the Curated Datasets

According to [Cruz et al. 2009], the concept of provenance initially referred to the explanation of the origins of art objects and has recently been incorporated by Computer Science and e-Science. It is being gradually disseminated and used in the areas of Data Science and Machine Learning. According to [Buneman et al. 2001], the classic definition of data provenance is complementary metadata about a

[7]Dataset: `https://doi.org/10.5281/zenodo.5193920`
[8]Best viewed at: https://doi.org/10.5281/zenodo.5193920

Table I. Representation of the new data dictionary of the Vaccination database[8].

Vaccination Database Data Dictionary				
About	**Field names**	**Description**	**Type**	**Category**
Record	document_id	Anonymized unique identifier of the vaccination record	standardized	
Paciente	paciente_id	Anonymized unique patient identifier	standardized	
	paciente_idade	Patient age at vaccination date	integer	
	paciente_dataNascimento	patient birth data	date and time	
	paciente_enumSexoBiologico	biological sex of the patient	category	'F': 'FEMALE'; 'M': 'MALE', 'I': 'UNINFORMED'
	paciente_racaCor_codigo	patient race and color code	integer	1: 'BRANCA'; 2: 'PRETA'; 3: 'PARDA'; 4: 'AMARELA'; 5: 'INDIGENA'; 99: 'SEM INFORMACAO'
	paciente_racaCor_valor	Vaccinated Race (White, Black, Brown, Yellow, Indigenous and No information)	category	1: 'BRANCA'; 2: 'PRETA'; 3: 'PARDA'; 4: 'AMARELA'; 5: 'INDIGENA'; 99: 'SEM INFORMACAO'
	paciente_endereco_coIbgeMunicipio	IBGE code of the municipality of the patient's address	integer	reference municipio.csv and uninformed: 999999
	paciente_endereco_coPais	country code of the patient's address nationality	integer	reference nacionalidadepaciente.csv and uninformed: 0
	paciente_endereco_nmMunicipio	city name of the patient's address	alphanumeric	reference municipio.csv and uninformed 'SEM INFORMACAO'
	paciente_endereco_nmPais	name of the country of nationality of the patient's address	alphanumeric	reference nacionalidadepaciente.csv and uninformed: 'SEM INFORMACAO'
	paciente_endereco_uf	federative unit of the patient's address	category	reference municipio.csv and uninformed: 'XX'
	paciente_endereco_cep	patient care group code	integer	não informado: 0
	paciente_nacionalidade_enumNacionalidade	enumerated list of the patient's nationality	category	'B': 'BRASILEIRO'(BRAZILIAN); 'E': 'ESTRANGEIRO'(FOREIGN), 'I': 'SEM INFORMAÇÃO'(NO INFORMATION)
Establishment	estabelecimento_valor	CNES code	integer	references the CNES table
	estabelecimento_razaoSocial	business name of the establishment	alphanumeric	
	estalecimento_noFantasia	business name of the establishment	alphanumeric	
	estabelecimento_municipio_codigo	IBGE code of the municipality of the establishment	integer	reference municipio.csv and uninformed: 999999
	estabelecimento_municipio_nome	name of the municipality of the establishment	alphanumeric	reference municipio.csv and uninformed 'SEM INFORMACAO'
	estabelecimento_uf	federative unit of the establishment	category	reference municipio.csv and uninformed: 'XX'
Vaccine	vacina_grupoAtendimento_codigo	patient care group code	integer	
	vacina_grupoAtendimento_nome	patient care group name	alphanumeric	
	vacina_categoria_codigo	patient category code	integer	
	vacina_categoria_nome	patient category name	alphanumeric	
	vacina_lote	vaccine batch	alphanumeric	
	vacina_fabricante_nome	name of vaccine manufacturer	alphanumeric	
	vacina_fabricante_referencia	alphanumeric vaccine manufacturer reference	alphanumeric	
	vacina_dataAplicacao	vaccine application date	date and time	
	vacina_descricao_dose	description of the applied dose	category	'PRIMEIRA DOSE'(FIRST DOSE); 'SEGUNDA DOSE'(SECOND DOSE); 'DOSE ÚNICA'(SINGLE DOSE)
	vacina_codigo	vaccine code	integer	81: 'Inválido'(Invalid); 85: 'Vacina Covid-19-Covishield'; 86: 'Covid-19-Coronovac-Sinovac/Butantan'; 87: 'Vacina Covid-19-BNT162b2-BioNTech/Fosun Pharma/Pfizer'; 88: 'Vacina Covid-19-Ad26.COV2.S-Janssen-Cilag'
	vacina_nome	description of the applied dose	category	
Record	sistema_origem	health system of origin of the vaccination record	alphanumeric	
	data_importacao_rnds	RNDS import date	date and time	

Table II. Representation of the City Data Dictionary.

Conteúdo de Dados do Município a partir do SIDRA/IBGE			
Field name	**Description**	**Type**	**Category**
coIbgeMunicipio	IBGE code of the municipality of the patient's address	integer	
nmMunicipio	city name of the patient's address	alphanumeric	
uf	acronym of the federative unit	category	UF acronym

Table III. Representation of the Dictionary of the patient's nationality.

Data Content in relation to the patient's nationality		
Field name	**Description**	**Type**
Code	Country code	integer
Nationality	name of the country	alphanumeric

data or process, containing information on how, when, where, and why the data were obtained and who produced them. Data Provenance captures information about what happened during the composition or execution of workflows to support the reasoning for trust and reproducibility. Currently, provenance

metadata is one of the indicators of data quality. However, it is still missing in many works in Data Science [Sikos and Philp 2020] and also in FAIRification workflows [Jacobsen et al. 2020].

In order to reduce this gap, in this work, we advocate the use of the W3C PROV model [Missier et al. 2013] as the conceptual method to illustrate the provenance of the Data Science pipelines. The PROV model was conceived to model the provenance of web applications, but it is flexible enough to be applied in many scientific domains, including Data Science and Machine Learning. The model represents the provenance of the steps of a pipeline as a graph whose records show how agents, processes, entities, and activities are related.

The provenance records are an indicator that shows the enhanced quality of this work. They describe the sequences of processes (tasks), the ingested inputs files used during the execution of the pipeline, and also the outputs datasets.

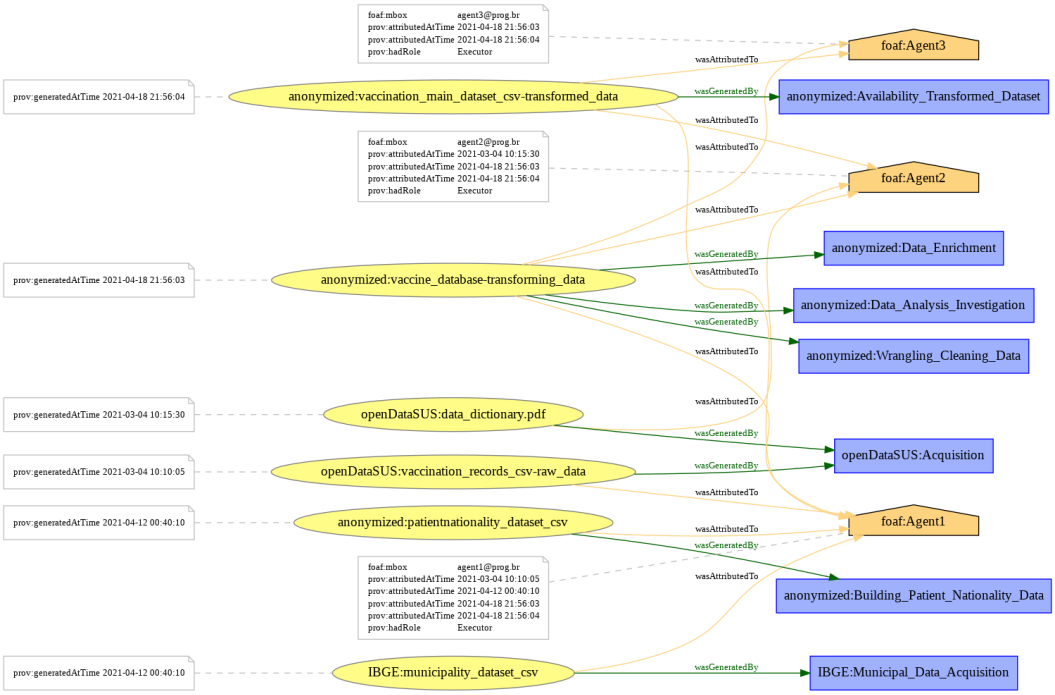

Fig. 2. Fragment of Provenance Graph of the Data Science pipeline according to W3C PROV model.

Figure 2 shows the history of the experiments in the form of a retrospective provenance graph based on the PROV model. It illustrates the relationship between the ingested raw datasets, the curated datasets, and the execution of the pipeline created during this work. The agents, activities, and entities and their notes and dependencies are also illustrated. Three different agents (orange polygons) were involved during the production of the cured vaccination datasets. These agents performed seven raw data processing tasks (blue boxes), represented as six entities (yellow ellipsis).

The Figure interpretation is made as follows: Agents 1, 2, and 3 carry out the Data_Analysis_Investigation task under the vaccine_database-transforming_data entity. In each execution of the pipeline, the vaccine_database-transforming_data entity will originate new data. Each step of an execution performs a single activity with a specific notation.

4.4 FAIR Data Pipeline

The FAIR principles were proposed as guidelines to make datasets more Findable, Accessible, Interoperable, and Reusable so that both researchers and machines are able to find, access, and (re)use data [Wilkinson et al. 2016]. Datasets should be made as FAIR as possible, even if it is not possible to make data open [Landi et al. 2020][Romain et al. 2020].

The principles indicate characteristics applicable to humans, machines, and technological resources like software tools, vocabularies, and data infrastructure to simplify the discovery and usage of any given data or metadata. The principles do not address issues related to data quality, while this is a significant concern when improving data usability. They focus mainly on mechanisms to facilitate data sharing. These principles are composed of sub principles and mainly aim to make the research data readable by machines and humans.

FAIR principles have been gaining increasing acceptance in the field of Data Science and workflows [Landi et al. 2020]. In order to be as compliant as possible with the principles, the following computational steps were followed: the cured dataset, the complementary datasets, and their files were described by rich metadata, registered and indexed in a searchable resource accessible to potential users (humans or machines). Zenodo[9] was used, whose metadata is in accordance with DataCite's Metadata Schema[10].

In Zenodo, a Digital Object Identifier (DOI) was assigned to the dataset. The identifier(DOI) allows persistent links to be established between data, metadata, and other related materials to aid in the registration and the indexing in search tools. Such an approach provides compliance with the Findable principle of FAIR.

Given the proper authorization, files can be browsed or downloaded through Zenodo's interface and may also be accessed through the REST API protocol. Zenodo keeps the metadata of hosted data even if the data becomes unavailable, which provides compliance with the Accessible principle of FAIR.

Datasets and their metadata were described using JSON Schema. In this way, knowledge is represented in a known standard vocabulary. The dataset files are available in open CSV and TXT formats, and this facilitates reuse and provides compliance with the Interoperable principle of FAIR.

For data to be reusable, it must be released with a clear and accessible license for use and must have provenance trails. In the case of this work, the Creative Commons CC-BY was used, which allows distributing, altering, or reinventing the datasets as long as the initial source is cited. This license was used to follow the license of the raw dataset. Rich metadata, data provenance, and complementary documentation were available and versioned to reuse. These items have been provided in Zenodo, which meets the relevant standards of the community of interest and provides provenance information. This provides compliance with the Reusable principle of FAIR.

5. DISCUSSION

The richness of the datasets about Covid-19 vaccination is essential in global or local epidemiological studies to understand the pandemic dynamics. Our datasets and data dictionaries are curated, uniquely identified, and reflect the temporal aspects of the beginning of the Brazilian vaccination campaign. Besides, they are findable, accessible, interoperable, and sharable, allowing civil and academic groups to manipulate them openly and allowing third-party investigations.

If one tried to use the raw vaccination dataset (available on OpenDataSUS) for analysis, visualization

[9]Zenodo: https://zenodo.org/
[10]DataCite's Metadata Schema: https://schema.datacite.org/

tools would not show information properly. This is because the raw dataset had several issues (as discussed in section 3.4). It would be necessary to discard many records to analyze the data correctly. However, in this work, we focus on the opposite. We developed standardized methods and used data linkage to preserve the highest number of vaccination records possible. Furthermore, we are increasing the geo-referenced content to offer a visual monitoring tool for health specialists and decision-makers to perceive the vaccination dynamics easily in digital geographical maps at the level of the Brazilian municipalities.

To evaluate the correctness of the curated dataset, we offer graphical representations that contain the information and the curated data. We analyze and visualize the following variables: age, gender, color, professional categories available.

Figure 3 illustrates the visualizations of the first phase of vaccination of the Brazilian people, which began on January 19, 2021, with the priority groups (health workers, institutionalized people (residing in nursing homes) aged 60 years or over, institutionalized people with disabilities, and indigenous population in villages[11]). We can observe that until April 2021, the amounts of unique dose applications in the country were low and barely visible.

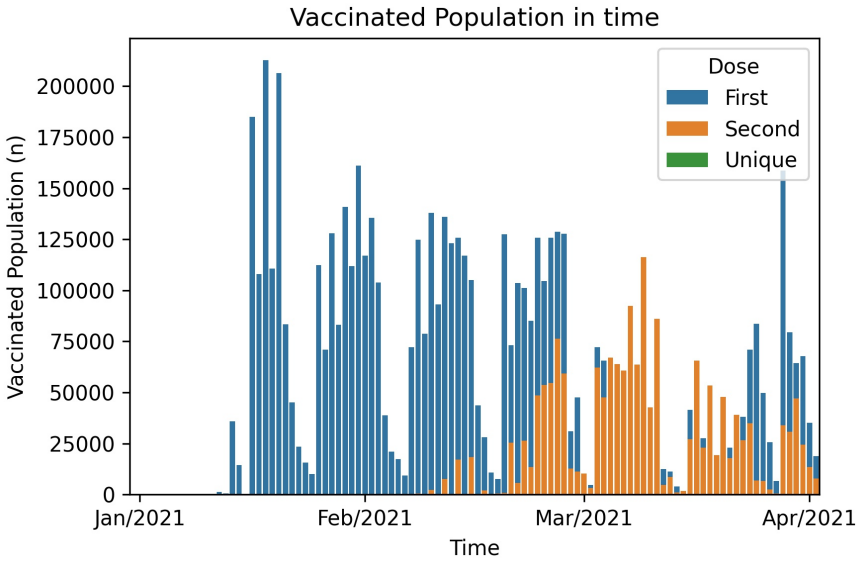

Fig. 3. Distribution of the Vaccination process of the Brazilian population in time.

Figure 4 illustrates the number of males and females divided by age groups, and we can observe that until April 2021, the number of vaccinated women was significantly higher than that of men. This disparity is nothing new, as women tend to care more about their health than men [12]. This reflects on the life expectancy, which is 80.1 years for Brazilian women and 73.1 years for Brazilian men[13].

From our point of view, the main weakness found in the original dataset was a large amount of invalid (null) data in the records. For instance, critical information like vaccine batch ID were invalid or blank. We also highlight that ZIP codes were fuzzy, as some Codes did not exist or were incorrect. This makes it challenging to analyze the geographic distribution of the vaccination campaign. We

[11]Agência Brasil EBC: https://www.unasus.gov.br
[12]COVID-19 Sex-Disaggregated Data Tracker: https://globalhealth5050.org/the-sex-gender-and- covid- 19-project / the-data-tracker/
[13]Folha de S.Paulo: https://folha.com/z370dxea

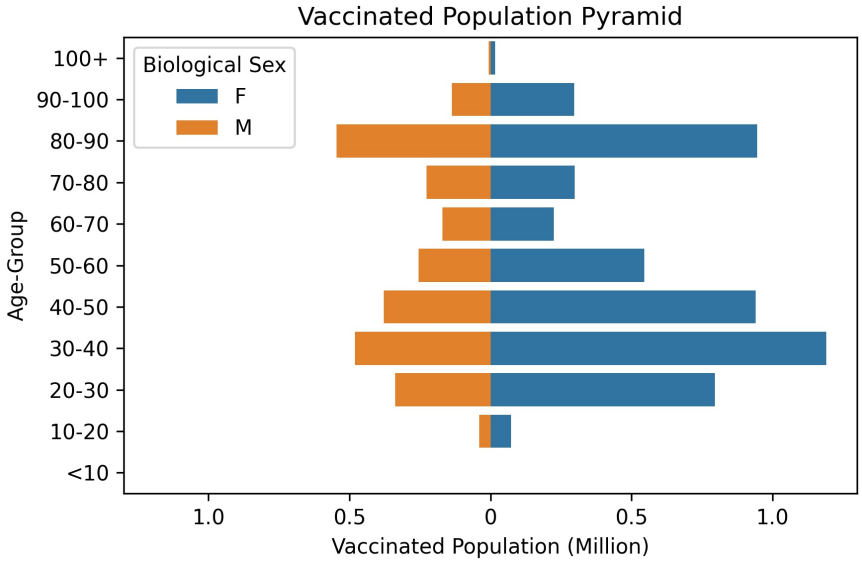

Fig. 4. Visualiztion of the distribution of the Vaccination as a population pyramid.

could circumvent these deficiencies if the National Data Network offered control files from vaccine batches and supplementary geographical data.

6. FINAL REMARKS

Brazil is a country that harbors great inequalities, especially in health and education, and this was reflected in the problems found in the data entries in the SIS used by the vaccination campaigns. The high inaccuracy of the raw dataset served as the initial motivation for developing the pipelines of this work, which involved data analysis, investigation, cleaning, linking, and correction, resulting in curated and open datasets. The resulting data represents an opportunity for society, especially researchers, to understand the dynamics and problems of the vaccination campaign and the fight against COVID-19 in Brazil regarding management, application of vaccines, temporal evolution, and geographic distribution.

This paper presents contributions to developing the data science pipelines that produce curated datasets enriched with provenance and aligned with the FAIR principles about the vaccination campaign against COVID-19 in Brazil. It includes the complementary datasets, crucial for understanding the data and for future studies. The new data dictionaries are other essential products of this work. The availability of these datasets with metadata allows easy reusability of the data for different purposes.

We emphasize that data provenance and FAIR principles played an essential role in this work, becoming more relevant if used by collaborative groups, allowing the creation of a baseline for future research and generating analyses, visualizations, contributions to investigations, and searches for continuous improvements on the data, which are licensed under Creative Commons CC-BY.

The activities were developed in an iterative way involving several pipeline executions on the data source in a computational infrastructure of servers. This allowed the verifiability of the developed pipeline and exposed the ability to reproduce the results in a consistent, authentic, and transparent way. Indirectly, this work makes it possible to enhance the discussions on actions to fight the disease and its impacts, inspect the vaccination process, and analyze the data to assess the decision-making process.

Understanding this fast-moving landscape is particularly challenging and relevant, as most of these datasets have not yet been investigated through a peer-review process or even correctly handled and processed by computer. These datasets are expected to assist in analyzing, researching, and monitoring vaccination progress against COVID-19 in Brazil, helping to understand the process and support health policymakers. We count on the collaboration of the science community to improve the method and continually update the datasets if significant inaccuracies are found.

Understanding this fast-moving landscape is particularly challenging and relevant, as most of these datasets have not yet been investigated through a peer-review process or even correctly handled and processed by computer. These datasets are expected to assist in analyzing, researching, and monitoring vaccination progress against COVID-19 in Brazil, helping to understand the process and support health policymakers. If significant inaccuracies are found, we count on your collaboration to improve the method and continually update the datasets.

As future works, we plan to include new datasets with vaccination records from the last one recorded in this work to the most recent ones, including vaccination boosters. This will allow the entire vaccination campaign to be covered, which will require continuous and periodic processing. We also intend to make specific cuts by regions or states, including geographic data, to expand analysis and visualization possibilities. This work may become a tool, which can be expanded as new data on the campaign are produced.

7. ETHICS STATEMENT

Health and vaccination data are of real value for COVID-19 scientific research, and there is an urgent need to reconcile the benefits of data sharing with privacy rights and constraints and ethical and regulatory requirements. The authors confirm compliance with the ethical policies of the journal and Brazilian laws. No ethical consents or approvals were required because this study did not involve any experimental protocol on humans or animals or used identification records of individuals. We used only anonymous publicly-available open data about vaccination (freely usable, reusable, and redistributable without restrictions and copyrights)

Acknowledgements

The Coordination partially funded this study for the Improvement of Higher Education Personnel - Brazil (CAPES-Tecnodigital) - Financial Code: 88887.514128/2020-0 and partially sponsored by the National Education Development Fund (FNDE), Educational Tutorial Program PET-SI / UFRRJ), National Council for Scientific and Technological Development (CNPq) - DT-II Scholarship (315399 / 2018-0), and Carlos Chagas Filho Research Foundation (FAPERJ) – Code E-26/210.192/2020.

The computational environment used was provided by the Sergio Arouca National School of Public Health (ENSP) of the Oswaldo Cruz Foundation (Fiocruz) and the logical network support provided by the logical network team of the IT Management Service (SGTI/ENSP/Fiocruz).

REFERENCES

Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P. B., Joe, B., and Cheng, X. Artificial intelligence and machine learning to fight covid-19, 2020.

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

Bernal, J. L., Andrews, N., Gower, C., Robertson, C., Stowe, J., Tessier, E., Simmons, R., Cottrell, S., Roberts, R., O'Doherty, M., et al. Effectiveness of the pfizer-biontech and oxford-astrazeneca vaccines on covid-19 related symptoms, hospital admissions, and mortality in older adults in england: test negative case-control study. BMJ 373 (1), 2021.

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

Christen, P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection / peter christen, 2012.

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

Cruz, S. M. S., Campos, M. L. M., and Mattoso, M. Towards a taxonomy of provenance in scientific workflow management systems. In 2009 Congress on Services - I. pp. 259–266, 2009.

Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A., and Larson, H. The pandemic of social media panic travels faster than the covid-19 outbreak, 2020.

Doyle, R. and Conboy, K. The role of is in the covid-19 pandemic: A liquid-modern perspective. International Journal of Information Management vol. 55, pp. 102–184, 2020.

Fridman, A., Gershon, R., and Gneezy, A. Covid-19 and vaccine hesitancy: A longitudinal study. PLOS ONE 16 (4): 1–12, 04, 2021.

Gonçalves, M. V., dos Santos, J., Ferreira, C., Zavaleta, J., Cruz, S. M. S., and Sampaio, J. Datasets curados e enriquecidos com proveniência da campanha nacional de vacinação contra covid-19. In Anais do III Dataset Showcase Workshop. SBC, Porto Alegre, RS, Brasil, pp. 148–159, 2021.

Ienca, M. and Vayena, E. On the responsible use of digital data to tackle the covid-19 pandemic. Nature medicine 26 (4): 463–464, 2020.

Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L. O. B., Mons, B., Schultes, E., Roos, M., and Thompson, M. A Generic Workflow for the Data FAIRification Process. Data Intelligence 2 (1-2): 56–65, 01, 2020.

Landi, A., Thompson, M., Giannuzzi, V., Bonifazi, F., Labastida, I., da Silva Santos, L. O. B., and Roos, M. The "A" of FAIR – As Open as Possible, as Closed as Necessary. Data Intelligence 2 (1-2): 47–55, 01, 2020.

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. Vacinação do Cardiopata contra COVID-19: As Razões da Prioridade. Arquivos Brasileiros de Cardiologia vol. 116, pp. 213–218, 2021.

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., and Rodés-Guirao, L. A global database of covid-19 vaccinations. Nature human behaviour 1 (5): 1–7, 2021.

Ministério da Saúde – Brasil. 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, 2021. [Acessado em 13 abr. 2021].

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

Oliveira, L. A., Muraro, R., Cristina, A. P., Andrade, A., Cecconello, S., and Lalucci, M. M. Vacinação contra a covid-19 em mato grosso: primeiros resultados. Nota Técnica - Universidade Federal de Mato Grosso, 06, 2021.

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. Plano nacional de vacinação contra a covid-19: uso de inteligência artificial espacial para superação de desafios. Ciência & Saúde Coletiva vol. 26, pp. 1885–1898, 2021.

Romain, D., Laurence, M., Alison, S., Stryeck, S., Mogens, T., Mohamed, Y., Clement, J., Laurent, D., Daniel, a. J., Daniel, e. B., Elena, B., Sophie, G., Hannah C., G., Jean-Eudes, H., Vassilios, I., Yvan, L. B., Emilie, L., and Anne, C.-T. Fairness literacy: The achilles' heel of applying fair principles. Data Sci. J. vol. 19, pp. 32, 2020.

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

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

Tagoe, E. T., Sheikh, N., Morton, A., Nonvignon, J., Sarker, A. R., Williams, L., and Megiddo, I. Covid-19 vaccination in lower-middle income countries: national stakeholder views on challenges, barriers, and potential solutions. Frontiers in Public Health, 2021.

Vasileiou, E., Simpson, C. R., Shi, T., Kerr, S., Agrawal, U., Akbari, A., Bedston, S., Beggs, J., Bradley, D., Chuter, A., et al. Interim findings from first-dose mass covid-19 vaccination roll-out and covid-19 hospital admissions in scotland: a national prospective cohort study. The Lancet 397 (10285): 1646–1657, 2021.

Victora, C. G., Castro, M. C., Gurzenda, S., Medeiros, A., Franca, G. V., and Barros, A. J. Estimating the early impact of immunization against covid-19 on deaths among elderly people in brazil: analyses of routinely-collected data on vaccine coverage and mortality. medRxiv, 2021.

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. The fair guiding principles for scientific data management and stewardship. Scientific data 3 (1): 1–9, 2016.