Opening Brazilian COVID-19 patient data to support world research on pandemics

AUTHORS: Luiz Eugenio Mello (https://orcid.org/0000-0002-6969-1108), Andrea Suman, Claudia Bauzer Medeiros (https://orcid.org/0000-0003-1908-4753), Claudio Prado, Edgar Rizzatti (https://orcid.org/0000-0002-4041-7543), Fatima L. S. Nunes (https://orcid.org/0000-0003-0040-0752), Gabriela Barnabé (https://orcid.org/0000-0001-5341-0984), João Eduardo Ferreira (https://orcid.org/0000-0001-9607-2014), José de Sá, Luiz Fernando Reis (https://orcid.org/0000-0002-4040-2456), Luiz Vicente Rizzo (https://orcid.org/0000-0001-9949-9849), Luzia Sarno, Raphael de Lamonica, Rui M. B. Maciel (https://orcid.org/0000-0002-6123-6098), Roberto Marcondes Cesar Jr (https://orcid.org/0000-0003-2701-4288), Rodrigo Carvalho

ABSTRACT

This paper describes the COVID-19 DataSharing/BR initiative, a pioneer public-private partnership to publish open data on Brazilian COVID-19 patients. Constructed in record time, it has been launched with clinical, laboratory and diagnostic data from approximately 177,000 Brazilian individuals, in answer to researchers' demand for quality data. COVID-19 DataSharing/BR was created by a consortium led by FAPESP (Sao Paulo Research Foundation) and USP (University of Sao Paulo, Brazil), with participation from three major private health institutions in Brazil – Fleury Institute, Sírio-Libanês Hospital and Albert Einstein Hospital. Launched on July 1st, 2020, within 10 days it had already been subject to 800 downloads from 14 different countries.This text provides a brief description of the initiative, and initial efforts for preprocessing and publishing the data according to legal and interoperability constraints. The COVID-19 DataSharing/BR repository took only one month from inception to delivery, thanks to the support of a pre-existing extensible open research data e-infrastructure.

1.Introduction and motivation

Ever since the beginning of the COVID pandemic, many initiatives have been launched to publish health data on COVID patients, to help researchers understand the virus, and work towards containing its spread, finding a cure, or preparing for its aftermath, to name but a few. Examples of such data sharing efforts include, for instance, the United States' National Covid

Cohort Collaborative (N3C) project [1], launched in the beginning of June, or the secure analytics mediator for the COVID-19 Future Operations Clearing Board [2, 3] in Austria.

N3C [1] is an analytics platform coordinated at NIH's National Center for Advancing Translational Sciences, being an initiative to provide analytics on clinical, laboratory and diagnostic data from hospitals and health care plans. Data are being transferred to the N3C secure servers from health facilities all over the United States. They are then preprocessed according to OHDSI/OMOP standards, to allow joint analyses, and stored in secure storage systems, to which access is given only to analytics software. Access to analytics facilities requires registration and authorization. The Austrian Platform [2], still under construction, centralizes a wide variety of patient, clinical, equipment and other kinds of COVID-19 related data, for research and policy purposes. The University of Vienna developed OSSDIP [3], a high-security data infrastructure that ensures that only authorized users have access to sensitive data, while non-sensitive information can be accessed freely. This preserves security even where anonymization is not possible.

Frequently, statistics are presented as dashboards or infographics, but data records are not provided. An exception seems to be the South Africa Dashboard [4], for which individual records, including travel information, are available until March 23, and from then on only aggregate data are made public via GitHub. Still other platforms offer a variety of open data sources, such as IEEE DataPort's COVID-19 datasets [5], in which researchers can upload a wide variety of COVID-19 related sources, such as tweets or images.

The abovementioned national initiatives [1,2,4] were cited because, like FAPESP's COVID-19 DataSharing/BR, they are concerned with clinical and patient data, as opposed to most platforms that provide data and/or analytics on pathogens, or gene sequences.

The FAPESP COVID-19 DataSharing/BR initiative [6], however, offers a distinct solution. Rather than controlling data access, publishing data aggregates, or preprocessing them for homogenization, it publishes raw data on patients that have undergone COVID-19 testing in Brazil, the vast majority of which Brazilian residents. Its datasets encompass clinical, laboratory and diagnostic information, as well as some demographic data. All records are pseudonymized to meet international standards and Brazilian laws. This repository was constructed following FAPESP's Open Science policies; its main goal is to contribute to accelerate COVID-related research all over the world through sharing data on Brazilian patients.

FAPESP (Sao Paulo Research Foundation) [7] is a public foundation, funded by the taxpayer in the State of São Paulo, with the mission to support research projects in higher education and research institutions and companies, in all fields of knowledge. The stability of the funding and the autonomy of the foundation allow for an efficient management of the resources that has had a sizable impact: while São Paulo has 22% of the Brazilian population and 30% of the scientists with a doctorate in the country, the state responds for 45% of the country's scientific articles published in international journals. Today FAPESP is one of Brazil's major funding agencies. The central role of Fapesp in research funding in Brazil and the credibility it amassed over many

decades lends itself to a central role on a data sharing initiative. The recent meeting of the Global Research Council (GRC) in 2019 held at FAPESP and the election of former FAPESP's Scientific Director, as Chair of the Governing Board of the GRC further underlines its importance in the global arena.

FAPESP responded very quickly to the pandemic crisis, through two major fronts - (a) creating a new program to fund COVID research as early as March 2020; and (b) launching the repository, in collaboration with the University of São Paulo, initially with data from three major private health institutions - the Fleury Group, the Albert Einstein Hospital and the Sirio Libanes Hospital. This second initiative was enabled through a multi-party agreement that ensured the appropriate legal and administrative conditions.

With more than 90 years of history, Grupo Fleury [8] is one of the largest and most respected medical and health organizations in Brazil, reference to the medical community and public opinion for its excellence in customer service, innovation and technical quality. Fleury's corporate values include a commitment to sharing knowledge and the ability to put ourselves in the place of others and truly understand their situation. Bearing this in mind, the Group develops social initiatives with its surrounding communities geared towards health and education and trains its employees to take part in the corporate volunteer program. Since its inception Grupo Fleury has had strong ties with academia and knowledge production. Its participation in the current data sharing initiative was hence a natural implication of this commitment.

The Albert Einstein Hospital [9] in Brazil was established by a group of Jewish community members in São Paulo in 1955. Its construction began three years later and the hospital was inaugurated on 28 July 1971.In 1999, it was the first health institution outside the United States to be certified by the Joint Commission International [10]. Since then, the hospital has become a reference in state-of-the-art treatments and humanized care, and has expanded its borders with social responsibility actions, education, and research activities. Today, the Sociedade Beneficente Israelita Brasileira Albert Einstein is at the forefront of important projects, which shows how the public-private partnership can yield fruits for the community, inspiring other institutions to join in for the health of Brazil. The creation of the Covid data sharing initiative was promptly embraced by the hospital as an important initiative.

Hospital Sirio Libanes Hospital (HSL) [11] is a private, not-for-profit, philanthropic institution established in 1921, as an initiative from the syrian and lebanese communities in Sao Paulo. It acts in medical assistance, research and education and, by virtue of its academic activities, HSL hopes to contribute, with knowledge and human resources, for the improvement of the public health system.

The logistics of creating and organizing the repository relied on a two-tiered management strategy - a group of 6 scientists was appointed to manage administrative and COVID-19 research issues, whereas a group of computer and data scientists was appointed to design and implement the repository, and preprocess the data. The authors of this white paper are the members of these two groups. Communications and discussions between them are regularly maintained through scientists that participate in both.

In its first stage, the repository made available curated data on approximately 177,000 patients, together with information on outcomes (primary endpoints), demographic data, and patient transfers, corresponding to approximately 4,2 million records that were preprocessed to meet Brazil's and international data protection laws - see Table 1 for the exact numbers. It will subsequently contain additional data types and sources, including medical images of COVID-19 patients, and associated reports. Data will be periodically updated throughout the pandemics, representing an exemplary open science effort within a public-private partnership. As discussed in section 4, this repository was delivered in a very short time, thanks to the availability of an extensible open data publishing platform, the Network of Research Data Repositories of the State of São Paulo.

The rest of this paper is organized as follows. Section 2 provides an overview of implementation infrastructure, and data preprocessing activities at each participating institution. Section 3 contains some statistics on downloads and data usage and Section 4 outlines ongoing work.

2. Implementing the Repository

2.1 Design decisions, and data characteristics

The pandemics gave rise to world-wide discussions on conditions and constraints under which COVID-19 data might be shared, and its ethical uses. In order to create the repository, the team followed the recommendations of the Research Data Alliance on sharing these specific kinds of data [12].

Given initial data requirements, provided by domain scientists, and the experience of the University of São Paulo in creating a major infrastructure for publishing open research data, the following design and implementation decisions were taken:

- The repository would be incorporated into the Open Research Data Repository Network of the State of Sao Paulo [13] (see section 2.2). Its data would be open, and preprocessed to avoid identification of individual patients, in compliance with international and Brazilian data protection laws;

- The kinds of records made available would be those for which all institutions already had significant amounts of data, thereby providing an initial sizeable sample for researchers all over the world;
- The records should include information on Patients (such as age, gender, spatial location), as well as clinical and laboratory Exams, together with information on Primary Endpoints and patient Transfers;
- Data would be provided on individuals who had undergone COVID-related testing in Brazil since the pandemic outbreak in the country;
- Exam records would cover exams starting November 1st, 2019, e.g., to allow trend analyses, and studies of comorbidities;
- Any Brazilian health institution would be welcome to join the initiative, depositing data in the repository, according to security and consistency rules defined by the computer science researchers invested in the design, implementation and maintenance of the repository.

Furthermore, following international standards for interoperability, data should be documented through associated metadata and data dictionaries, providing all mandatory fields established by the Sao Paulo Repository Network (namely Author, Title, Description, Keywords/Subject, URI, Persistent Identifier, File type and Funding information) to allow findability.

2.2 Computational Infrastructure

The computational infrastructure for the COVID-19 DataSharing/BR repository relies on the computational platform of the Open Research Data Repository Network of the State of Sao Paulo [13]. This platform was designed and implemented by a Working Group instituted by FAPESP in 2017, composed of representatives of the state's six public universities (University of Sao Paulo, University of Campinas, Sao Paulo State University, Federal University of Sao Carlos, Federal University of ABC, and Federal University of São Paulo, Unifesp), and of the Aeronautics Institute of Technology. In 2018, the workforce was joined by CNPTIA (the Informatics Research Center of the Brazilian Agricultural Research Corporation, EMBRAPA). This network, available since december 2019 at https://metabuscador.uspdigital.usp.br/, hosts scientific data from research produced by the network members in all scientific domains.

Its architecture is similar to that of a federated database system [14], being composed of two major elements: (a) repositories of the participating institutions, which are designed and maintained independently; and (b) a metadata search engine, that daily harvests and exposes through a common interface information on metadata available at each institution. All metadata of the "federation" are stored at a metadata repository. This search engine was developed by the University of São Paulo, where the metadata repository is maintained. Metadata are transferred using the standard OAI-PMH protocol for metadata harvesting [15].

Thanks to the availability of this open research data infrastructure, which took 2 years to design and build, FAPESP was able to give a prompt response to the pressing need for reliable data for research on COVID-19. Through it, the FAPESP COVID-19 Data Sharing/BR repository became yet another member of the federated architecture.

The design and implementation of the COVID-19 repository was developed and maintained by the University of São Paulo, considering the following design and implementation decisions:

- Data selection and preprocessing are performed by the institutions providing data, so that data deposited are already de-personalized. Users have no access to source data (which remain at the respective institutions), just to the repository data. Thus, the security of the original data sources is ensured;
- The repository is partitioned per institution, thereby allowing each partner to independently version and deliver their data, without requiring cross-institutional controls. In the future, this will also allow institutions to add new kinds of data to their partition, when available;
- Also aiming to ensure data security, each institution has an authorized, secure, access protocol to upload its dataset in the repository by using a standard interface where metadata are provided. Figure 1 provides a screen copy of the data upload interface, through which the depositing institution must also provide associated metadata;
- Before actual publication, each uploaded dataset is checked against a set of rules, to ensure its compliance with the repository's depersonalization and naming schemes, and additional consistency checks. Non-compliant datasets are returned to the depositing institution for verification;
- To ensure the physical security of the data, the repository stores the datasets in a cloud structure, with redundancy provided by two datacenters. Additionally, a database team and an infrastructure team perform continuous monitoring to check the data volume growth, to provide additional disk space when necessary;
- A development team periodically performs tests related to scalability and correctness in order to ensure the continuous operation of the repository. This team is also responsible for registering new institutions and including new functionalities related to usability and efficient user access to the data (namely, downloads);
- To further ensure independence among depositing institutions, the datasets provided are self contained. In other words, each institution must always deposit a package of datasets, containing not only the data themselves (e.g., Patient records, Exam records), but also a data dictionary that describes the data in the package. This will ensure that, in the long term, each institution may tailor the contents of its package beyond the initial configuration (e.g., adding images or diagnostic information).

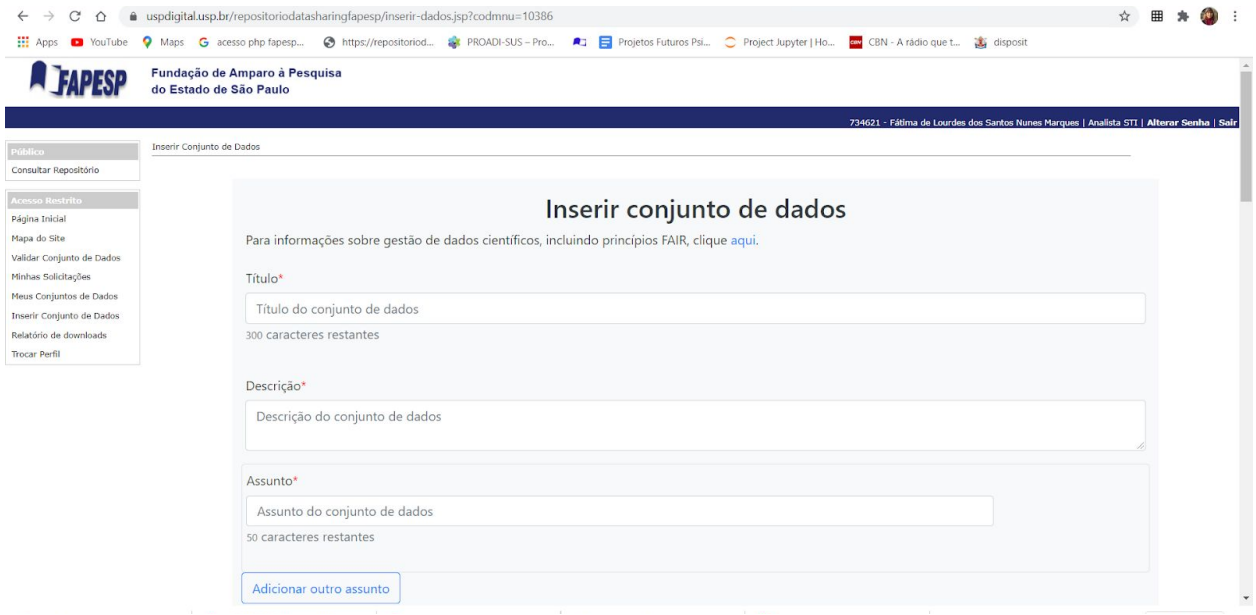

Figure 1 - Interface for uploading datasets. Institutions are prompted to provide metadata for the uploaded datasets - this partial screen copy shows requests for fields for “Title”, “Description”, “Subjects”.

These design and implementation decisions ensure that some of the main challenges are currently successfully bypassed. Storage, availability and maintainability are currently provided by University of São Paulo, but the growth of data volume must be continuously monitored to ensure the allocation of additional resources when necessary.

There are still countless challenges to be met, including the heterogeneity among depositing institutions. Indeed, not only does each have its own data management practices and systems (and thus a wide variety of attributes and records), but many differences across available information, e.g., the types of exams, or value ranges. This heterogeneity will increase as new institutions join the repository. In its first stage, heterogeneity issues were circumvented by limiting the set of records and attributes to those common to all institutions, and establishing basic naming conventions. Nevertheless, as the repository grows, it is expected that each depositing institution will be able to deposit additional kinds of records, thereby supporting a wider range of research questions. This solution is naturally enabled by letting each institution configure and deposit its datasets in self-contained packages. In some sense, it is as if the COVID-19 Datasharing/BR repository were itself a set of repositories, interlinked by a minimum set of predefined attributes and files, naming conventions, and consistency constraints.

To help external access and usage, researchers from the University of São Paulo designed an initial package containing basic documentation and software to allow researchers to download the data and create their own relational databases, which they can then use as a

basis for their analyses. This package is itself published as a research data item within the Sao Paulo repository network [16].

3. Initial Usage Statistics

The COVID-19 DataSharing/BR repository was published in two stages. First, a sample of data on 300 patients was made publicly available for tests and feedback, on June 17 2020. Throughout 2 weeks, these data were downloaded approximately 600 times, mostly from Brazil, but also from North America, Europe and Asia. User feedback was mostly of two natures - (1) request for additional attributes, some of which impossible to provide without violating patient privacy (e.g., through fine-grained spatial accuracy); and (2) request for more detailed metadata.

Given this feedback, the second stage consisted in publishing on July 1st 2020 the first full set of records on approximately 180,000 individuals, for which metadata contents were more detailed (see Table 1). Within two weeks, this repository had been viewed over 1,500 times, with approximately 800 data downloads from 15 different countries. User feedback continues to arrive and, if possible, will be acted upon - see section 4.

Table 1 presents a data summary of the initial full dataset. More than 750 types of clinical and laboratory exams are included. The numbers tend to increase since the dataset will be periodically updated with new records and health data providers.

TABLE 1: Data summary of the initial full dataset.

Data Source	**#Patients**	**#Clinical Exams**	**#Primary endpoints**
Fleury Group	129,597	2,496,592	0
Sirio Libanes Hospital	2,732	371,358	9,634
Albert Einstein Hospital	44,879	1,867,091	0
Total	177,208	4,735,041	9,634

Table 2 shows the terms most frequently used by search strings, two weeks after launching the second stage. Table 3 shows the top ten countries that have visited the repository in the same

period. Figure 2 shows the different types of entity that downloaded datasets from the repository.

Table 2 - The ten most searched terms in the repository, in Portuguese - e.g., pandemic, sorology, exam results and so on (Date: 2020, July 17th).

Order	Search term	Visited Pages	% of Total
1	subject_keyword:pandemia	13116	80.24%
2	subject_keyword:sorologia	10482	64.13%
3	dateIssued_keyword:2020	10204	62.43%
4	subject_keyword:resultados exames	9796	59.93%
5	subject_keyword:coronavirus	9659	59.09%
6	subject_keyword:PCR	9511	58.19%
7	subject_keyword:covid-19	9157	56.02%
8	has_content_in_original_bundle_keyword:true	4180	25.57%
9	author_keyword:Fleury, Grupo	4108	25.13%
10	author_keyword:Albert Einstein, Hospital Israelita	2851	17.44%

Table 3 - The ten countries with most accesses (Date: 2020, July 17th).

Country	# of visits
Brazil	1639
USA	73
Portugal	13

Netherlands	10
Germany	9
Australia	8
Canada	7
Ireland	7
India	7
France	6

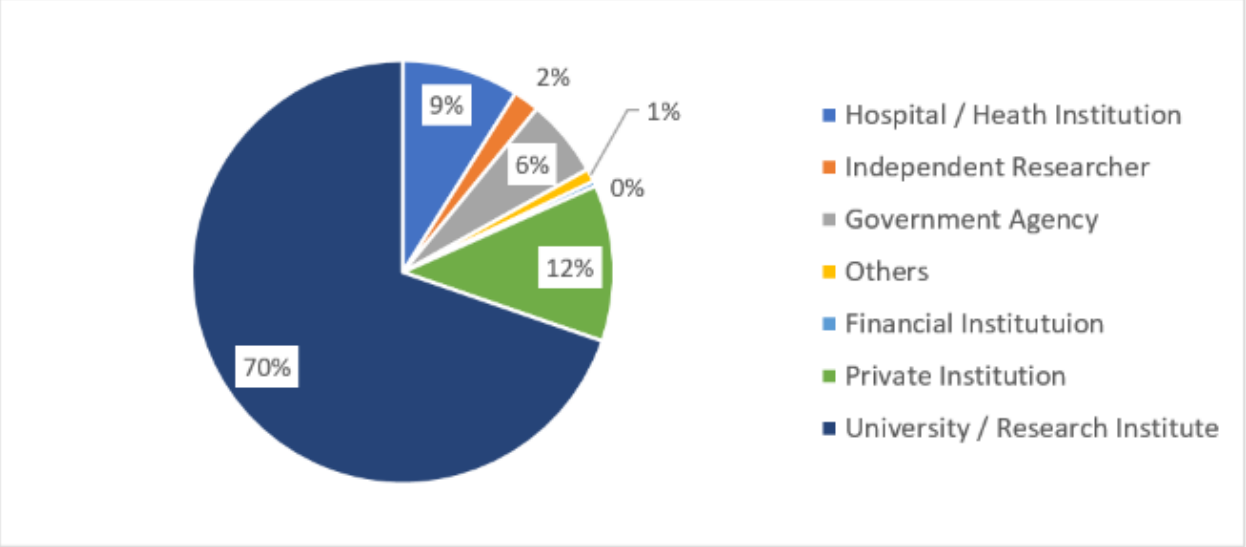

Figure 2 - Categories of entities that downloaded datasets from the repository.

4. Ongoing and future activities

This paper described the COVID-19 DataSharing/BR initiative, a Brazilian coalition that was created to publish open data on Brazilians that have been tested and/or hospitalized for COVID-19 within the health institutions that participate in the coalition. The main goal of this initiative is to make available in a timely manner large sets of curated patient clinical and health

records, and associated demographic information, to enable data-intensive pandemics-related research. To the best of our knowledge, this is the first initiative of its kind in Brazil.

As such, it is rapidly gaining new adhesions, and new Brazilian health institutions are already signing agreements to add their data to the repository. This, in turn, has triggered new design and implementation efforts, to ensure that all data will continue to be made available in a secure and timely manner. Additionally, new measures are being taken to facilitate overall data management, including definition of basic rules for naming and encoding criteria to be followed by all medical data providers.

Given the pressing need for quality COVID19 data, we had to reach a compromise between openly publishing the records in a timely manner, and user friendliness and interoperability with other initiatives, including international efforts. Thus, ongoing work concerns three main directions. The first refers to designing and implementing mechanisms for periodically updating the repository with additional records while maintaining overall integrity (e.g., involving issues such as data versioning and incremental integrity checks). The second involves the design of a complementary repository for medical images of COVID-19 Brazilian subjects. Besides the challenges posed by the present collection, there will be concerns about data transfer loads (given the expected data volume), designing additional mechanisms for individual privacy, and linkage to other records.Last but not least, we are concerned with user-friendliness on accessing data, based on feedback we are receiving with suggestions for more functionality. This also involves further systematization of its data and metadata dictionaries, to support aggregating data from other institutions.

Authorship attribution notes:

Following the authorship credit taxonomy of https://casrai.org/credit/, authors Mello, Medeiros, Prado, Nunes, Barnabe, Rizzatti, Ferreira, Reis, Rizzo, Maciel and Cesar directly contributed to writing and editing this text. Authors Numan, Sa, Sarno, Lamonica and Carvalho participated in the implementation and data selection and preprocessing at the respective institutions to enable the creation of the repository. The authors compose the group of people who coordinate and implement the initiative.

References

[1] National Covid Cohort Collaborative - N3C. 2020. accessed July 2020 <https://ncats.nih.gov/n3c>

[2] COVID-19 Future Analytics Clearing Board. 2020. accessed July 2020 <https://www.bundeskanzleramt.gv.at/themen/think-austria/covid-19-future-operations-clearing-board.html>

[3] Open Source Secure Data Infrastructure and Processes (OSSDIP) Supporting Fully Controlled Data Visiting for Sensitive Data. 2020. accessed July 2020 <http://www.ifs.tuwien.ac.at/~andi/secure_data_infrastructure.html>

[4] Vukosi Marivate et al. Coronavirus Disease (COVID-19) case data - South Africa. 2020. DOI: https://doi.org/10.5281/zenodo.3938381,

[5] IEEE. IEEE DataPort - COVID19. 2020. Accessed July 2020. <https://ieee-dataport.org/topic-tags/covid-19>

[6] FAPESP COVID-19 DataSharing/BR. FAPESP COVID-19 DataSharing/BR Initiative. 2020. Accessed July 2020. <https://repositoriodatasharingfapesp.uspdigital.usp.br/>

[7] FAPESP. Sao Paulo State Foundation. Accessed July 2020. <http://www.fapesp.br/en>

[8] Grupo Fleury. Fleury Group. Accessed July 2020. <http://www.grupofleury.com.br>

[9] Sociedade Beneficente Israelita Albert Einstein. Albert Einstein Israelite Hospital. Accessed July 2020. <http://www.einstein.br>

[10] The Joint Commission. Accessed July 2020 <http://jointcommission.org>

[11] Hospital Sirio Libanes. Sirio Libanes Hospital. Accessed July 2020. <http://hospitalsiriolibanes.org.br>

[12] RDA COVID-19 Working Group. Recommendations and Guidelines on data sharing. Research Data Alliance. 2020. DOI: https://doi.org/10.15497/rda00052

[13] Open Research Data Network of the State of Sao Paulo. 2019. Accessed July 2020 <https://metabuscador.uspdigital.usp.br/>

[14] Amit Sheth and James Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys, 22(3):183-236, 1990

[15] Open Archives Initiative. Open Archives Initiative Protocol for Metadata Harvesting. Accessed July 2020 <https://www.openarchives.org/pmh/>

[16] Danilo Carlotti, Joao Eduardo Ferreira, Fatima L. S. Nunes. Relational data model and programs to use and view data from the FAPESP COVID-19 Data Sharing/BR repository. 2020 URI: http://repositorio.uspdigital.usp.br/handle/item/243