6. ANALYSIS OF THE DATASETS AVAILABLE IN THE COVID-19 DATA SHARING/BR REPOSITORY IN CONFORMANCE TO THE FAIR PRINCIPLES

Anderson Rafael Castro Simões[125]
Renata Lemos dos Anjos[126]
Guilherme Ataíde Dias[127]

6.1 INTRODUCTION

The constant use of Digital Information and Communication Technologies (DICTs) by various individuals and sectors of society, including the academic-scientific community, contributes to a growing and continuous production of data. The academic-scientific scenario is configured both as a large producer and consumer of these, which are now considered primary sources for new scientific investigations, providing the development of science.

In this perspective, Sales *et al*. (2020) affirm that the advancement of science in various areas of knowledge is strongly linked to the reuse of scientific data, which points to a demand for managing and preserving them through digital curation activities for as long as necessary, to enable their effective reuse in future research.

These data, in addition to gaining value and importance in social, political, and economic scenarios, become crucial components in facing serious social and environmental challenges of the 21st century, through their sharing (Leonelli, 2019).

Meanwhile, on March 11, 2020, the World Health Organization (WHO) declared a COVID-19 pandemic caused by the SARS-CoV-2 coronavirus. As a result, issues regarding scientific data sharing and collaboration among different sources of investigation have gained prominence and have been evidenced. The pandemic presents a real and urgent need to gather efforts on a global scale, especially in the scientific field, so that the gaps about the new coronavirus are quickly and effectively resolved (Almeida *et al*., 2020).

Therefore, considering this reality, the São Paulo State Research Foundation (FAPESP), in cooperation with other institutions, announced the creation of the country's first open, anonymized Research Data Repository (RDR) re-

125 Master in Management of Learning Organizations from the Professional Master's Program in Management of Learning Organizations at the Federal University of Paraíba – MPGOA/UFPB. E-mail: anderson.simoes@estudantes.ufpb.br.

126 Master in Information Science from the Postgraduate Program in Information Science at the Federal University of Paraíba – PPGCI/UFPB. E-mail: renata.anjos@academico.ufpb.br.

127 PhD in Communication Sciences (Information Science) from the School of Communication and Arts of the University of São Paulo - ECA/USP. Professor of the Department of Information Science of the Federal University of Paraíba. E-mail: guilhermeataide@ccsa.ufpb.br.

lated to COVID-19, the **COVID-19 Data Sharing/BR**[128], which aims to make available COVID-19 related research data in Brazil, to contribute to this topic (FAPESP, c2016).

In this context, COVID-19 Data Sharing/BR materializes the conception that sharing, using, and reusing datasets available in a repository efficiently help to solve the challenges posed by the pandemic, which emphasizes the current role of scientific data.

Some initiatives were proposed to contribute with effective data reuse by other researchers and scientific investigations, among which FAIR principles stand out. Those principles stand out as being an approach that aims at making data easily ***F****indable*, ***A****ccessible*, ***I****nteroperable* and ***R****eusable*. Such principles encourage, among other practices, the use of metadata that, when properly used, can contribute to increase findability, access, interoperability, and reuse of different datasets (Dias; Anjos; Rodrigues, 2019).

Given the importance of collaboration among scientists during the pandemic and the potential of adhering to FAIR principles to increase the sharing of scientific datasets available in digital repositories and maximize their use and reuse, the following research question is posed: **How well do the datasets available in the COVID-19 Data Sharing/BR repository conform to the FAIR principles?**

To answer the research question, the following general objective was elaborated: evaluate the adherence of datasets available in COVID-19 Data Sharing/BR data repository according to FAIR principles.

6.2 EXPANDING THE ACCESS TO SCIENTIFIC DATA IN THE CONTEXT OF COVID-19 PANDEMIC

The COVID-19 Data Sharing/BR data repository was created with the aim of promoting sharing and collaboration, and was established through a partnership between FAPESP, the University of São Paulo (USP), the Fleury Institute, as well as Sírio-Libanês and Israelita Albert Einstein hospitals, all located in the state of São Paulo. The repository contains demographic data on 75 million patients, 1.6 million clinical and lab exams, and data on the outcomes of 6,500 patients, all of which can be used to support scientific research on COVID-19. The datasets available in the repository are classified into three categories: demographic data (gender, year of birth, and region), data on clinical and/or lab exams, and data on the patient's history (hospitalization, recovery, and death) (FAPESP, c2016).

Regarding this matter, the FAIR principles initiative was created to serve as guidelines for academic-scientific, industry, financing agencies, and publisher scenarios that aim to improve their data support infrastructure and promote the use and reuse of their data. Unlike other initiatives that focus on improving data usage for humans, the FAIR principles seek to improve machine capacity to automatically find and use data, thus enabling its reuse by users (Wilkinson *et al.*, 2016). The FAIR principles aim to provide distinct considerations for contemporary data

128 Available from: https://repositoriofapespcienciasaude.uspdigital.usp.br/. Access on: feb. 25, 2021.

release environments, including support for data input, exploration, sharing, and manual and automated reuse (Wilkinson *et al.*, 2016).

The first FAIR principle (***F****indable*) addresses the need to make data findable, which is an essential prerequisite for the effectiveness of the other three FAIR principles. A dataset must have a unique and persistent identifier, allowing its discovery at any time. Additionally, data must be described with rich metadata in a way that the researcher can find the desired data, even if they do not have access to its identifier (Dias; Anjos; Rodrigues, 2019).

The ***A****ccessible* principle focuses on the need to make data and metadata more accessible from the moment they are found. These entities must be accessible to users and/or machines at all times. For this purpose, it is important to use open, free, and universally implementable protocols (Go FAIR, [202-?]). It is recommended that metadata be available and accessible, even when the dataset does not allow free access to its content (Wilkinson *et al.*, 2016; Go FAIR, [202-?]).

The third FAIR principle (***I****nteroperable*) addresses the need to make data and other digital assets more interoperable. This issue is related to the need to integrate data with other datasets and with a wide range of applications throughout their lifecycle. To make interoperability among datasets possible, it is important to have instruments to semantically standardize the systems involved in this process, such as thesaurus and ontologies (Go FAIR, [202-?]).

The fourth and final FAIR principle (***R****eusable*) addresses the need to make data reusable. The implementation of data reuse requires a multifaceted approach and enables data to be reused by new communities of users for new needs and applications. In this regard, data can become more valuable to individuals in a wide range of organizations, including open-source communities and private organizations (Wise *et al*., 2019). It is recommended that policies for data and metadata access be explicit, thus ensuring understanding of access, use, and reuse rights, as well as details that indicate the origin of these objects (Go FAIR, [202-?]).

In this perspective, it is noteworthy the importance of implementing FAIR principles, which, when implemented, can result in several developments, including the possibility of process automation through the capacity of machine-automated data and metadata reading. This contributes to increasing their reuse and scalability, besides providing a more rigorous data and metadata management with potential to benefit the entire academic community. Thus, FAIR principles become a premise to support scientific findings and innovation (Wilkinson *et al*., 2016; Wise *et al*., 2019).

6.3 METHODOLOGICAL PATH

The research aims to evaluate the adherence of datasets available in the COVID-19 Data Sharing/BR data repository to the FAIR principles. From the objective standpoint, it is characterized as an exploratory and descriptive study. Regarding the problem approach, it is structured, with all stages of the investigative process previously determined. As for the type of investigation, it is a mixed analysis, using qualitative analysis to verify the datasets' adherence to the FAIR principles and quantitative analysis to evaluate the FAIRness adherence score to the FAIR principles (Richardson, 2017).

The *corpus* of this research was constituted by datasets available in COVID-19 Data Sharing/BR repository, in a total of three sets, each one of them coming from one of the collaborating institutions, namely: Fleury Group, Sírio-Libanês Hospital and Israelita Albert Einstein Hospital.

To verify the FAIRness score, we used the *online Self-Assessment Tool to Improve the FAIRness of Your Dataset,* SATIFYD[129], proposed by *Data Archiving and Networked Services* (DANS). This tool serves as a tool for dataset auto evaluation before its publication.

The SATIFYD tool consists of 12 questions that address the FAIR principles, divided equally into sections corresponding to the letters of the acronym FAIR. Specifically, the *Findable* section includes questions one to three; the *Accessible* section includes questions four to six; the *Interoperable* section includes questions seven to nine, and the *Reusable* section includes questions ten to twelve.

As a means of evaluation, the tool provides both a score by letter/principle and a visual representation of the corresponding letter. The more "blue" each letter is, the more adherent the dataset is to FAIR principles in that respective dimension. The tool also provides a general score – FAIRness – calculated from the average score associated with each principle.

At the beginning of the analysis, it was observed that the three datasets in COVID-19 Data Sharing/BR were available in the same manner and followed the same structure. As a result, these datasets received similar scores by the end of the analysis. Therefore, in the presentation and analysis of the results, it was decided to present and analyze the results obtained from only one dataset, as they were deemed similar. The analysis was conducted jointly and simultaneously by the authors, with the goal of providing a comprehensive assessment from different perspectives.

It should be noted that to conduct the analysis, it was necessary to download the datasets, and access the metadata sets available in the respective repository under the option "Completed Record". It is imported to mention that there are no records of changes or updates to the datasets in their respective completed records. Therefore, it is necessary to identify the dates associated with these alterations or updates through metadata related to the transferred files.

6.4 PRESENTATION AND ANALYSIS OF RESULTS

The first section of the tool, corresponding to the *Findable* principle, consists of three questions. The first question pertains to the availability of sufficient metadata, and the tool provides an informative text for each question with an icon **"i"** next to it. The tool presents a list of 13 items indicating the parameters to be met when sufficient metadata is sought. During the analysis, it was observed that the metadata of the analyzed datasets did not satisfy four items on the list: people who contributed to the datasets, target group of the datasets, license indicating data accessibility, and spatial coverage (geographic location in which the research was conducted). It is worth

129 https://satifyd.dans.knaw.nl/

noting that there is no information available on the individuals who contributed to the research that led to the creation of the datasets. Instead, only the collaborating institutions are mentioned as authors. Regarding the use license, the repository's initial page mentions that all datasets adopt the Creative Commons CC-BY open data license. However, users who access the datasets directly through their persistent identifiers may not have access to information on the adopted license.

The second question discusses the use of standards such as controlled vocabulary, taxonomies (thesaurus), or ontologies for describing sets. As analyzed, the datasets in question did not provide clues about the use of controlled vocabularies. It is noted that not using terminology control resources/tools causes a deficit, both at the moment of finding datasets and in ensuring that other researchers reuse them.

The third question pertains to providing additional documentation, such as a README file. Although it was not found with the same nomenclature, it is noteworthy that the repository clearly shows concern in developing a data dictionary for each of the datasets published. These data dictionaries are considered additional documentation that describes and explains the way data is structured, enabling any researcher and/or institution that can access it to understand and reuse it in their investigations.

Thus, regarding the *Findable* principle, the datasets obtained a FAIRness score of 38%.

The second section, which concerns the *Accessible* principle, also consists of three questions (questions four, five, and six). The fourth question addresses whether metadata is accessible to the public even when data is no longer available. As no information was found regarding this possibility of accessing metadata, even when data is no more available, it was decided to select the option "I can't find this information".

The fifth question addresses whether datasets contain personal data. According to the law Nº 13.709/2018, the Data Protection Law – DPL, personal data refers to natural individuals who are identified or identifiable. Since the data used to create the datasets were anonymized and conformed to the DPL's anonymized data classification, which pertains to data that cannot be traced back to an identifiable individual (Brasil, 2018), this question was answered negatively.

The sixth question discusses which use licenses were chosen to ensure access rights. As previously stated, the repository specifies the use license assigned to all its datasets on its initial page; however, the license is not informed in the sets of metadata. Among the available answer options, we have: open access to all, open access to registered users, restricted access through approval request, restricted access to specific groups and other types of access. It was decided to choose the option of open access to registered users, as users are required to complete a brief registration (name, e-mail, and institution) and agree to the statement of responsibility regarding the ethical use of data and the obligation to give proper credit to the datasets through citations to download the datasets.

It was observed that the complete registrations of datasets do not reveal metadata that facilitates the contact between users and data holders. It should be noted that this contact may be necessary to clarify future and potential questions.

As a result, regarding the *Accessible* principle, the datasets received a FAIRness score of 55%.

The third section, which corresponds to the *Interoperable* principle, also comprises three questions (questions seven, eight, and nine). The seventh question discusses whether the datasets are stored in preferred formats. The tool provides informative texts about these preferred formats, in which, for spreadsheets, the tool states that the preferred formats are ODS and CSV. The analyzed datasets can be downloaded in CSV format, so all the data is in the group of formats indicated as preferred.

The eighth and ninth questions discuss the linkage to other (meta)data and whether they are accessible online and whether there is the provision of contextual information (reference to other sets or publications) about datasets. It was noted that the datasets in question do not contain contextual information or links to other (meta) data, which goes against the recommendation of the Interoperable principle.

As a result, regarding the *Interoperable* principle, the datasets obtained a FAIRness score of 50%.

The fourth section, corresponding to the *Reusable* principle, is composed of three questions (questions ten, eleven, and twelve). The tenth question discusses if there is information about where data comes from, such as data origin, citation for reused data, description of the workflow, history of data processing, and version. In this question, only the option of data origin was selected since it is informed in its description. Concerning citations for reused data, description of workflow, and history of data processing, and version, no information was found.

It is valid to point out the relevance of these other information for datasets. The description of workflow allows other researchers to have a deeper understanding of how data was created or reused through citations of these datasets.

The eleventh question discusses once again the access and use license adopted by datasets, as previously discussed in the sixth question, with the same options of answer, in which open access was selected.

The twelfth question discusses if the (meta)data meets the domain standards concerning standardized data organization. The option of using domain standards was selected, considering the structured and standardized way in which data was organized.

Thus, regarding the *Reusable* principle, the datasets reached the FAIRness score of 74%.

For a better understanding and visualization of the questions and how they were answered, Table 1 was developed with this respective description.

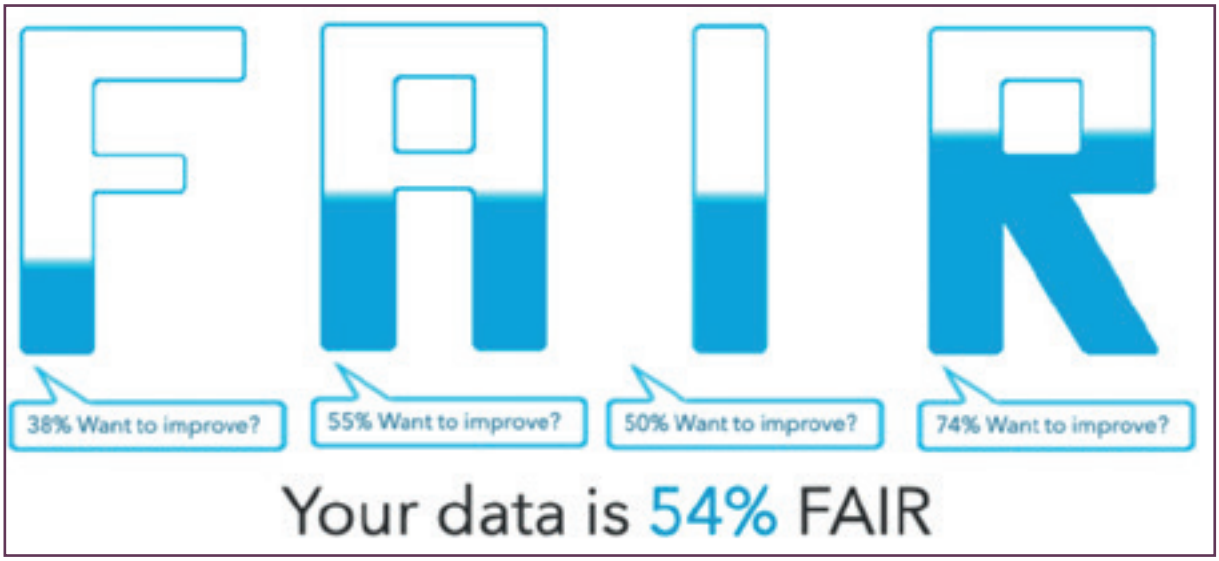
Table 1 – SATIFYD questions and respective options of answers selected.

PRINCIPLE	QUESTIONS	OPTION OF ANSWER SELECTED
FINDABLE Section 1	Did you provide metadata (information) enough about your data so that other people find, understand, and reuse it?	Mandatory metadata fields and some additional fields.
	Did you use standards such as controlled vocabularies, taxonomies (thesaurus) or ontologies to describe your dataset?	No standards were used.
	Did you provide rich and detailed additional documentation?	README file.
ACCESSIBLE Section 2	Is metadata accessible to the public even if data is not available anymore?	Yes.
	Does your dataset have personal data?	No.
	Which use license did you choose to fulfill the access rights attached to data?	Open access (registered user).
INTEROPERABLE Section 3	Is data in your dataset stored in preferential formats?	All data is in preferential formats.
	Do you link to other (meta)data? Can this (meta)data be accessed *on-line*?	No.
	Did you provide contextual information about your dataset?	With no contextual metadata.
REUSABLE Section 4	What kind of information did you provide about your data origin?	Data origin.
	Which use license did you use for your dataset?	Open access (user registered).
	Does your (meta)data meet the domain standards?	Domain standards in metadata.

Source: Data Archiving And Networked Services (c2024); Research data (2020).

In the end, the tool provided a final FAIRness score of 54%, which is calculated as the average score of each principle. Figure 1 illustrates how the tool presents the analysis result.

Figure 1 – Result of analyses in SATIFYD.



Source: Research data (2020).

6.5 FINAL CONSIDERATIONS

It was noted that, despite the datasets adopting some practices proposed by the FAIR Principles, the COVID-19 Data Sharing/BR does not mention or recommend the adoption of the principles before the act of publishing the data by its holders (Santos; Sant'ana, 2019).

As observed, the repository informs, only on its initial page, that all datasets published there adopt Creative Commons CC-BY open data license and that all and any publication or presentation that uses data in the repository must cite it. It is suggested that all datasets inform, in their metadata, the access and use license adopted so that there are no misunderstandings, given the possibility of users directly accessing datasets through their persistent identifiers.

Another point observed was that, repeatedly, only on its initial page, the repository informs that the datasets are periodically updated, so it must be frequently verified for downloading new data. On the other hand, in metadata of datasets published, there is no information about the version history, so it is up to the user to verify, after downloading, its creation date and confirm is there was an update. It is suggested that the version history is informed in metadata.

It is understood that the more evaluation tools available for use, the more opportunities to improve and rethink the ways of evaluating datasets according to FAIR principles; the need to mature the evaluation process is similar to its improvement. A relevant fact for the repositories as well is that the more evaluations are made, more proposals for improvements will be suggested.

During the research, only tools for self-assessment of datasets were found. That is, the researcher himself performs the self-assessment of his datasets before being published in the repositories. Thus, it is important to create tools that enable the analysis of datasets already published to investigate how the adoption of FAIR principles is being taken by the academic-scientific community.

An initiative such as COVID-19 Data Sharing/BR, which in this case was developed in face of a pandemic context, is very welcome; the rationale for the idea and the importance of collaborations from institutions such as FAPESP, University of São Paulo (USP), and the participation of Fleury Institute, besides Sírio-Libanês and Israelita Albert Einstein hospitals, show the importance of making data available in issues regarding public health. However, it is recommended that more efforts are put on increasing the adherence of FAIR principles in datasets stored in this repository.

REFERENCES

ALMEIDA, B. de A. *et al*. Preservação da privacidade no enfrentamento da COVID-19: dados pessoais e a pandemia global. **Ciência & Saúde Coletiva**, [*s. l.*], v. 25, p. 2487-2492, 2020. Available from: https://doi.org/10.1590/1413-81232020256.1.11792020. Access on: 15 aug. 2020.

BRASIL. **Lei nº 13.709 de 14 de agosto de 2018**. Lei Geral de Proteção de Dados (LGPD). Brasília: Presidência da República, 2018. Available from: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm. Access on: 10 oct. 2020.

DATA ARCHIVING AND NETWORKED SERVICES. **Self-Assessment Tool to Improve de FAIRness of Your Dataset (SATIFYD)**. c2024. Available from: https://dans.knaw.nl/en/satifyd/. Access on: 07 oct. 2024.

DIAS, G. A.; ANJOS, R. L.; RODRIGUES, A. A. Os princípios FAIR: viabilizando o reuso de dados científicos. *In*: DIAS, A. D.; OLIVEIRA, B. M. J. F (org.). **Dados Científicos**: perspectivas e desafios. João Pessoa: UFPB, 2019. p. 177-187.

FAPESP. **FAPESP COVID-19 Data Sharing/BR**. c2016. Available from: https://repositoriodatasharingfapesp.uspdigital.usp.br/. Access on: 10 aug. 2020.

GO FAIR. **FAIR Principles**. [202-?]. Available from: https://www.go-fair.org/fair-principles/. Access on: 10 aug.. 2020.

SALES, L. *et al*. GO FAIR Brazil: a challenge for brazilian data science. **Data Intelligence**, [*s. l.*], v. 2, n. 1-2, p. 238-245, 2020. Available from: https://doi.org/10.1162/dint_a_00046. Access on: 20 aug. 2020.

LEONELLI, S. Data-from objects to assets. **Nature**, [*s. l.*], v. 574, p. 317 - 320, 2019. Available from: http://dx.doi.org/10.1038/d41586-019-03062-w. Access on: 20 aug. 2020.

RICHARDSON, R. J. **Pesquisa Social:** Métodos e Técnicas. 4 ed. São Paulo: Atlas, 2017. 424 p.

SANTOS, P. L. V. A. C.; SANT'ANA, R. C. G. Camadas de Representação de Dados e suas Especificidades no Cenário Científico. *In*: DIAS, A. D.; OLIVEIRA, B. M. J. F (org.). **Dados Científicos**: perspectivas e desafios. João Pessoa: UFPB, 2019. p. 53-66.

WILKINSON, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, [*s. l.*], v. 3, n. 1, p. 1-9, 2016. Available from: https://doi.org/10.1038/sdata.2016.18. Access on: 15 aug. 2020.

WISE, J. *et al*. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. **Drug Discovery Today**, [*s. l.*], v. 24, n. 4, 2019, p. 933-938. Available from: https://doi.org/10.1016/j.drudis.2019.01.008. Access on: 15 aug. 2020.

How to cite this chapter: SIMÕES, Anderson Rafael Castro; ANJOS, Renata Lemos dos; DIAS, Guilherme Ataíde. Analysis of the datasets available in the COVID-19 data sharing/BR repositor in conformance to the FAIR principles. *In*: SALES, Luana Faria; VEIGA, Viviane Santos de Oliveira; VIDOTTI, Silvana Aparecida Borsetti Gregório; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **FAIR Principles Applied To Research Data Management**: Brazilian Experiences. Brasília, DF: Editora Ibict, 2024. cap. 6, p. 80-89. DOI: 10.22477/9788570131959.cap6.