17. VODAN BR - a platform for supporting COVID-19 data following FAIR principles

Maria Luiza Machado Campos[185]
Vania Borges [186]
Giseli Rabello Lopes [187]
Maria Claudia Cavalcanti[188]
João Moreira[189]
Sergio Manuel Serra da Cruz[190]

17.1 INTRODUCTION

The COVID-19 pandemic has made clear the importance of having the results of scientific research more easily available for immediate and wide reuse. Several groups already involved in these themes were mobilized, seeking to discuss and speed up the definition and building of supporting infrastructures capable of facing this challenge. In particular, participants from the Research Data Alliance (RDA)[191], World Data Systems (WDS)[192], GO FAIR network[193] and the Committee on Data (CODATA)[194], linked to the International Science Council (ISC)[195], launched a call for action called Data Together (Data Together, [2020?]), which accelerated and promoted cooperation between different ongoing initiatives.

The implementation network GO FAIR *Virus Outbreak Data Network* (VODAN) was designed to initiate a "community of communities" to quickly design and build an infrastructure for a network of interoperable and shareable international data, and to offer support in the search for evidence-based responses to viral outbreak cases (Go Fair, 2020). As it is an initiative of the GO FAIR consortium, data, and services generated must meet FAIR principles (Mons, 2020), which provide guidelines to make data findable, accessible, interoperable and reusable. In the case of VODAN network, the starting point is clinical data of patients with COVID-19, carrying out, in the first phase,

185 PhD, Graduation Program in IT– PPGI/UFRJ, mluiza.campos@gmail.com

186 Doctoral Student, Graduation Program in IT – PPGI/UFRJ, vjborges30@gmail.com

187 DSc, Graduation Program in IT – PPGI/UFRJ, giseli@dcc.ufrj.br

188 DSc,Military Institute of Engineering– IME, yoko@ime.eb.br

189 PhD, University of Twente - UTwente, j.luizrebelomoreira@utwente.nl

190 DSc, Graduation Program in IT – PPGI/UFRJ, serra@ppgi.ufrj.br

191 Available on: https://www.rd-alliance.org/. Access on: 20 Sept. 2024.

192 Available on: https://www.worlddatasystem.org/. Access on: 20 Sept. 2024.

193 GO FAIR – initiative that encourages the availability of FAIR data and services (Findable, Accessible, Interoperable and Reusable) for scientific research projects. Available on: https://www.go-fair.org. Access on: 20 Sept. 2024.

194 Available on: https://codata.org. Access on: 20 Sept. 2024.

195 Available on: https://council.science/. Access on: 20 Sept. 2024.

the transformation and treatment of these data, according to the Clinic Research Form (CRF), developed and standardized by the World Health Organization (WHO).

The CRF-WHO is a clinical research protocol developed with the help of specialists to obtain relevant information in epidemic and pandemic cases. This form is dived in three modules: the first aims at collecting patient admission data; the second aims at follow-up data during hospitalization; and the third aims at treatment outcome data, whether by discharge, hospital transfer or death.

According to Manifesto Vodan (2020), the original proposal consists in developing a solution that allows health professionals to record data observed in the format established by CRF-WHO, storing them in repositories or data banks. Later, metadata of these repositories must be available in a FAIR Data Point (FAIR DP). A FAIR DP is a component of the FAIR data support infrastructure through which software agents can have access to descriptors that allow them to find and visit data locally and execute queries on them (Mons, 2020). Local data curator will give the permission or not for the query/analysis to be performed. This structure allows that the patient's information to remain protected in databases of health facilities, respecting the legislation for health data in each country.

In Brazil, VODAN BR project[196] began concurrently with the advance of the pandemic in the country, during the first months of 2020, as part of GO FAIR Brazil Health[197], linked to Oswaldo Cruz Foundation (Fiocruz), in multi-institutional partnership with the Federal University of Rio de Janeiro (UFRJ) and the Federal University of the State of Rio de Janeiro (UniRio), among other institutions. The development of the infrastructure is under the responsibility of GRECO Research Group[198], from UFRJ, and its pilot test partners are Gaffreé Guinle Federal Hospital in Rio de Janeiro, and São José Municipal Hospital, in Duque de Caxias. Data are collected from their original systems and treated to be in line with the standard established, that is, the WHO questionnaire format, aiming at their later availability as well as their metadata, meeting FAIR principles, Semantic Web standards and following licensing and anonymization criteria established.

This chapter aims to present an overview of computational assets being developed to support VODAN BR project. This scalable, distributed and generic infrastructure aims to meet an intensive data collection with high heterogeneity, making them available in platforms that offer data and metadata interoperable and processable by software agents, supporting the discovery of other resources that can be associated with them. Thus, it is possible to obtain greater agility in the discovery and generation of knowledge from a more effective reuse of research results.

The next sections are structured as follows: section 2 addresses the VODAN implementation network and the technologies it supports; section 3 presents VODAN BR platform, describing the process and infrastructure being developed; and section 4 presents the conclusion and future possibilities identified.

196 Available on: https://vodanbr.github.io/. Access on: 20 Sept. 2024.

197 Available on: https://portal.fiocruz.br/go-fair-brasil-saude. Access on: 20 Sept. 2024.

198 Available on: http://dgp.cnpq.br/dgp/espelhogrupo/634046. Access on: 20 Sept. 2024.

17.2 VODAN IMPLEMENTATION NETWORK AND ASSOCIATED TECHNOLOGIES

Although the volume of information available on the Web since the beginning of the pandemic has grown far above expectations, it is observed that, for the most part, they refer to total people infected, hospitalized, recovered and deaths. In addition to these aggregated data, data referring to clinical picture, the treatment of patients and their outcome constitute an important support for more detailed studies in clinical research. However, in general, it is observed that despite being extremely valuable to the scientific community, these data are not generally accessible.

Two main problems immediately emerge from dealing with data at this level of detail. The first is the confidentiality of medical records, which can be circumvented by providing anonymized and structured data to meet the demands of clinical investigations or specifically defined licensing. Another problem, more technical and more difficult to solve, resides in the use, by a large part of the Hospital Units (HU), of software for electronic medical records without greater structuring for data entry, with many free text fields, which make analysis and later extraction difficult.

Added to these problems are the challenges of developing and implementing an infrastructure that supports FAIR data release. Although the proposal of the principles is already some Years old, providing a conceptual basis and guidelines that quickly became popular, there are still few technological alternatives that have already been tried together. Certainly, the use of Semantic Web approaches and standards constitutes a solid contribution to the solutions being prospected and developed, but complementary mechanisms and technologies are still necessary. The next two subsections describe VODAN network in more detail, as well as some main technological resources and solutions that support it.

17.2.1 VODAN Implementation Network

VODAN implementation network emerged in early 2020, as a joint effort to implement, experiment and expedite solutions (some already being independently tried in other domains), to support FAIR data release and exploration in the context of research associated with COVID-19 and to other future viral outbreaks. The network proposes an effort for the so-called FAIRfication[199] of COVID-19 data, even after the fact, employing the CRF-WHO model to establish the standardization of information (Satti *et al.*, 2020). The Data FAIRfication process promotes the application of FAIR principles to data and metadata, as well as to the infrastructure that supports them. For that purpose, in general, the process contemplates two stages of: (i) collection of non-FAIR data; (ii) analysis of data collected; (iii) definition of a semantic model for the dataset that allows describing the meaning of entities and their relations; with accuracy and with no ambiguity; (iv) definition of metadata associated to data collected, including, among others, provenance, distribution and location, types of access; (v) treatment to make data potentially interlinkable with other sources, with assignment of persistent identifiers, annotation based on controlled vocabularies and/or ontologies, employing technologies and standards of Semantic Web and Linked Data (Heath; Bizer, 2011); (vi) definition of metadata associated to data (and their treatment so they are also FAIR); data and their metadata release.

199 FAIRification of data – process for turning non-FAIR data in FAIR data. Available on: https://www.go-fair.org/fair-principles/fairification-process/. Access on: 20 Sept. 2024.

After the FAIRfication process, a set of data and metadata adhering to the FAIR principles is obtained. This well-structured data and metadata can be explored through mechanisms that use machine learning techniques and other artificial intelligence (AI) approaches to discover significant patterns in epidemic outbreaks, supporting decisions and actions to face them. As presented in (Satti *et al*., 2020), it is vital to ensure that the data, metadata, and vocabularies used are FAIR, in the original sense of the acronym, but also in the sense of "*Federated, AI- Ready*", that is, federated data for AI.

At the end of the developments associated with the VODAN initiative, the establishment of a federated network of epidemiological FAIR DPs is expected, promoting FAIR services and data, accessible to researchers, for studies on the COVID-19 pandemic and other epidemics that may arise in the future.

VODAN Africa&Asia network[200] was the first initiative of implementation of VODAN network. It is funded by the Philips Foundation[201] and aims at promoting the access distributed to CRF data from Africa and Asia, to support the fight against the COVID-19 pandemic, assisting Universities and Hospitals in Uganda, Ethiopia, Nigeria, Kenya, Tunisia and Zimbabwe, among other countries. This initiative directed its activities towards training researchers and data designers in the creation of FAIR DPs. The trainings guided the participants on the FAIR principles and on the process of building the FAIR DPs, ensuring that data and metadata released are linked and can be available and processed by software agents. As a result, on July 22, 2020[202], the world's first FAIR DP was made available in Uganda. Since then, other FAIR DPs have been activated, based on data and metadata from partners HUs.

Differently from VODAN Africa&Asia network, VODAN BR project opted for a smaller scope, aiming to develop a supporting environment initially focused on data from two partner hospitals, adjusting them to CRF -WHO and creating a computational infrastructure for its dissemination through a first FAIR DP in Brazil. Subsequently, other hospitals will be contemplated, which can already make use of the results of the pilot experience conducted, and the infrastructure developed and tested.

17.2.2 Semantic Web and FAIR Principles

The Semantic Web proposes that data on the Web is defined and connected in a way to be interpreted by both human beings and machines, promoting their sharing and reusing by applications, companies, and community. For that purpose, the proposal of representation of connected data establishes a set of standards and best practices for releasing and interconnecting data structured on the Web, based on data annotation in controlled vocabularies and ontologies, making the identification of new connections among items from different sources easier, aiming to form a global data space, the so-called Data Web (Heath; Bizer, 2011).

200 Available on: https://www.vodan-totafrica.info/. Access on: 20 Sept. 2024.

201 Available on: http://www.digitaljournal.com/pr/4626217. Access on: 20 Sept. 2024.

202 Available on: https://kiu.ac.ug/special-news-page.php?i=covid-19-computer-readable-observational-data-installed-at--kampala-international-university_1595432235. Access on: 20 Sept. 2024.

FAIR principles, initially aimed at managing research data, add to what is already proposed for the Semantic Web, with the objective of making digital objects findable, accessible, interoperable and reusable. In essence, these principles add to the standards established by W3C[203] the importance of using metadata to facilitate the discovery and understanding of data, especially by machines (software agents). It should be noted that FAIR principles do not establish standards or supporting technologies, but rather guide the creation of FAIR data and metadata.

Metadata standards and content annotations are established to promote the common understanding of data meaning, guaranteeing the right interpretation and its adequate use. In order for this metadata to be machine-interpretable, they need to be findable and structured. Machine-actionable metadata, essential to FAIR principles, has led members of GO FAIR and RDA, start in 2018, to foster discussion on *Metadata for Machine* (M4M), in a series of events[204] to assess the state of the art and encourage the creation and reuse of metadata components and metadata templates for machine processing. In VODAN implementation, M4M has been involved in the standardization of metadata referring to catalogs and datasets that will be made available via FAIR DPs, as well as in a series of services associated with them.

In any case, it is not trivial to unequivocally explain a shared semantics about these digital assets, and ontologies play a fundamental role in this. Currently, ontologies are considered in areas of computing (Studer; Benjamins; Fensel, 1998), two of which are: (i) in the area of conceptual modeling, where, through the process of ontological analysis, models well-grounded on top-level ontologies are built; and (ii) in the area of Web Semantic, where both lightweight ontologies, in the line of vocabularies, taxonomies and thesauri, as well as robust ontologies are used, preferably following well-founded models and represented in expressive languages, which can be explored by inference mechanisms, to generate more knowledge.

Considering the definition of ontology as *"... a formal and explicit specification of a shared conceptualization"* (Santos, [2020?]), it follows that : an ontology is considered formal because it is machine interpretable; it is explicit because it presents specifications of concepts, properties, relations, functions, restrictions, and axioms very well-defined; it is a conceptualization for defining and abstract model and a vision of a phenomenon of the world that one wants to represent; and it is shared because it is consensual knowledge among those who work with the domain or applications in question.

The approach to ensure formalism and flexibility for the creation and availability of data and metadata uses RDF (Resource Description Framework) language[205], including RDFS (RDF Schema) in this context [206]. The OWL (Web Ontology Language) language[207], developed for the creation of robust ontologies, also uses this pattern.

203 Available on: W3C Semantic Web Activity https://www.w3.org/2013/data/. Access on: 20 Sept. 2024.

204 Available on: https://www.go-fair.org/resources/go-fair-workshop-series/metadata-for-machines-workshops/. Access on: 20 Sept. 2024.

205 Available on: https://www.w3.org/wiki/RDF. Access on: 20 Sept. 2024.

206 Available on: https://www.w3.org/TR/rdf-schema/. Access on: 20 Sept. 2024.

207 Available on: https://www.w3.org/OWL/. Access on: 20 Sept. 2024.

The formalism of the RDF specification is associated with the structural pattern used to describe and store data. This pattern is defined by triples consisting of the following elements: <*subject*> <*predicate*> <*object*>. Each triple constitutes a declaration, the basic unit of RDF, it is a set of declarations that describe a web resource. Each resource, in turn, has a unique identifier called Universal Resource Identifier (URI). The Uniform Resource Locators(URLs) associated to URIs are dereferenced, that is, they can be accessed through browsers, providing information about the resource. This unique identifier allows the reuse of resources between different data sources, streamlining implementations, providing interoperability and facilitating integrations

By describing data and its metadata, RDF allows flexibility in the construction and evolution of schemas not available in the usually used Database Management System (DBMS), such as those based on relational technologies. The set of statements represented by RDF constitute an RDF Knowledge Graph.

17.3 VODAN BR PROJECT AND THE PERSPECTIVE OF FAIR DATA AND METADATA MANAGEMENT

The VODAN BR project established a set of premises to be respected during its implementation phases. These premises guide the activities related to data and metadata management, aiming to establish a structure capable of being quickly adjusted, which significantly reduces the need for changes in applications/tools with each evolution and version of CRF or of the terminological instruments of reference. Among the established premises, it is worth mentioning:

- create an infrastructure capable of implementing and making available a digital CRF (application), centered on users of health services, which is capable of responding to epidemic episodes of this pandemic;

- store information established in the CRF-WHO, anonymously, considering possible versions of the current CRF for inclusion, alteration, or exclusion of form elements;

- enable the creation of CRFs or the inclusion of specific additional questions. This need was presented considering the different types of survey forms used in Brazil, which, in addition to the elements established by the CRF-WHO, are concerned with specific information, relevant to research in the country, such as, for example, participation in campaigns of vaccination and date of last dose;

- promote a conceptual modeling that allows the alignment of the elements of the forms to the ontologies (semantic models), helping the process of data FAIRfication;

- provide a flexible infrastructure, modular, scalable and agile infrastructure, to support software and database adaptations;

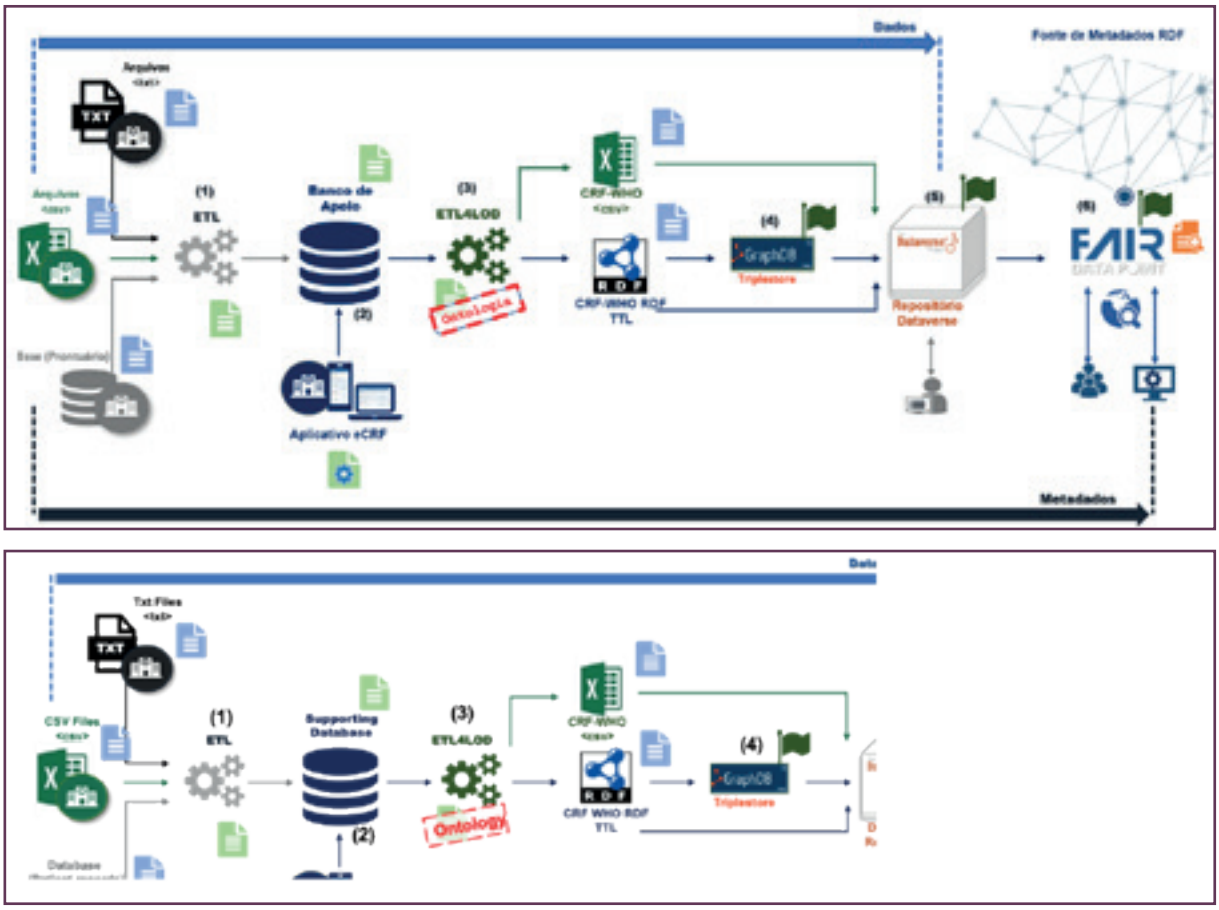
- transform the collected data, that is, "non-FAIR data" into linked data, mapping them to machine-readable formats, using RDF, making them available in datasets and releasing their metadata, also in RDF, in a FAIR DP;

- publicly make available a FAIR DP set to meet the conditions agreed with the participants, enabling access to data through controlled queries and not through traditional downloads.

Respecting these premises, a platform was designed for data processing and services that range from the availability of clinical research data by HUs to metadata release FAIR DP. The platform, represented in Figure 1, has as main requirements to be modular, distributed, scalable, and flexible. Modular, because the planned activities are organized in the form of modules that interact in a chained way, with the result of a module the input of the subsequent module. Scalable and distributed because the idea is that a supporting database is made available in each HU, as well as triple store repositories and/or databases will host data structured according to CRF-WHO in its different distributions or formats. In this way, as more hospital participate of the Project, more computational infrastructure will be added, causing a natural horizontal scaling. In addition, it is a flexible platform, as the heterogeneous data produced by the HUs are treated and transformed into a RDF graph representation, which is one of the formats that facilitates data interconnection.

Initially, as shown in Figure 1, (1) the platform captures data that can be in different formats, such as txt, cvs, or even in the format used in each HU, and via an Extraction-Transformation-Loading (ETL), performs the debugging and transformation of data, storing them in a supporting database (2) which can also directly receive data through a mobile application (eCRF) specially developed. The data stored in the supporting database then undergoes a transformation to connected data, (3), being annotated in vocabularies and ontologies, to meet the interoperability principle. They are then loaded into a graph database (4), in the role of a triple store, or, in the form of an RDF dataset, made available for download in a repository (5). The associated metadata also undergoes a processing and transformation process (3) being loaded and made available in a FAIR DP (6).

Figure 1 – Representation of VODAN BR Platform



Source: Designed by authors.

As established in VODAN network, the datasets must be "visited" by algorithms, respecting the access established by HUs. The metadata associated, contemplating, for example, information on the origin of existing data, types of distribution and the access policies, will be available and accessible in FAIR DP.

Of the elements that make up the platform, the following can be distinguished, due to their relevance in the project and the attention and challenges in the treatment of data: (i) the mechanisms for capturing data, contemplating different requirements and systems of the HUs; (ii) the supporting database, responsible for storing data from these heterogeneous data sources; (iii) the tool to support treatment, transformation, and annotation for inter-connected data and metadata; (iv) alternatives for releasing data; and (v) the creation and feeding of FAIR DP, part of the international VODAN general access point federation.

The tasks performed and the technological choices for these 5 elements of the VODAN BR platform are described in the following subsections.

17.3.1 Data collection

The project envisages three different ways of collecting data from clinical trials of patients with COVID-19:

- by using the application (eCRF) created for recording information, according to the CRF – WHO;
- through an ETL tool for anonymized data uploads from files in txt or csv formats made available by HUs;
- through ETL processes connecting data bank to data bank, with the purpose of transferring information from the patient records to supporting data banks, in the format established by CRF-WHO.

The collection from existing digital records represents an additional challenge. Despite the use of the supporting database by HUs, as a transition bank for CRF-WHO format, and all the facilities it offers, one of the main problems in the analysis and extraction of clinical data for research stems from the flexibility of existing medical records systems that enable textual fields for recording certain aspects of the treatment. The lack of standardization in this record and the large volume of these unstructured data (which includes each procedure performed on the patient, including medications and lab tests), make the collection and transformation process difficult, requiring the support of a health professional for its interpretation and recording. This problem is not new and has been a constant in studies on the interoperability of health treatment data (Santos, [2020?]; Cruz; Campos; Mattoso, 2009). Another important aspect considered was the diversity of information on data provenance (Cruz; Campos; Mattoso, 2009) to be managed.

17.3.2 Creation and Maintenance of Supporting Data banks

Due to the heterogeneity of the data sources and the data per se, it was decided to develop a supporting base for the treatment and formatting of data, aiming to adapt them to the CRF-WHO structure and to support and accelerate the transformation process for connected data.

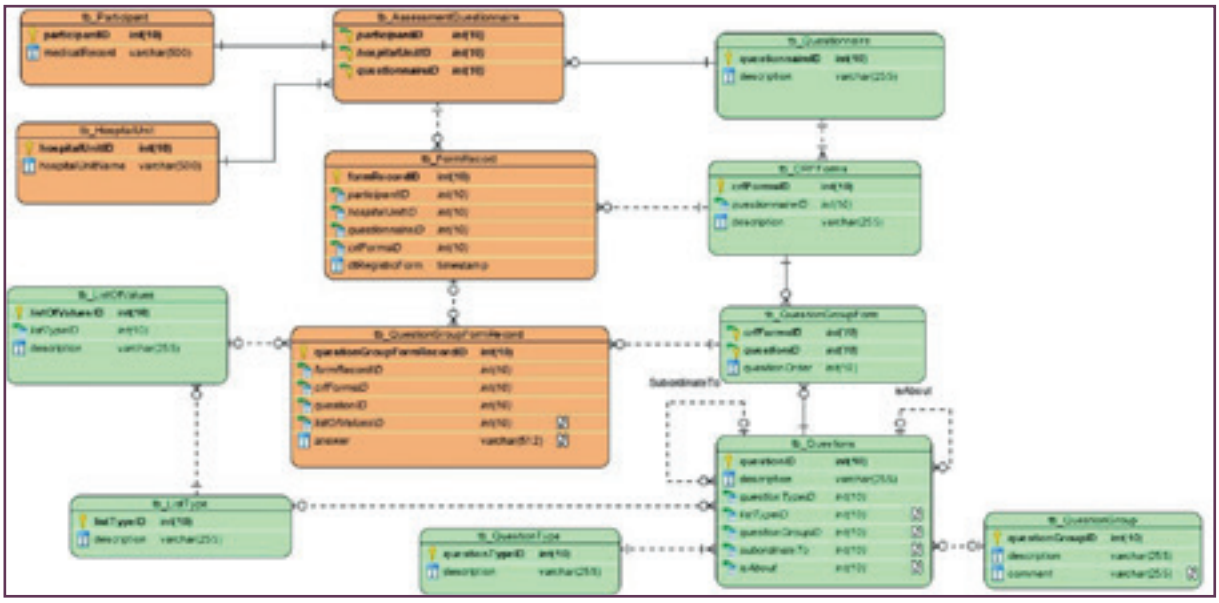
We emphasize that, although part of the data sources come from the medical records of patients with similar systems to a certain extent, it was chosen to follow a model that adheres to the questionnaire for data collection from the CRF-WHO. This decision was critical, as it allowed the establishment of a structure of questions and answers associated with the forms/modules oriented to the service and, consequently, to the collection of information. The form represents a survey, in our case a clinical data survey. It consists of a set of questions, grouped into well-defined categories, collected by a health agent, in this case at a HU, considering observations made about an element of interest, the patient. This research requires a spatio-temporal view, having for that purpose: an admission module, destined to the questions currently the patient arrives; a follow-up module, intended for questions about the patient during hospitalization; and an outcome module, with an overview of the treatment provided and the patient's final situation.

The use of questions with mostly standardized answers helps the discovery of vocabularies and the adoption of Semantic Web, such as ontologies that can be used to define a semantic model.

Another important aspect refers to the combination of the hierarchical structure Module/Group/Question and Subordinate Question that embeds an organization of knowledge by categories, allowing to define different views for analysis. An example of a possible analysis would be the evaluation of cases (from admission to outcome), considering the comorbidities identified at the moment of admission and the medications administered during hospitalization. The result of this analysis could help in the process of guiding drugs indicated or not in a treatment, given the patient's comorbidity.

For the modeling and implementation of this supporting database, it was decided to use a modeling based on relational technology, due to the ease of maintenance and interaction with the mechanisms and applications for data uploading and manipulation. A partial view of this database schema is presented in Figure 2, where the entities in green represent the hierarchical structure of the CRF-WHO and those in orange represent the record of patient information collected by hospital units.

Figure 2 – Extract of the Logical Model of the supporting data bank– CRF-WHO View



Source: Modelled by authors.

In this model, information on CRF-WHO and its versions are registered in the *tb Questionnaire* table; the three modules associated to CRF are registered in the *tb_CRFForms* table; the questions are stored in the *tb_Questions* table, where the type of question is associated (*tb_QuestionType*), the group to which it belongs (*tb_QuestionGroup*), if any, and the type of list of standardized answers (*tb_ListType*), in case the question demands a standardized answer.

The information of the research participants (patients) is entered in the *tb_Participant* table and the information of the Hospital Unit in the *tb_HospitalUnit* table. The table *tb_AssessmentQuestionnaire* records the opening of a CRF-WHO for a participant. For each record in this table, the modules enabled for the patient are launched in the *tb_FormRecord* table. The table *tb_QuestionGroupFormRecord* stores, per module, the questions and answers obtained from the evaluation of each patient, considering textual and standardized answers (*tb_ListOfValues*).

17.3.3 Transformation for Connected Data

After the treatment and the upload of data from each source (HU) in its supporting data bank, the next step refers to a process of transformation for a semantic model, following the paradigm of connected data.

For the VODAN project, the FAIR Data Team made available a model of semantic data that represents the fast version of the CRF-WHO. This semantic model (or ontology) was denominated WHO-COVID-CRF[208] and, in addition to the representation of CRF-WHO, it established a set of entities in the health field domain, to which the questions in the form are related. These entities refer to other existing and well-documented ontologies, providing quality and additional information to guide the users in filling out the form.

As it was developed oriented to the CFR-WHO form, the analysis of this semantic model identified a series of similarities that made it possible to extend the modeling of the supporting database, including ontology information referring to identification, structuring and valuation, to speed up the data FAIRfication process.

The structuring of WHO-COVID-CRF ontology allowed the use of its information for the initial upload of the tables that represent the questionnaire, with very few adjustments. Through this upload, the alignment of the information in the tables referring to the questionnaire with the ontology was implemented, allowing the creation, by the data administrator, of views that present the questionnaire and its ontological information, as well as views that help the stage transformation to linked data performed later.

In order to make data connected, the tool ETL4LOD[209] was used. This tool was initially developed through a partnership between URFJ and UFES universities, in the *LinkedDataBR*[210] Project, aiming at building an infrastructure to support open data release using Semantic Web standards and technologies. An ETL4LOD consists in a set of plugins, developed in JAVA, which extends the *Pentaho Data Integration* functionalities, an ETL tool widely used, providing transformation of data from different sources for connected data.

In the same way that the tool has been adapted for the treatment of data, it also includes the treatment of metadata, to support the FAIRfication process as a whole.

It should be noted that the adopted modeling allows ontologies of interest that may arise to be incorporated into the database, serving to make additional annotations that will contribute to reducing the ambiguity of the data and metadata treated.

208 Available on: WHO-COVID-CRF: https://github.com/FAIRDataTeam/WHO-COVID-CRF. Access on: 20 Sept. 2024.

209 Available on: ETL4LOD: available in https://github.com/johncurcio/ETL4LODPlus. Access on: 20 Sept. 2024.

210 Available on: https://memoria.rnp.br/_arquivo/gt/2010/GT-LinkedDataBR_fase1.pdf. Access on: 20 Sept. 2024.

17.3.4 Data release

Following VODAN network guidelines, survey data must be made available in the connected data format, using RDF standard. Following trends in the research data management and its availability in institutional or thematic repositories, one of our alternatives for data release was the use of a repository platform. In VODAN BR, we chose Dataverse[211], as it is the platform previously selected by the coordinating institution of GO FAIR Health Brazil, Oswaldo Cruz Foundation, for releasing research data.

Data verse is an open-code data repository, developed by the Institute of Quantitative Social Sciences of Havard (IQSS), to store, share, release, cite, explore and analyze research data. The repository hosts several virtual archives called data verses. Each data verse contains sets of datasets, and each dataset contains metadata and descriptive data archives (including documentation and ode that accompany data). As a method of organization, a data verse can also contain other data verses.

To increase the reuse of data, in addition to datasets in the RDF standard, the project established other two formats of distribution: the first, in a triple store supported by a DBMS graph using the GraphDB[212] tool, and the second, in .csv format, for a more traditional use of data.

GraphDB is a DBMS for data banks in graph structure, also used as triple store RDF, which provides an agile structure for release and consumption of connected data. This consumption is performed through SPARQL[213] language (SPARQL Protocol and RDF Query Language), a language for semantic queries with a protocol for accessing data in RDF. Therefore, in an initial proposal, each participating hospital can have their data available in different formats and platforms of distribution, according to their convenience and licensing it defines.

17.3.5 Publication in FAIR DP VODAN BR

As previously mentioned, a FAIR DP is an infrastructure to store and access data that aims to: (i) allow data holders to expose their datasets in accordance with FAIR principles; (ii) make discovery of information on FAIR DP easier by data consumers, in a FAIR DPs network; (iii) establish mechanisms that manage consumers' access, according to licenses and restrictions imposed to data by its managers; (iv) provide data holders with indicators of access on (meta)data available; and (v) provide interaction of data for humans through the Graphical User Interface – GUI, and for software agents, using Application Programming Interface – API (Santos *et al.*, 2016).

To promote standardization for the VODAN FAIR DPs, the reference metadata was established and structures in the RDF model by the FAIR Data Team. This standardization defines a set of rich metadata that describe infor-

211 Available on: The Dataverse Project – Available in https://dataverse.org. Access on: 20 Sept. 2024.

212 Available on: GraphDB. Available in: http://graphdb.ontotext.com/. Access on: 20 Sept. 2024.

213 Available on: SPARQL 1.1 Query Language. https://www.w3.org/TR/sparql11-query/. Access on: 20 Sept. 2024.

mation such as, for example, structural and internal coherence of data, licenses and reference sources, access conditions, context, and provenance (Santos, [2020?]).

Another important aspect considered was the diversity of information from data sources (Cruz *et al*., 2020) to be collected, managed and made available in FAIR DP. In short, data provenance plays a significant role in scientific or even commercial projects. It can be defined as a historical documentation of an artifact (object, data, or dataset) generated by an agent-driven procedure (person, process or computational system). It enables scholars to understand and be able to assess the importance and context of the creation, application, or reuse of that artifact more accurately. Provenance is a type of metadata that enhances the quality assurance and veracity of data or datasets. It assists in managing project data as well as supporting reproducibility and reliability.

The provenance of data in the VODAN BR project could be useful for researchers and health professionals who seek to understand the effects of the pandemic. For example, provenance information can be incorporated at the record level by assigning descriptors as part of the data transformation process (e.g., whether a diagnosis was entered by a physician or derived from a version of the form, or whether the data comes from the ELT process). These details are important because datasets that include records from multiple sources that are indistinguishable in general-purpose databases end up generating very different profiles and analyses.

Following the guidelines of VODAN implementation network and the tutorials developed by VODAN Africa&Asia, the VODAN BR FAIR DP will release the metadata referring to the repositories and their datasets, describing in detail, the data sources and their items. It will thus become part of the FAIR DPs VODAN federation, which aims to facilitate the dissemination/release of metadata about COVID-19 data, promoting access to this data by software agents and humans (Santos *et al.*, 2016).

17.4 FINAL CONSIDERATIONS

The development of a computational asset in the form of a platform for the availability of research data regarding viral outbreaks in the middle of a pandemic is a great challenge, both in terms of computational aspects and public health aspects. As presented throughout this chapter, VODAN BR Project has been working continuously to implement its platform, maintaining an overview of the data, from the moment of its capture to the availability of metadata associated with repositories and datasets FAIR DPs.

The experience in the project reinforces the importance of FAIR DP in the infrastructure, not only as an essential element for the federation of access points and search and reuse mechanisms for FAIR data, but also for supporting sensitive research data that requires some degree of confidentiality, as is the case with patients' data. In this aspect, through the FAIR DPs, access to metadata referring to research data is made available, giving them visibility and accessibility. However, effective access to data respects well-defined conditions, promoting "data as open as possible and as close as necessary" (Wilkinson *et al*., 2016).

Among the lessons learned regarding the platform established for VODAN BR project, the lack of tools that help the FAIRfication process as a whole was observed. Some solutions adopted can be automated, improving the process

and making the platform more stable to meet new challenges. An example that can be automated occurs in the process of publishing distributions of a dataset in the *Dataverse* repository and its associated metadata in FAIR DP.

Finally, we have end goals similar to those of the VODAN Africa&Asia network. The challenges being experienced during all phases of the Project provide different and complementary visions, providing a wealth of experiences that must be observed and analyzed, for the establishment of good practices to be taken to other FAIR implementation networks.

ACKNOWLEDGEMENTS

This work has been elaborated through multiple efforts. The authors would like to thank the VODAN-BR team, the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Financing Code 001, 88887.613048/2021-00, CNPq – Financing Code 315399/2018-0 and 306115/2021-2, FAPERJ, Gaffreé Guinle Federal Hospital in Rio de Janeiro, São José Municipal Hospital in Duque de Caxias, professor Mauro Martin from ESDI-UERJ, and, in particular, the group of undergraduate and graduate students from PPGI/UFRJ program and other volunteer students who have been dedicating to the project since April 2020.

REFERENCES

CRUZ, S. M. S.; CAMPOS, M. L. M.; MATTOSO, M. Towards a taxonomy of provenance in scientific workflow management systems. *In*: INTERNATIONAL CONFERENCE ON WEB SERVICES, 2009, Los Angeles. **Anais** [...], 2009. DOI: 10.1109/SERVICES-I.2009.18. Available on: https://ieeexplore.ieee.org/document/5190667. Access on: 01 dec. 2023.

DATA TOGETHER COVID-19 Appeal and Actions. [S.l.: s.n.], [2020?]. Available on: https://www.go-fair.org/wp-content/uploads/2020/03/Data-Together-COVID-19-Statement-FINAL.pdf. Access on: 01 Dec. 2023.

GO FAIR. **Declaration**: Virus Outbreak Data Network (VODAN) GO FAIR Implementation Network, 2020. Available on: https://www.go-fair.org/wp-content/uploads/2020/03/VODAN-IN-Manifesto.pdf. Access on: 01 dec. 2023.

HEATH, T.; BIZER, C. **Linked Data**: Evolving the Web into a Global Data Space. Germany: Springer, 2011.

MONS, B. The VODAN IN: support of a FAIR-based infrastructure for COVID-19. **Eur J Hum Genet** v. 28, pp. 724–727, 2020. DOI: https://doi.org/10.1038/s41431-020-0635-7. Available on: https://www.nature.com/articles/s41431-020-0635-7. Access on: 01 dec. 2023.

SANTOS, L. O. B. S. *et al*. **FAIR Data Points Supporting Big Data Interoperability, Enterprise Interoperability in the Digitized and Networked Factory of the Future**,: Lisbon: ISTE Press, 2016.

SANTOS, L. O. B. S. **FAIR Data Point design specification**, [2020?]. Available on: https://github.com/FAIRDataTeam/FAIRDataPoint-Spec. Access on: 01 dec. 2023.

SATTI, F. *et al*. Semantic Bridge for Resolving Healthcare Data Interoperability *In:* INTERNATIONAL CONFERENCE ON INFORMATION NETWORKING, 2020, Barcelona, **Anais**..., 2020 p. 86-91.

STUDER, R., BENJAMINS, R., FENSEL, D. Knowledge engineering: Principles and methods. **Data & Knowledge Engineering**, v. 25, n. 1–2, p. 161–198, 2018.

WILKINSON M. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Sci Data**. v. 3 n. 160018, 2016. Available on: https://pubmed.ncbi.nlm.nih.gov/26978244/. Access on: 01 dec. 2020.

How to cite this chapter: CAMPOS, Maria Luiza Machado; BORGES, Vania; LOPES, Giseli Rabello; CAVALCANTI, Maria Claudia; MOREIRA, João; CRUZ, Sergio Manuel Serra da. VODAN BR - a platform for supporting COVID-19 data following 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. 17, p. 205-219. DOI: 10.22477/9788570131959.cap17.