Design and implementation of FAIR principles in digital health data in Uganda, Africa

Basajja, M.

Chapter 6

Proof of Concept and Horizons on Deployment of FAIR Data Points in the COVID-19 Pandemic

This chapter is based on the following publication:

Basajja, M., Suchanek, M., Taye, G.T., Amare, S.Y., Nambobi, M., Folorunso, S., Plug, R., Oladipo, F.O., Van Reisen, M.: Proof of Concept and Horizons on Deployment of FAIR Data Points in the COVID-19 Pandemic. Data Intelligence 4(4) (2022). doi: `https://doi.org/9.1162/dint_a_00179`

Abstract

Rapid and effective data sharing is necessary to control disease outbreaks, such as the current coronavirus pandemic. Despite the existence of data sharing agreements, data silos, a lack of interoperable data infrastructures, and different institutional jurisdictions hinder data sharing and accessibility are hindered. To overcome these challenges, the Virus Outbreak Data Network (VODAN)-Africa initiative is championing an approach in which data never leaves the institution where it was generated, but instead, algorithms can visit the data and query multiple datasets in an automated way. To make this possible, FAIR Data Points—distributed data repositories that host machine-actionable data and metadata that adhere to the FAIR Guidelines (that data should be Findable, Accessible, Interoperable, and Reusable)—have been deployed in participating institutions using a dockerised bundle of tools called VODAN in a Box (ViB). ViB is a set of multiple FAIR-enabling and open-source services with a single goal: to support the gathering of World Health Organization (WHO) electronic case report forms (eCRFs) as FAIR data in a machine-actionable way, but without exposing or transferring the data outside the facility. Following the execution of a proof of concept, ViB was deployed in Uganda and at Leiden University. The proof of concept generated a first query, which was implemented across two continents. A SWOT (strengths, weaknesses, opportunities, and threats) analysis of the architecture was carried out, and established the changes needed for specifications and requirements for the future development of the solution.

Keywords: Digital health, Data in residence, FAIR Guidelines, Machine-actionable, VODAN-Africa

6.1 Introduction

The emergence of the novel coronavirus (COVID-19) pandemic has highlighted the critical need for rapid data sharing to enable an effective pandemic response [100]. Acknowledging this urgency, the World Health Organization (WHO) has emphasized the vital role of public access platforms in disseminating complete viral genome sequences and assay protocols. These efforts have significantly contributed to early and accurate infection diagnosis during the ongoing global emergency [101, 102]. However, the sensitive nature of real-time data collected, referred to as Real World Observation Data, poses challenges in centralized data warehousing, such as at the WHO [100, 103]. In response, a distributed data-sharing approach has emerged, enabling data querying within the country or institution where it is collected. The approach aligns with the principle of making data “as open as possible, as closed as necessary” [103].

The Ebola virus outbreak in Africa (2013–2016) highlighted the critical role of efficient information sharing during public health crises (2013-2016) [104, 105]. Advances in data-sharing mechanisms during this crisis fostered a global consensus on the necessity of transparent information sharing, particularly during emergencies [102, 105] but 2015]. An insight at a WHO advisory meeting in September that merely agreeing on information-sharing principles is not enough highlighted the need for practical implementation of technical challenges, including improving the accuracy of data capture, ensuring data quality, and aligning data structures [104, 105]. Drawing from the lessons of the West African Ebola outbreak, it became evident that individual deployment of various digital solutions for data collection lacked interoperability. This lack of coordination was seen in communication and efficiency within the international response to the epidemic [101, 105].

Limited public access to medical information is a major challenge in Africa during the current COVID-19 pandemic. Restricted access to this information hinders understanding and analysis of epidemic progress. For example, although the first reported outbreak in East Africa occurred in Kenya on March 13, 2020, the lack of publicly available SARS-CoV-2 genome data from East African Member States, including Tanzania, Burundi, Uganda, Rwanda, and South Sudan, raises concerns. Similarly, the first case was reported in Nigeria on February 2, 2020, but SARS-CoV-2 genome data remain elusive. The delay in sequencing and sharing Ugandan SARS-CoV-2 genomes until 2 July 2020 underscores the difficulties in obtaining critical genomic, epidemiological, and clinical data essential for an effective pandemic response [101, 106]. The absence of comprehensive, digitally recorded, and interconnected health data poses a substantial obstacle to understanding, analyzing, and effectively responding to the complexities of the COVID-19 pandemic. The following sections delve into a proof of concept and potential horizons for deploying FAIR Data Points as a solution to address these challenges.

6.2 Relevance of the Study

The urgent need for reliable and accurate data to combat the global spread of COVID-19 has prompted national and international cooperation [107]. To support data-driven decision-making, the dissemination of information, whether to the public, research institutions, hospitals, or government offices, must adhere to usefulness, usability, and desirability principles. Traditional data collection methods have introduced health risks for domain experts, data collectors, and stewards, particularly in isolation measures affecting groups and individuals. As the

COVID-19 crisis unfolds, technological solutions have become essential for mitigating risks associated with the collection, access, and distribution of high-quality data. Yet, navigating these challenges is further complicated by data generated across multiple jurisdictions, each governed by varying policies on data sharing, accessibility, and compliance with the EU General Data Protection Regulation (GDPR).

Contrary to expectations given Africa's communal lifestyles, reported COVID-19 cases on the continent have remained relatively low [108]. Potential factors contributing to this unexpected trend include minimal imported cases from COVID-19 hotspots and the advantageous experience gained from combating previous outbreaks like Ebola [109, 110]. However, the looming threat of COVID-19 on Africa's already fragile health systems highlights the critical need for meticulous data collection, storage, and analysis within sovereign local repositories. The pivotal question emerges: How can data be effectively utilized to drive transformative change in Africa's fragmented health systems? Collaborative efforts among all stakeholders are essential to leverage valuable data for understanding COVID-19 trends and geographical distribution. A significant hurdle in addressing these challenges lies in the poor data handling and reporting mechanisms prevalent in Africa [111]. The lack of accurate, reliable, and timely data hinders the effective monitoring and evaluation of COVID-19 interventions. Electronic health information systems, such as Ethiopia and Uganda, remain highly fragmented and lack interoperability, hindering the seamless exchange of information. Despite the plethora of data collected by non-governmental organizations and foreign researchers in Africa, a critical limitation arises as the data is often relocated to Western research centres, rendering it underutilized for local improvements [112].

The existing data collection methods, whether through paper-based health management information systems (HMISs) or automated tools like electronic medical records (EMRs), often fall short of being machine-readable. In hierarchical systems where data is not federated, health professionals contribute data but seldom receive feedback or have control over their data. Establishing a system that empowers health professionals to own and utilize their data, presented in a machine-readable format, offers the potential for local data engineering and analytics. This approach not only minimizes duplicate efforts but also addresses the imperative need for FAIR-ification [170]. The subsequent sections present a proof of concept and explore the potential for deploying FAIR Data Points during the ongoing COVID-19 pandemic, aiming to transform the landscape of data management and utilization in Africa.

6.3 FAIR Guidelines

At the heart of making Real World Observation Data 'Findable,' 'Accessible,' 'Interoperable,' and 'Reusable' (FAIR) lie the FAIR Guidelines [102, 29], an essential framework championed by a diverse group of data stewards and stakeholders. These guidelines underscore the importance of achieving data localisation, ensuring data is reachable over the Internet, adopting machine-readable formats using linked-data approaches for human and machine comprehension, and fostering data convergence through controlled vocabularies within collaborative data communities. The resulting machine-readable information, known as 'metadata,' encapsulates a set of data that describes and informs about other data. However, the realization of these objectives entails navigating diverse political, institutional, and regulatory landscapes governing privacy, security, confidentiality, and the rights associated with health and personal data. Such complexities often act as impediments to seamless data sharing for research and other purposes.

The existence of data silos and non-interoperable data infrastructures further exacerbates the challenge by hindering the machine-assisted querying of multiple datasets.

Addressing these challenges, VODAN-Africa, in collaboration with the GO FAIR Foundation, has introduced a groundbreaking paradigm for data curation, aiming to render data FAIR. In a FAIR ecosystem, data ownership remains within the local context, granting sovereignty over the data [114]. This paradigm shift empowers algorithms to visit data for analysis. The VODAN-Africa initiative, spearheaded by researchers from six African countries—Ethiopia, Uganda, Kenya, Zimbabwe, Tunisia, and Nigeria—in partnership with Leiden University, specifically focuses on combating COVID-19. The primary objective is to establish distributed access to COVID-19 data through the Internet of FAIR Data and Services (IFDS) [114]. IFDS envisions a scalable and transparent framework for routing data, tools, and services [115], aligning with the principles of FAIR to enhance the accessibility and utility of critical COVID-19 information.

6.4 Research Questions and Objective

Research Questions

Main Research Questions: How can the deployment of FAIR Data Points in the context of the COVID-19 pandemic enhance data interoperability and facilitate distributed analytics across different geographical areas and jurisdictions? **Sub-Research Questions:**

Sub-Research Question 1:
What are the essential components and technical specifications necessary for developing and deploying a VODAN-In-a-Box solution for FAIR data management?

Sub-Research Question 2:
How effective is the deployment of FAIR Data Points in real-life situations, particularly in African countries such as Ethiopia and Uganda, in generating and accessing machine-readable semantic clinical patient data?

Sub-Research Question 4:
What are the strengths, weaknesses, opportunities, and threats (SWOT) of the FAIR Data Point architecture in managing COVID-19 data across diverse jurisdictions?

The primary objective of the VODAN-Implementation Network (VODAN-IN) is to establish a robust, international, and interoperable distributed data network infrastructure. This infrastructure addresses the immediate need for evidence-based responses to the ongoing viral outbreak while also facilitating the long-term reuse of data and service infrastructures in future epidemic scenarios [103]. As part of their collaborative efforts within VODAN-IN [102, 116], four major international data organizations—Committee on Data for Science and Technology (CODATA) [117], Research Data Alliance (RDA) [118], World Data System (WDS) [119], and GO FAIR [120]—have united under the consortium named 'Data Together.' Their collective commitment aims to optimize the global research data ecosystem, identifying opportunities and needs that will drive federated infrastructures to meet the evolving demands of data-driven science [115, 121]. Recognizing data and science as a global public good, Data Together, through its 'Data Together COVID-19 Appeal and Actions,' emphasizes the imperative of aligning data and science platforms with the FAIR Guidelines to provide the best available science for pivotal decision-making in the current and future crises [121].

Amidst the urgency of the COVID-19 pandemic, the accelerated implementation of a FAIR ecosystem becomes imperative [31]. This ecosystem is envisioned to enhance the ability to combine, visualize, and utilize data from diverse sources, facilitate nuanced data access and protection, and enable decentralized and machine-assisted analysis [30]. The formation of VODAN-IN stems from the recognition that the valuable data amassed from past and ongoing epidemics is not consistently accessible across different populations and countries [100, 122].

The FAIR-Implementation Network Africa and the VODAN-Africa Research Network were established with a shared mission—combating and containing the COVID-19 pandemic through distributed access to critical data from Africa. The primary focus of VODAN-Africa is to build resources facilitating digital access to COVID-19 pandemic-associated data, adhering to existing governance regulations [290]. The overarching objective of VODAN-IN Africa is to ensure the FAIRness of COVID-19 (meta)data. This FAIR data can then be analyzed in conjunction with globally available data to effectively combat the ongoing COVID-19 pandemic and future infectious disease outbreaks.

Employing participatory ethnography as the research methodology, the study aimed to understand how data analytics of digital health data in residence could be propagated, integrated, and sustained through a FAIR Data Point (FDP). As a participatory ethnography design, the researchers actively engaged in the development of the solution, providing access to key documents and involving critical stakeholders in the process [290]. The investigation delved into the deployment of a FAIR Data Point, accompanied by the installation of Data Stewardship Wizard (DSW) software prepared for this effort, to produce FAIR data and test the hypothesis that visiting data held in repositories in health facilities would be possible. The proof-of-concept criteria included:

Production of machine-readable semantic clinical patient data in residence.

Visitation of metadata through the Internet by deploying FAIR Data Points in six African countries and Leiden University Medical Center.

Querying FAIR Data Points installed in different geographies and jurisdictions using distributed analytics.

Integration of COVID-19 data between FAIR Data Points through data visiting, validating the feasibility of an Internet of FAIR Data and Services (IFDS).

Testing the architecture and tools in a real-life situation.

To assess the practical aspects of the deployment, this article provides a detailed exploration of the deployment experience in two of the six countries, namely Ethiopia and Uganda. Additionally, a SWOT analysis was conducted on the architecture to discern its strengths, weaknesses, opportunities, and threats (SWOT) and pave the way for potential future development [290].

6.5 Development of VODAN In a Box

To fortify data ownership in Africa, VODAN-Africa advocates an innovative approach where data remains within the local institution's underlying database. In this paradigm, algorithms, encapsulated in virtual machines, visit the data locally to conduct analyses addressing specific inquiries. The outcome is pseudonymized aggregate data, allowing combination with results from analogous queries at other locations, contingent upon the authorization of the respective institution. Addressing this challenge, the technical team of the Data Stewardship Wizard (DSW) developed tools with the following specifications:

Emphasis on data ownership and residency in health facilities.

Secure data-visiting over the Internet.

Focus on COVID-19 clinical patient data.

Integration with COVID-19 research data on prevalence.

Deployability in diverse real-life African settings.

This comprehensive toolset, coined "VODAN in a Box" (ViB), encompasses vital components aimed at capturing COVID-19 clinical data and publishing metadata for the datasets [123].

6.5.1 Components

The ViB consists of three essential components: the FAIR Data Point (FDP), the Case Report Form (CRF) Wizard, and the triple repository/query interface, illustrated in Figure 6.1. [123] The ViB tools, DSW, World Health Organization-sponsored (WHO) Electronic Case Report Forms (eCRFs). It allows you to read what machines can read. Features include:

FAIR Data Point (FDP), a tool for publishing definitions for FDP, and FAIR metadata for catalogues, data types, and classifications. It acts as a metadata repository, enabling data owners to disclose the metadata of their digital products fairly, and making it easier to find for consumers,[124].

Case Report Form (CRF) Wizard: This is a tool for managing semantic data. The data entry form is enclosed with the template K. WHO COVID-19 eCRF. It enables the implementation of efficient data management systems based on FAIR data management and provides highly searchable, accessible, usable, and reusable data [125].

The AllegroGraph Triple Store: it works as a knowledge graph repository that provides query interfaces that allow the retrieval of data [126].

The DSW, originally designed for data management planning, now supports FAIR data stewardship. The ViB's components, including the FDPs deployed in each local institution, allow virtual machines to visit them akin to trains docking at stations. This innovative approach enables virtual machines to visit multiple FDPs to obtain answers to their queries [122, 127]. ViB, therefore, functions as a collection of applications supporting the acquisition and exposure of [290]. In essence, ViB encapsulates the technical infrastructure needed for the FAIRification of data, emphasizing interoperability and secure data access within a distributed framework.

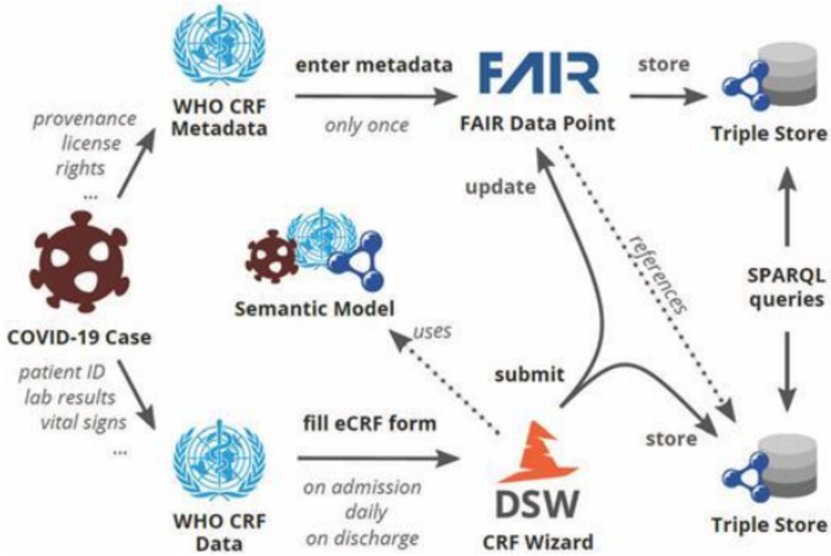

Figure 6.1: Architectural framework for VODAN in a Box platform showing the flow of clinical data, metadata, and semantic annotations within the FAIR data ecosystem

Source: [123]

The COVID-19 electronic Case Report Form (eCRF) is an integral part of the VODAN in a Box (ViB) framework, designed to facilitate machine interpretation of patient data. The eCRF comprises three-page forms, each housing three modules: admission, follow-up, and discharge. These modules group logically related data elements, encompassing vital signs, demographic characteristics, and laboratory test results. The VODAN-Africa initiative sought to develop an eCRF template capable of generating machine-readable data in compliance with FAIR Guidelines.

The ViB deployment architecture is deliberately designed to be user-friendly, even while integrating multiple interconnected services. Hosted as an open-source GitHub repository [128], ViB includes essential configuration files and a Docker-compose.yml. To run ViB, Docker and Docker-compose are prerequisites, applicable for both local testing and production mode. The latter enhances security and availability, offering options such as an HTTPS proxy or a triple store with authentication. ViB seamlessly connects the CRF Wizard, a FAIR Data Point (FDP), and an AllegroGraph triple store. Both the FDP and DSW are actively developed, highly configurable open-source projects. Transitioning ViB to production mode requires preparatory steps such as creating DNS records and generating SSL certificates to ensure secure access. Some configurations involving security measures and local adjustments require manual handling. Collaborative efforts within the network, supported by comprehensive documentation [129], enabled a smooth deployment process, with only minor issues related to misconfigurations or certificate acquisition. On July 22, 2020, Kampala International University successfully deployed the first ViB instance in production.

The AllegroGraph triple-store database, responsible for storing linked data subject predicate object triples, empowers sophisticated queries using the SPARQL query language, including federated queries [126, 130]. By default, the FDP uses an in-memory store, with alternative options including the Native store, AllegroGraph Repository, Graph DB Repository, and Blaze Graph Repository [131]. In Uganda, the ViB implementation specifically utilizes AllegroGraph triple-store storage. User account information for the FDP resides in the MongoDB database. Packaged as a Docker image under FAIR data/FDP, the FDP serves as the core component for managing business logic and operations related to semantic data. The FDP provides an application programming interface (API) for data manipulation in various formats and offers a user interface for human interaction. In a production deployment setup, a reverse proxy ensures secure connections to the FAIR Data Point by handling HTTPS certificates.

6.5.2 Technical Specifications.

The development of the VODAN in a Box (ViB) initiative commenced during the early stages of the COVID-19 pandemic, driven by a strategic focus on COVID-19-related data. In response, the project primarily concentrated on producing machine-readable metadata, with a specific emphasis on the COVID-19 World Health Organization (WHO) electronic Case Report Form (eCRF). Consequently, the scope of ViB was limited to the COVID-19 WHO eCRF, aligning with the project's overarching goal.

The ViB's Case Report Form (CRF) Wizard, designed as a web-based form derived from WHO's Rapid COVID-19 eCRF, facilitates the export of data in the Resource Description Framework (RDF) using the semantic data model of the WHO COVID-19 eCRF. This RDF data can be downloaded as a file or directly submitted to the pre-configured triple store attached to the system. The triple store is specifically designed to facilitate distributed querying across multiple datasets. Although the CRF Wizard is inherently connected to the integrated triple store, its versatility allows for seamless configuration with any triple store that supports a standard SPARQL endpoint for CRF data insertion.

The ViB's CRF Wizard serves as a user-friendly interface for data entry, streamlining the process of converting patient data into the machine-actionable format specified by the CRF. Patient data, derived from sources such as Case Report Forms (CRFs), remains localized, residing within the jurisdiction of the originating hospital or clinic. In the routine workflow of ViB, data entry users or data stewards can:

Enter Data into the CRF Wizard: The interface facilitates the entry of patient data into the CRF Wizard.

Generate a CRF Report: The system enables the generation of a comprehensive CRF report based on the entered data.

Submit the CRF Report to the Triple Store: Once created, the CRF report can be exported to the triple store, which is pre-configured and associated for later storage and retrieval.

The produced data, which is now in a machine-readable format, becomes usable for queries through data-visiting algorithms. It is important to note that successful data visiting relies on the access and control arrangements that were established, although the complexity of these arrangements is beyond the scope of this article.

6.5.3 Metadata Component Choices

The selection of metadata components for the VODAN in a Box (ViB) program was achieved through a collaborative effort by the VODAN-Africa team, conducted over six months of weekly Metadata for Machines (M4M) workshops. The process resulted in nine metadata products with ViB FAIR Data, so we developed standard Points (FDPs), which have been implemented in six African countries. The selected metadata components described in Table 6.1 were intended to meet a variety of requirements to ensure that the data is findable, accessible, interoperable, and reusable in compliance with FAIR principles.

Table 6.1: VODAN metadata component choices for Kampala International University, Uganda FDP 1.1 [29].

#	Metadata Component	FAIR Guidelines	Question for the domain community participating in the Metadata for Machines (M4M) workshop	Kampala International University Uganda FDP 1.1
1	F1	What is the persistence policy for the identifier systems used for digital assets?	Policy statement from the domain community regarding the requirements for identifier systems.	PURL/hospital persistent URL.
2	F2	What is the minimum acceptable profile needed to ensure findability?	Examples: people, institutions, geographical locations, date/time, funding source, publishers, repositories.	Geographical locations, institutions, date/time, repositories, publisher.
3	F4	What are the minimum acceptable indexing requirements and limitations?	Metadata that complies with indexers and search algorithms.	Google.
4	A1.1	What access policies and protocols are in place for this community?	Self-describing access restrictions, based on policy.	Hospital patient data is only accessed by the hospital data management personnel.

Continued on next page

Table 6.1 continued from previous page

#	Metadata Component	FAIR Guidelines	Question for the domain community participating in the Metadata for Machines (M4M) workshop	Kampala International University Uganda FDP 1.1
5	A2	What is the persistence policy for the metadata?	Policy statement from the domain community regarding the requirements for metadata longevity.	In Uganda, archiving metadata to be saved permanently is done.
6	I2	How FAIR are the vocabularies?	Metadata linking to canonical descriptions of terminology systems, vocabularies, and ontologies.	Open Science Framework.
7	R1.1	What usage license(s) will be used?	Metadata linking to machine-actionable license.	CC License.
8	R1.2	What are the minimum provenance metadata needed to ensure reuse?	Examples: laboratory methods, analytical methods, computational platforms.	
9	R1.3	What FAIR Implementation Profile (FIP) is used to make data/metadata FAIR?	A machine-actionable list of the FAIR-enabling resources used to make the data and metadata FAIR.	VODAN FIP.

During a six-month series of Metadata for Machines (M4M) workshops, the VODAN-Africa team carefully selected nine customized metadata components for each Fair Data Point (FDP) in six African countries. These metadata components addressed crucial aspects such as the persistence policy for identifier systems (F1), the minimum profile for findability (F2), acceptable indexing requirements (F4), access policies (A1.1), metadata persistence policy (A2), FAIRness of vocabularies (I2), usage licenses (R1.1), minimum provenance metadata (R1.2), and the FAIR Implementation Profile (FIP) for making metadata FAIR (R1.3). Over time, additional metadata features have been refined and updated across different countries to enhance their applicability and alignment with local contexts.

The active participation of five data managers played a pivotal role in the six-month trainer preparation program. These data managers were instrumental in the preparatory training process, actively collaborating with health agencies, regional health authorities, and university officials. Their responsibilities include setting up local servers, purchasing domain names, setting up domain names via DNS, installing ViB on Virtual Private Servers (VPS), analyzing ViB for data generation, populating COVID data from various sources, and integrating with the first SPARQL.

This period marked a critical initiation where data stewards, by engaging in data collection and FDP setup, retained ownership of data within their respective institutions. Through internal surveys, practical assignments, and discussions, it was observed that data stewards found the WHO standard COVID-19 eCRF Wizard user-friendly. Once COVID-19 data was entered into the eCRF, the template automatically converted the raw data into RDF format, storing it in the triple store as FAIR-compliant data. This entire process, from training to installation and query, spanned four months. Subsequently, a phase of documentation, assessment, and reflection ensued, preparing for the upcoming stages of the project.

6.6 Proof of Concept

6.6.1 Deployment in A Real-life Situation

For the proof of concept, deployment in a real-life situation was critical. The research team developed the following steps for the deployment: Step 1: Testing of the proposition in a real-life situation Step 2: Presentation of a clear proposition to stakeholders in universities, university hospitals, other hospitals, clinics, and ministries of health Step 3: Analysis to ensure compliance with relevant Regulatory Frameworks Step 4: Approval by stakeholders in all locations Step 5: Establishment of FAIR Data Points in two African and one European countries reachable over the Internet Step 6: Machine-actionable data production (test data and real data) Step 7: Running queries over the Internet across the FAIR Data Points that visit the machine-actionable data Step 8: Completion of proof of concept

This describes the initiation of the VODAN-Africa project, highlighting the crucial role of capacity building through the training of trainers. The objective was to empower African experts, including researchers and traditional data stewards from health facilities, with the skills needed to generate FAIR (Findable, Accessible, Interoperable, and Reusable) machine-actionable COVID-19 data. The training sessions focused on using the Electronic Case Report Form (eCRF) wizard, a component of the ViB, to facilitate the entry of structured and standardized data. Participants were trained not only in data entry but also in the subsequent steps of publishing metadata on FAIR Data Points (FDPs) and querying multiple triple stores across FDPs using SPARQL (SPARQL Protocol and RDF Query Language).

At the same time, Dockerized FDP and ViB were deployed together with training workshops. Technical dockerization was used for packaging applications and dependent containers for scalability and ease of use. On 22 July 2020, Kampala International University reached a milestone with the launch of the first FAIR Data Point for COVID-19 data called VODAN in a Box (ViB). This is a significant step being made towards realizing the VODAN-IN Africa initiative that provides a platform for collecting and disseminating data on COVID-19 that is suitable for device operations in line with FAIR principles.

6.6.2 Procedure for Proof of Concept

The proof of concept is inclusive of the World Health Organization (WHO) Electronic Case Report Form (eCRF), which can be read by humans and machines and distributed to partners in different countries. This Electronic Case Report Form is the basis for creating machine-readable metadata on FAIR Data Points (FDPs). The FDPs were configured to be viewable and accessible through the Internet. This allowed them to be discovered by algorithms through the VODAN FDP community.

The steps that are key in designing a proof of concept:

Preparing the WHO Electronic Case Report Form: A WHO Electronic Case Report Form that can be read by humans and machines was developed and distributed to partners in various locations.

Visibility and Accessibility of FDPs: FDPs were designed to be visible and accessible through the Internet. This, along with the appropriate licenses, ensured that algorithms running on the Internet could be accessed.

Automation of Data Exchange: The already installed FDPs supported the automated cross-country exchange of observational COVID-19 patient data. This facilitated querying of FAIR-compliant data by hospitals, facilities, and researchers across different countries while adhering to local data protection and governance laws.

Mock Data Entry: Mock data was entered into the ViB using the CRF Wizard tool, which adhered to the semantic model based on the WHO eCRF. This was achieved using SPARQL, an RDF query language.

Example Query: An example query was formulated to retrieve information from the triple store. The query was designed to present the names of healthcare facilities that admitted male patients with a fever (body temperature greater than or equal to 38 degrees). This query was executed through the AllegroGraph query interface.

SPARQL Query Execution: SPARQL queries were entered into the AllegroGraph query interface, and service call requests were made to the SPARQL endpoints of the three countries involved (Uganda, Kenya, and the Netherlands). The results displayed the facility names against the number of COVID-19 patients.

The proof of concept aimed to demonstrate the functionality of VODAN in a Box (ViB) in a real-world setting, showcasing the viability of querying FAIR-compliant data across different geographical locations.

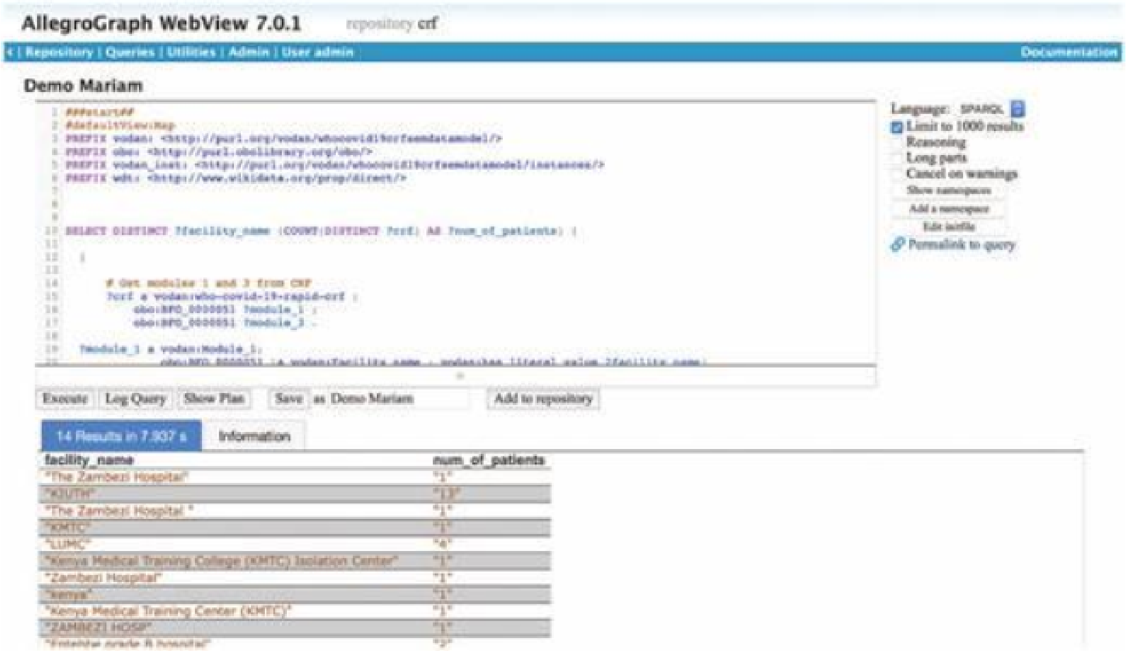

Figure 6.2: Output produced by the VODAN in a Box proof of concept demonstrating federated querying and retrieval of distributed health data from participating repositories

6.7 SWOT Analysis

The reflection period was used to reflect on the potential of the architecture, but also to identify limitations and alternatives. To assess the architecture, an analysis was carried out of strengths, weaknesses, opportunities, and threats (SWOT). The analysis was carried out as an internal evaluation by the data stewards involved in the research.

6.7.1 Strengths

The strength of the architecture was identified as its potential to organize data ownership and data visiting, and generate potential for the data within the health facilities to contribute to the quality of healthcare. Together, the FDPs form a distributed data network through which queries are able to visit the data and pose research queries on the data therein without the necessity of downloading or importing the data to another location. This distributed network of human and machine-readable data will constitute the Internet of FAIR Data and Services. The tools were deployed in diverse real-life African settings and generated interest.

Of the original requirements for the proof of concept, the following assessment was made by the resident data stewards within the hospitals and facilities. Data ownership and residency in health facilities were feasible, and secure data-visiting over the Internet was possible.

In terms of strengths identified, the data stewards found that the design and deployment of the ViB, together with the FDP project, uses some of the best practices in the modern computing world, such as distributed computing, software virtualization/containerization, and semantic interoperability. Scientific fields that generate high-dimensional data sets have embraced the principle of data visitation (i.e., moving the code to data instead of data to code). This is because moving big data across networks takes massive computational resources and time. Besides, such data transfers incur myriad bottlenecks, including network issues, particularly with internet speed, especially in Africa. Code and data usually exist on different machines/servers, so either the code or the data has to be moved to execute the code on the data. As the code (usually contained in an executable file) is almost always smaller than the data, it is better to move the code to the data (or closer to the data). Apache Hadoop, the distributed computing framework for big data storage and processing, which consists of a storage layer (Hadoop Distributed File System) and a batch processing engine (MapReduce) [132], is based on the design principle that moving computation is cheaper than moving data [133]. In addition to this, further ViB development is planned to allow local customizations and other data types.

6.7.2 Weaknesses

A critical issue was the inflexibility of the DSW Wizard tool towards the FAIR metadata production of research data and health data types. The limited scope of the WHO eCRF within the DSW Wizard did not allow the production of any alternative data in a machine-readable format. The alternative data, including research data on COVID-19, could not be queried despite the availability of the FDP. This was a setback in the integration of both the patient data and research data. The focus on COVID-19 clinical patient data needed more flexible data entry templates. The DSW tool did not allow for bulk input, which would be a more efficient way of data production. The DW Wizard also did not allow seamless integration with the workflows and requirements in the health facilities, especially regarding the requirements to upload data in the HMISs (such as the District Health Information System, DHIS2). Moreover, the integration with COVID-19 research data on prevalence was limited by the inflexibility of the tools used to produce semantic data.

Subsequent to the proof of concept, and to overcome the limitations encountered, VODAN-Africa has extended collaboration with developers in the semantic world to enhance and develop education and development. Moreover, to create greater flexibility in the production of semantic languages, the VODAN-Africa group has engaged with the digital online learning courses developed to help advance these skills among computer science and data science graduates so that data managers will obtain the skills for the development of machine-readable semantic data in Africa [134].

Other weaknesses included a lack of flexibility in creating new templates and an inability to import bulk data when the immediate requirements go beyond the WHO eCRFs. The DSW toolset also had further weaknesses: the bulk inputting of data was not possible; it was not adaptable to seamlessly integrate into clinics and hospitals; it could not integrate different types of data on COVID-prevalence, including those derived from scientific projects; it could not solve the potential problems of data export from the Wizard into the DHIS; and it could not be easily extended to include the functionalities based on differentiated situations on the ground. Key considerations for local deployment include the need for the localization of user interfaces and the semantic content, and adaptation to accommodate languages and cultures within the system. Ease of use and flexibility, as well as ease of deployment through Dockerization or virtualization [135] for local deployment, with due consideration for infrastructure (such as bandwidth and storage), are also equally important. The system also needs to be customized for it to seamlessly integrate with FPDs and to be able to add new features beyond eCRFs.

6.7.3 Opportunities

Opportunities were identified in response to the lack of flexibility in creating templates in existing tools. The Center for Expanded Data Annotation and Retrieval (CEDAR) Workbench [136] was approached as an alternative to the implementation for data entry. CEDAR provides an easy-to-use and intuitive template creation feature.

In addition, it was identified that both the CRF Wizard and the CEDAR Workbench data entry tools lacked the ability to import bulk data, as reported by a number of health facilities. To overcome the limitation on bulk data import within the CEDAR platform, the VODAN team developed a custom script for the two available solutions. The script allows for the upload of bulk comma-separated data to the CEDAR server using its API, as shown in Figures 3 and 4. While it is easy to install the ViB using Docker, it comes bundled with the DSW eCRF Wizard.

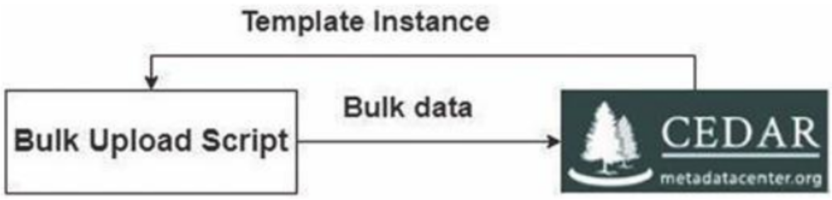

Figure 6.3: Schematic diagram illustrating Bulk data upload workflow for the CEDAR platform, depicting bulk data ingestion into the platform using a dedicated upload script

The construction of metadata templates and the filling in templates to generate high-quality metadata seems to be versatile in the CEDAR Workbench [136]. The Representational State Transfer (REST) based environment provides the CEDAR web-based tool to ease communication with other tools. Multiple file formats are available as JSON, JSON-LD, or RDF to produce machine-readable and actionable metadata and data.
Given the project's interest in going beyond eCRF, there was a need for the localized deployment of the CEDAR Workbench, which is currently a cloud service. This is essential in the context of low-resource settings and the need for federated data stores deployed on-premises. The team is currently working on a locally deployable editor for CEDAR.
Further opportunities were identified in the possibility of using and visualizing data at the point of care through data dashboards. An aggregate dashboard based on data visualization of the locally deployed data was also conceivable, strengthening the possibility of monitoring the data for early health risk detection.

6.7.4 Threats

Threats identified pertained to the lack of ability to integrate the requirements of the ministries of health to deploy data in HMISs. The lack of an output component to the required systems would mean that data clerks would have to input the data twice, which would undermine the objective intended: to edit data once for interoperable and reuse across different needs. The lack of integration with other systems was regarded as a serious threat that undermined the purpose of the innovation. In addition, training and capacity-building was identified as critical to support the system with knowledgeable data managers.

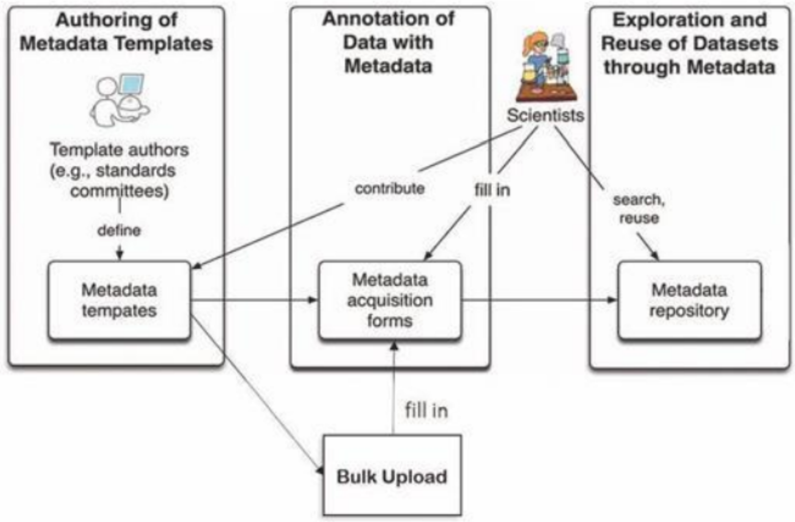

Figure 6.4: Schematic rendering of the CEDAR ecosystem supporting metadata management showing interaction with the bulk upload mechanisms

Source: [137]

6.8 Conclusion

During the ongoing COVID-19 pandemic, there remains a significant scarcity of publicly accessible data across Africa. It is challenging to ascertain whether sufficient data exists or if existing data is inaccessible due to limitations in availability and sharing mechanisms. The establishment of FAIR Data Points (FDPs), designed to support diverse data types, presents a viable solution to the challenges of making data Findable, Accessible, Interoperable, and Reusable (FAIR). This may enhance Africa's role in the fight against the coronavirus pandemic; ensure data ownership (i.e., that African data resides in Africa and is not removed to storage facilities elsewhere); and enable Africa to be a huge resource of verified data and strengthen data-informed health systems for Africa and the world.

This study explored the deployment of the VODAN in a Box (ViB) tool to enhance the interoperability of COVID-19 clinical patient data, ensuring that the data remains locally stored and accessible as machine-readable semantic metadata. The proof of concept generated a first query, which was implemented across two continents. These results are promising and highlight the potential for further development and refinement of the ViB tool. However, a SWOT analysis revealed several limitations associated with the ViB tool that require attention. The most significant limitation identified was the lack of flexibility in semantic data production. This could be solved by localizing existing platforms, such as CEDAR, to locally deployable versions.

Additionally, the study highlighted the necessity of enhancing the capacity of data stewards through targeted training in semantic data science to support the effective development and deployment of such tools.

6.9 Discussion

The task at hand was to FAIRify the COVID-19 data (i.e., make it Findable, Accessible, Interoperable, and Reusable), with the long-term vision of also making it federated and artificial intelligence (AI)-ready, in the sense that machine learning and AI approaches can discover meaningful patterns in data from future epidemic outbreaks [122].

The ViB-FDP was designed to host WHO electronic Case Report Forms (eCRFs) for patients, enabling data querying while ensuring the data remains localized within the jurisdiction of the hospital or clinic where it is generated. Data stewards identified a critical limitation in the WHO eCRF, noting its restricted capacity to handle diverse data types, such as FAIR metadata for both research and health data. The limited scope of the WHO eCRF within the DSW Wizard did not allow for the production of this data in a machine-readable format. Other data types collected within the different hospitals and clinics could not be queried despite the availability of the FDP. This limitation posed a significant challenge to the integration of patient and research data. It was concluded that the integration of the data on COVID-19 requires facilitation from semantic platforms, such as CEDAR, which supports the creation of customized metadata templates, with convergence developed to facilitate queries.

There also needs to be a change for the quick adoption of the system by health facility management procedures and HMIS requirements. Technical support should also be locally available for sustainability. Proper documentation and the availability of training and administration manuals are essential for the technical team, as well as end users. Governance documents also help in setting the rules of engagement within the context of the health system. As the end goal of a digital health system is to be able to use data for decision-making, there have to be features that facilitate data utilization, such as analytics tools, dashboards, and other reporting tools and formats.

Pandemics necessitate rapid responses in both health services and data management to support effective decision-making and strategic planning. While the WHO eCRF serves as an effective tool for collecting COVID-19 data, it lacks the additional HMIS elements required by routine health systems. This emphasises the need for a tool capable of integrating additional HMIS elements while adhering to FAIR principles. This would include new tools for creating new and adapting existing templates to accommodate the information needs of health systems. These tools need to work well with existing tools without creating much pressure on health workers and avoiding duplication of effort. Hence, other data production tools should be explored, and new integration tools created beyond the ViB, which is dependent on the WHO's COVID-19 eCRF. This requires the enhancement, customization, and localization of existing and new tools to accommodate demand. It was concluded that the DSW toolset was not the best match with the requirements and specifications, and alternative tools, such as the CEDAR Workbench, should be explored in further research.