RESEARCH PAPER

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

Mariam Basajja[1†], Marek Suchanek[2], Getu Tadele Taye[3], Samson Yohannes Amare[3], Mutwalibi Nambobi[4], Sakinat Folorunso[5], Ruduan Plug[1], Francisca Oladipo[6,7], Mirjam van Reisen[7,8,9]

[1]Leiden Institute of Advanced Computer Science, Leiden, 1011NC, Amsterdam, the Netherlands

[2]Faculty of Information Technology (FIT), Czech Technical University, 15000 Prague, Czech Republic

[3]Mekelle University, P.O. Box 1871 Mekelle, Ethiopia

[4]Kampala International University, 256, Uganda

[5]Department of Mathematical Sciences, Olabisi Onabanjo University, P.M.B 2002, Ago-Iwoye, Ogun State, 120005 Nigeria

[6]Federal University, 260101 Lokoja, Nigeria

[7]Virus Outbreak Data Network-Africa

[8]Tilburg University, P.O. Box 90153 5000, the Netherlands

[9]Leiden University Medical Centre (LUMC), Leiden University, 1310 Leiden, the Netherlands

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

Citation: 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), 917–937 (2022). doi: 10.1162/dint_a_00179

Submitted: March 10, 2021; Revised: June 10, 2022; Accepted: July 15, 2022

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, lack of interoperable data infrastructures, and different institutional jurisdictions hinder data sharing and accessibility. 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

† Corresponding author: Mariam Basajja, Leiden University (Email: mariam.basajja@gmail.com, m.basajja@liacs.leidenuniv.nl; ORCID: 0000-0001-7710-8843).

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 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.

ACRONYMS

AI	artificial intelligence
API	application programming interface
CEDAR	Center for Expanded Data Annotation and Retrieval
CRF	case report form
DHIS	District Health Information System
DNS	Domain Name System
DSW	Data Stewardship Wizard
eCRF	electronic case report form
FAIR	Findable, Accessible, Interoperable Reusable
FDP	FAIR Data Point
FIP	FAIR Implementation Profile
HMIS	health management information system
IFDS	Internet of FAIR Data and Services
M4M	Metadata for Machines
RDF	Resource Description Framework
SWOT	strengths, weaknesses, opportunities and threats
ViB	VODAN in a Box
VODAN	Virus Outbreak Data Network
VODAN-IN	VODAN Implementation Network
WHO	World Health Organization

1. INTRODUCTION

In the wake of the ongoing novel coronavirus (COVID-19) pandemic [1], institutions worldwide have realised the need for rapid data sharing in order to swiftly and effectively combat the pandemic. Recently, scientists from the World Health Organization (WHO) reported that "the release of full viral genome sequences through a public access platform and the polymerase chain reaction assay protocols that were developed as a result made it possible to accurately diagnose infections early in the current emergency" [2, see also 3]. Due to the sensitivity of the data being collected during the pandemic in real time (Real World Observation Data), sending the data to a central data warehouse/repository such as at WHO, is not possible [1, 4]. This automatically enables the querying of data in a distributed way, as it cannot leave the country or institution

at which it is collected. It is, therefore, only under controlled circumstances that such data can be accessed and such access requires that the data is "as open as possible, as closed as necessary" [4].

The obstacles to data availability and data access in public health emergencies came to the forefront during the Ebola virus outbreak in Africa from 2013 to 2016 [5, 6]. Taking lessons from the deficiencies in data-sharing mechanisms highlighted during the Ebola epidemic, agreements were reached on the need for the open sharing of data and results, especially in public health emergencies [3, 6] However, at a WHO consultative meeting in September 2015, participants acknowledged that "it is not enough for parties to simply agree, in principle, on sharing primary data" [5]. Although there is a growing general consensus on the need for timely data sharing, fuelled by the urgency of public health emergencies, the technical challenges involved in implementing data sharing agreements (such as simplifying and standardising data capture procedures, ensuring data quality, and harmonising disparate data platforms) have crippled efforts to develop a rapid and effective public health response based on the best available evidence on major disease outbreaks [5, 6]. During the West-African Ebola outbreak, different digital solutions deployed individually in the collection of various datasets (e.g., mobile phone data, diagnostic app data, social network data, geographic mapping of epidemiological data) were not interoperable, making coordination and communication between them very difficult. In turn, this limited the efficiency and effectiveness of the combined international response to the epidemic [2, 6].

In the ongoing COVID-19 pandemic, very few clinical data collections are publicly available in Africa. It is not easy to tell whether these data exist or not, as access to them is restricted or rather difficult. For example, it was noted that even though the first reported case of COVID-19 in East Africa was on 13 March 2020, in Kenya, 10 weeks later, there was not a single SARS-COV-2 genome publicly available from any of the East African Community countries (Kenya, Tanzania, Burundi, Uganda, Rwanda, South Sudan). Likewise, in Nigeria, the onset of COVID-19 was reported on 2 February 2020, but the SARS-COV2 genome data is still not publicly available. It was not until 2 July 2020 that the first Ugandan SARS-COV-2 genomes were sequenced and made publicly available [2]. A combination of genomic sequence data, epidemiological and clinical data is necessary not only to help design an appropriate response to the pandemic and inform public health and social measures employed to counter the spread of the virus, but also to study and monitor changes in the behaviour of the virus that may have a bearing on transmissibility, severity, diagnosis, and the development of vaccines and other therapies [7]. Such data are also important in ensuring that the virus does not spread undetected. A lack of digitally recorded and connected health data makes the ability to understand and analyse the situation an uphill task.

2. RELEVANCE OF THE STUDY

National and international cooperation has been striving for reliable and accurate data to curb the spread of COVID-19 globally [8]. To bring about data-driven decision making, information needs to be shared (with the public, research institutions, hospitals, and government offices) in a way that is useful, usable, and desirable. However, collecting data in a traditional way has become a health risk to domain experts, data collectors and data stewards. The isolation of groups and individuals hinders the collection, access,

and distribution of information. As the full scope of the COVID-19 crisis is unclear at the moment, technological solutions are expected to overcome some of the risks involved in collecting, accessing, and distributing good quality data. Further challenges are faced during attempts to use data generated by individual institutions falling under multiple jurisdictions, which have different policies on data sharing and accessibility, privacy, and confidentiality, as well as different degrees of compliance with the EU General Data Protection Regulation (GDPR).

Given the communal lifestyle in many parts of Africa, the prevalence of COVID-19 cases was expected to be high. However, the number of COVID-19 cases and deaths reported has been relatively low, compared to other places [9]. There are various reasons as to why COVID-19 is not spreading as expected in Africa. Firstly, it could be due to the low level of imported cases from COVID-19 hotspots [10] and, secondly, previous exposure to outbreaks such as Ebola may have given African countries a comparative advantage in tackling it. For example, the experience gained in fighting Ebola in the Democratic Republic of Congo can be applied to other health emergencies, such as COVID-19 [11].

However, the COVID-19 pandemic could have tragic consequences for people in Africa due to already poor health systems. To better understand and respond to the vast range of issues contributing to the COVID-19 crisis, data needs to be collected, stored, and analysed with care in a sovereign local repository. Perhaps the central question is how data could be used more widely in a way that brings about a fundamental change in the disparate health systems in Africa. A collaboration by all responsible stakeholders needs to be supported with valuable data to understand the trends and geographical distribution of COVID-19.

The data handling and reporting mechanisms in Africa are very poor [12]. Thus, lack of accurate, reliable, and timely data has become a recurring issue in most African countries. This poor culture of data use has created constraints on the effective monitoring and evaluation of the interventions to fight COVID-19. The electronic health information systems in Ethiopia and Uganda, as well as in some other African countries, are severely fragmented and not yet interoperable. Various policy and strategy documents have been drafted to improve health information systems. However, most of these documents have not been implemented on the ground. Although a mammoth amount of data is being collected in Africa by non-governmental organisations and foreign researchers, this data is not used locally to improve the wellbeing of the people. This is because the data is usually removed from where it was produced and taken to Western research centres [13].

The collection of data in health management information systems (HMISs) in Africa is being done either on paper or using automated tools to inform decision making. Starting from point of care systems, electronic medical records (EMRs) and other mobile and desktop-based systems are in place to record information at an individual level—and this data is aggregated in HMISs, such as the District Health Information System 2 (DHIS2), but tend not to be machine readable. In countries where data is not federated, health professionals collect data and report it up the system hierarchically, but seldom receive feedback and, hence, neither use it for decision making nor have control over their own data. Having a system that allows for health professionals to own and use their own data and that makes it available in a machine-readable format would allow them to carry out data engineering and analytics locally. This avoids the duplication of efforts when there is a need for FAIRification [14].

3. FAIR GUIDELINES

The FAIR Guidelines, which were promulgated by a group of diverse data stewards and stakeholders, are crucial to making this Real World Observation Data 'Findable', 'Accessible', 'Interoperable and 'Reusable'—or FAIR. The FAIR Guidelines [3, 15] promote data localisation through the repositing of data within residence; data visiting by making data reachable over the Internet; machine readability by making use of linked-data approaches under the proviso that data is readable by both humans and machines; and data convergence by introducing controlled vocabularies within communities that collaborate on data. Such data are referred to as 'metadata'—which is a set of data that describes and gives information about other data in machine-readable format. Achieving all these things involves tackling the different political and institutional jurisdictions and regulatory environments regarding the privacy, security, confidentiality, and rights of health data and personal data, which can hinder data sharing for research and other purposes. Data silos and data infrastructures that are not interoperable make it impossible to query more than one dataset in a machine-assisted way.

VODAN-Africa, in collaboration with the GO FAIR Foundation, introduced a new paradigm for data-curation that aims to make data FAIR. When data is FAIR, data ownership is retained in residence and it becomes sovereign. Then, algorithms can visit the data for analysis. VODAN-Africa [16], started with researchers from six African countries—Ethiopia, Uganda, Kenya, Zimbabwe, Tunisia, and Nigeria—together with Leiden University, to fight against COVID-19. The aim was to establish distributed COVID-19 data access via the Internet of FAIR Data and Services (IFDS) [17], which envisions the scalable and transparent routing of data, tools, and services [18].

4. RESEARCH OBJECTIVE

The VODAN-Implementation Network (VODAN-IN) is focused on building an international and interoperable distributed data network infrastructure, not only to support evidence-based responses to the current viral outbreak, but also to reuse the resulting data and service infrastructure for future outbreaks [4]. With VODAN-IN as one of their joint activities, four major international data organisation have come together to form a consortium known as 'Data Together' and laid down their joint commitment "to optimize the global research data ecosystem and to identify the opportunities and needs that will trigger federated infrastructures to service the new reality of data-driven science" [17, 18]. These organisations include the Committee on Data for Science and Technology (CODATA) [19], Research Data Alliance (RDA) [20], World Data System (WDS) [21] and GO FAIR [22]. Galvanised around the consensus that data and science is a global public good, Data Together, in its 'Data Together COVID-19 Appeal and Actions', stipulates that in order to provide the best available science to make pivotal decisions in the current COVID-19 crisis in the short term, and similar crises in the long term, it is imperative that data and science platforms and infrastructures are based on the FAIR Guidelines [23].

The current situation requires the accelerated implementation of a FAIR ecosystem [24], which would "maximize the ability to combine, visualize, and use data from many sources; facilitate fine-grained data

access and protection; and allow for decentralized and machine-assisted analysis" [25]. The formation of VODAN-IN was inspired by the fact that the immensely valuable data of past and current epidemics is not always equally accessible for different affected populations and countries [1, 26].

The FAIR-Implementation Network Africa and the VODAN-Africa Research Network were established to fight and contain the COVID-19 pandemic using distributed access to critical data from Africa. The main aim of VODAN-Africa is to build resources that allow digital access to the data associated with the COVID-19 pandemic within the pre-existing governance regulations [27]. The objective of VODAN-IN Africa is, therefore, to ensure that the COVID-19 (meta)data is made FAIR so that it can be analysed in combination with data available globally, so as to effectively combat the current COVID-19 pandemic and future outbreaks of infectious disease.

This research project employed participatory ethnography, which requires the researcher to use, or observe, others using a designed object or system. Participatory ethnography provides a tool for the study of design as a social practice. Being a participatory ethnography design, the researchers participated in the development of the solution, allowing them access to key documents and resource persons [27]. Critical stakeholders were identified and included in the process of development of the solution by the VODAN-Africa team [27].

This investigation sought to understand how the data-analytics of digital health data in residence could be propagated, integrated and maintained through a FAIR Data Point (FDP). In order to test the hypothesis that the visiting of data held in repositories in health facilities would be possible, a query of clinical patient data on COVID-19 held in residence was developed and implemented, after deployment of a FAIR Data Point, together with a Data Stewardship Wizard (DSW) software prepared for this effort and installed for the production of FAIR data. The team carried out a proof of concept against the following criteria, namely that:

- Machine-readable semantic clinical patient data could be produced in residence.
- The meta-data could be visited through the Internet by deployment of FAIR Data Points in six African countries and Leiden University Medical Center.
- The FAIR Data Points installed in different geographies and jurisdictions could be queried using distributed analytics.
- COVID-19 data could be integrated between FAIR Data Points by data visiting, proving the feasibility of an IFDS.
- The architecture and tools developed could be tested in a real-life situation.

Deployment in a real-life situation was important to test the idea and identify the factors that would emerge in a real-life setting [27]. To assess the practical features of the deployment, this article zooms in on the experience of deployment in two of the six countries: Ethiopia and Uganda. A SWOT analysis was carried out on the of the architecture to determine its strengths, weaknesses, opportunities and threats (SWOT) in order to establish its potential future development.

5. DEVELOPMENT OF VODAN IN A BOX

In order to strengthen data ownership in Africa, VODAN-Africa proposes an approach whereby the data never leaves the underlying database of the local institution, but, instead, algorithms packaged in virtual machines visit the data to perform analyses regarding a specific question. What is returned is pseudonymised aggregate data, which can be combined with the results of similar queries at other locations, upon being granted permission by the respective institution. In order to solve the problem at hand, the technical team of the DSW developed tools according to the following specifications:

- Data ownership and residency in health facilities
- Data-visiting over the Internet in a secure way
- Focused on COVID-19 clinical patient data
- Integration with COVID-19 research data on prevalence
- Deployable in diverse real-life African settings

This resulting toolset of components developed by DSW is called VODAN in a Box (ViB).

5.1 Components

ViB was developed to facilitate the capture of virus outbreak COVID-19 clinical data and the publication of the metadata of the datasets for the same. The ViB has three critical components: the FDP, the case report form (CRF) Wizard which is comprised of the WHO electronic case report forms (eCRFs) in a machine-readable format, and the triple store/query interface, as depicted in Figure 1 [28]. The machine-readable production of the WHO eCRF is realised through the DSW. The toolset comprises a data entry interface, the DSW, the output subsystem, the FDP, and the eCRF modelled and embedded in the DSW to facilitate the provision of semantically rich data. In addition, the ViB is supported by the AllegroGraph triple store to run eCRF data and queries. These are combined in one product in the ViB.

The FDP is a tool used for publishing definitions for the FDP itself, FAIR metadata for catalog, datasets, and distribution [29]. The FDP is a metadata repository that allows data owners to expose metadata on their digital objects in a FAIR way and consumers to discover information about the accessible digital objects [29]. The CRF Wizard is a data entry form coupled with a semantic data model of the WHO COVID-19 eCRF [30]. And the AllegroGraph is a triple store that stores a knowledge graph and has a query interface [31].

DSW is a tool that was originally created to have efficient data management planning based on FAIR data stewardship to optimise the findability, accessibility, interoperability, and reusability of data [32]. Subsequent to this, FDPs, which are data repositories, were deployed in each local institution such that virtual machines can visit them in a similar way to how trains dock at stations. In this way, the virtual machines can visit more than one FDP to get their questions answered [33]. For this approach to work, firstly, the data at the local institution has to be FAIRified through retrieving non-FAIR data.

Hence, ViB is a collection of applications that support the acquisition and exposure of the CRFs locally on FAIR Data Points. This research highlighted three components which were created based on the WHO

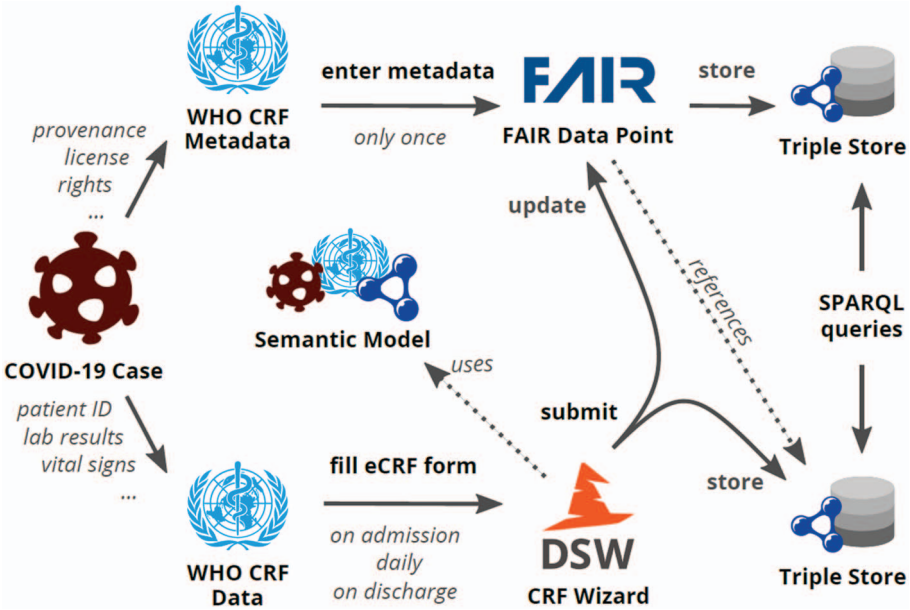

Figure 1. VODAN in a Box architectural framework [28].

COVID-19 eCRF to allow machines to interpret patient data. The wizard has three-page forms that consist of three modules: admission, follow-up, and discharge. A module is a group of logically related data elements collected in a form (e.g., vital signs, demographics characteristics and laboratory tests results). VODAN-Africa sought to create this eCRF template to produce machine-readable data, which conforms to the FAIR Guidelines.

The deployment of the ViB is designed to be relatively simple, even though it consists of multiple interconnected services. It is distributed as an open-source GitHub repository [34] with necessary configuration files and Docker-compose.yml. Therefore, Docker and Docker-compose are required to run ViB, both locally to try it out, as well as in production mode. The production mode provides additional options and services for higher security and availability such as HTTPS proxy or triple store with authentication. The ViB connects the CRF Wizard, an FDP, and a triple store. Both the FDP and DSW are highly-configurable and actively developed open-source projects. To deploy ViB in production mode, the teams from VODAN-Africa had to prepare Domain Name System (DNS) records for those services, including generating SSL certificates for secure access. Some configuration steps needed to be done manually due to security reasons (e.g., setting accessibility, passwords, and permissions) or local configuration (e.g., own metadata layers in FDP). Thanks to successful collaboration within the network and documentation provided [35], there were only minor issues during the deployment related to misconfiguration or problems acquiring certificates. Kampala International University deployed the first ViB in production on 22 July 2020.

The AllegroGraph triple store database, which stores the linked data subject predicate object triples, allows for elaborate queries using the SPARQL query language, as well as federated queries [31]. The FDP

stores its semantic data in a triple store. By default, the FDP uses the In-memory store. The list of possible triple stores include: In-memory store, Native store, AllegroGraph Repository, Graph DB Repository and Blaze Graph Repository [36]. The Uganda ViB uses the AllegroGraph triple store storage. Besides the semantic data, the FDP needs information about the user accounts. These are stored in the MongoDB database. The FAIR Data Point is distributed in Docker image FAIR data/FDP. This is the core component that handles all the business logic and operations with the semantic data. It also provides the application programming interface (API) for working with data in different formats. It provides the user interface for humans. In production deployment, there is a reverse proxy that handles HTTPS certificates, so that the connection to the FAIR Data Point is secure.

5.2 Technical Specifications

The development process started at the beginning of the COVID-19 pandemic and, therefore, researchers agreed that the focus of the project would be on data relevant to COVID-19. For the production of data in the form of machine-readable metadata, a human and machine-readable COVID-19 WHO eCRF was prepared and provided to the partners in each of the locations to produce machine-readable metadata on the FAIR Data Point. The ViB was, therefore, limited to the COVID-19 WHO eCRF. The CRF Wizard is a web form based on WHO's Rapid COVID-19 eCRF. It exports data in the Resource Description Framework (RDF) based on the WHO COVID-19 eCRF semantic data model. The RDF data can be exported as a downloadable file or submitted directly to the pre-configured and attached triple store. The attached triple store supports distributed querying. Nevertheless, the CRF Wizard can be easily configured to use any triple store that uses standard SPARQL endpoint for inserting the CRF data.

The data entry CRF Wizard of the ViB makes it easy to import the patient data into the machine-actionable format of the CRF. The patient data obtained from patient records, such as CRFs, is put in a machine-actionable format. The data remains local, within the jurisdiction of the hospital or clinic that produces the data. In its daily operation, the ViB allows the data entry user/data steward to do the following using the CRF Wizard, with the following process:

- Enter the data into the CRF Wizard
- Generate a CRF Report
- Submit the CRD report to the triple store

Once the data is produced it can be subjected to queries through data visiting algorithms. A precondition for successful visiting at this point remains the access and control arrangements, which is outside the scope of this article.

5.3 Metadata Component Choices

The ViB metadata components were selected for each of the FDPs by the VODAN-Africa team during weekly Metadata for Machines (M4M) workshops over a period of six months (see Table 1). FDPs were successfully deployed in six African countries and nine metadata elements on the FDP were standardised.

Table 1. VODAN metadata component choices for Kampala International University, Uganda FDP 1.1 [15].

Metadata component	FAIR Guidelines	Question for the domain community participating in the M4M workshop	Metadata	Implementation considerations	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	Unless the domain community builds its own identifier services, a widely reused policy will simply default to the policy of existing services (such as DOIs or PURLs) that may be chosen based on a range of other considerations.	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	The schema for many of these metadata elements are often highly generic and reusable.	Geographical locations, institutions, date/time, repositories, publisher
3	F4	What are the minimum acceptable indexing requirements and limitations?	Metadata that comply with indexers and search algorithms	Exceptions for established repositories and domain communities will likely rely on commercial indexing and search such as Google and Bing. Rich and semantically enabled metadata search solutions are emerging in third parties/companies.	Google
4	A1.1	What access policies and protocols are in place for this community?	Self-describing access restrictions, based on policy	Existing authentication and authorisation technologies will likely be reused, coupled with ontologically enabled access conditions.	Hospital patient data is only accessed by the hospital data management personnel.
5	A2	What is the persistence policy for the metadata?	Policy statement from the domain community regarding the requirements for metadata longevity	A metadata standard for this element does not yet exist, but version 1.0 can be easily created referencing the policy document and widely reused.	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	The FAIRness of these resources can be measured using FAIR evaluation tools and these evaluations could also be reported as metadata.	Open Science Framework
7	R1.1	What usage licence(s) will be used?	Metadata linking to machine-actionable licence	See, for example, the licences available in the Open Science Framework.	CC Licence
8	R1.2	What are the minimum provenance metadata needed to ensure reuse?	Examples: laboratory methods, analytical methods, computational platforms	The schema for these metadata elements are often highly specific and idiosyncratic to the use case. In general, these elements will require the most time and resources to develop.	
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	Example: VODAN FIP	VODAN FIP

These included: a persistence policy for identifier systems used for digital assets (F1)[①], the minimum acceptable profile needed to ensure findability (F2), the minimal acceptable indexing requirements and limitations (F4), access policy and protocols in place for this community (A1.1), the persistence policy for the metadata (A2), the FAIRness of the vocabularies (I2), usage licences to be used (R1.1), the minimum provenance metadata needed to ensure reuse (R1.2), and the FAIR Implementation Profile (FIP) to be used to make the metadata FAIR (R1.3). More metadata elements were updated and improved over time in the different countries.

Five data stewards participated in the training of trainers for six months and were involved in the preparatory training, drafting questionnaires to prepare for FDP implementation, and setting up the ViB in the universities, including engaging in the following:

- Discussions with ministries of health, regional health bureau and university administrations
- Setting up local servers
- Purchasing domain names and configuring the domain names with DNS
- Installation of ViB on the virtual private server (VPS)
- Analysis of the ViB for data production purposes
- Populating COVID-19 data collected form several centres
- Participated in the first SPARQL query test

As a critical first step, the data stewards were engaged in collecting data and making it available in FPDs, while retaining data ownership in residence. Through an internal survey, a practical assignment, and discussions, it was found that many of the data stewards appreciated how easy it is to use the WHO standard COVID-19 eCRF Wizard. Once the COVID-19 data was entered into the eCRF, the eCRF template converted the raw data to RDF and stored it in the triple store as FAIR.

The entire process from training to installation and query lasted four months. This was followed by a period of documentation, assessment and reflection to prepare for the next phase.

6. PROOF OF CONCEPT

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

[①] F1, F2, F3, F4; A1, A1.1, A1.2, A2; I1, I2, I3; R1, R1.1, R1.2, and R1.3 refer to facets of the FAIR Guidelines relating to Findability, Accessibility, Interoperability and Reusability.

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 of queries over the Internet across the FAIR Data Points that visit the machine-actionable data
Step 8. Completion of proof of concept

The VODAN-Africa project began by providing capacity building to African experts from the partner countries through training of trainer workshops with the stakeholders (researchers and traditional data stewards from health facilities). This training produced FAIR data stewards who are able to produce machine-actionable COVID-19 data using eCRF, publish metadata on the FDPs, and query multiple triple stores across FDPs using SPARQL.

The DSW eCRF wizard was implemented as part of the ViB, as shown in Figure 1. The training of trainer workshops and the implementation of the dockerised FDPs, at first, and then scaled up to dockerised ViB [35], were conducted side by side. On 22 July 2020, the first FAIR Data Point for COVID-19 data, called VODAN in a Box, was deployed at Kampala International University by the VODAN-IN Africa [23].

6.2 Procedure for Proof of Concept

For the production of data in the form of machine-readable metadata, a human and machine-readable WHO eCRF was prepared and provided to the partners in each of the locations to produce machine-readable metadata on the FDP. The machine-actionable FDPs were visible and reachable over the Internet, calling home to the VODAN FDP community, meaning that, with the correct permissions, they were findable by algorithms run over the Internet. The already installed FDPs support the automation of the cross-country exchange of observational COVID-19 patient data. This allows hospitals, facilities and researchers in different countries to query FAIR compliant data at the same time, while fully complying with local data protection and governance laws. Mock data was fed into the ViB using the CRF Wizard tool whose semantic model was originally standard to capture data based on the limited WHO eCRF. This was done using SPARQL, which is an RDF query language.

An example query on the data held in the triple store is as follows: 'Present the names of the health care facilities that have had male patients that were admitted with fever (temperature => 38)' [37]. A database SPARQL query was entered into the AllegroGraph query interface, which is also part of the ViB. Within this query, three service call requests were made using the SPARQL endpoints of the three countries (Uganda, Kenya and Netherlands) to obtain and display the facility name against the number of COVID-19 patients. To execute a query, first navigate to https://sparql.kiu.ac.ug:8443/#/repositories/crf/overview and login using 'anonymous', which requires no password. Click to select any of the displayed pre-defined queries, then execute for results or clear to input a query of your choice (https://sparql.kiu.ac.ug:8443/#/repositories/crf/).

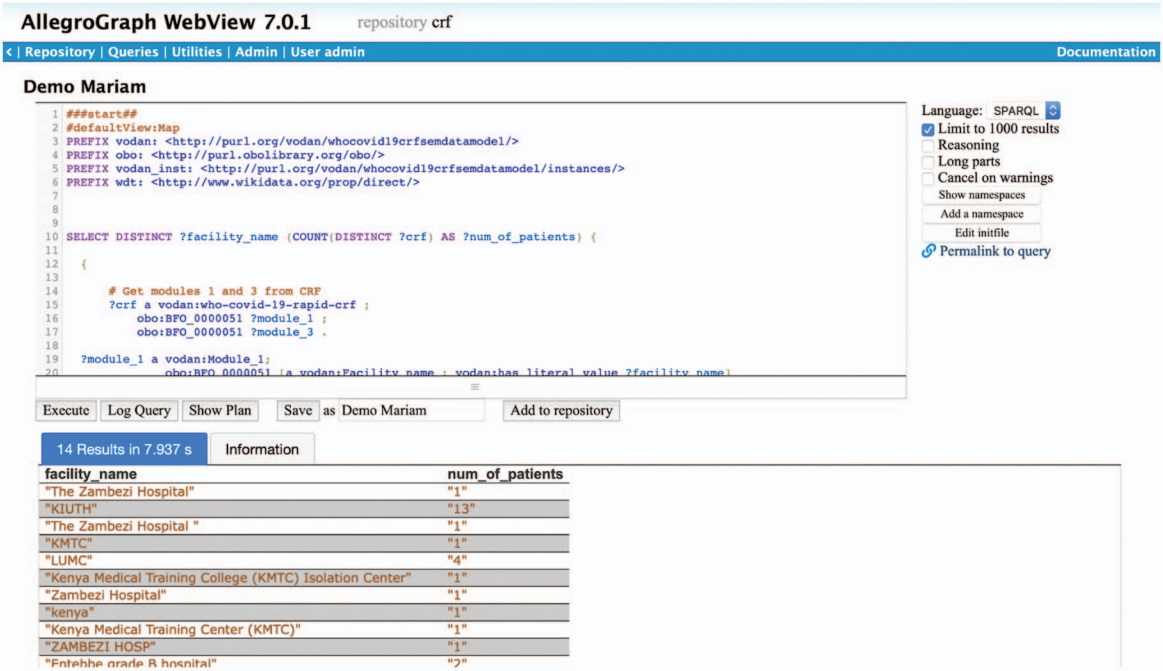

Figure 2. VODAN in a Box proof of concept (Source: Screenshot M. Basajja).

7. SWOT ANALYSIS

The reflection period was used to reflect on the potential of the architecture, but also to identify limitations and alternatives. In order 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.

The strength of the architecture was identified as its potential to organise data ownership and data visiting, and generating 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 was feasible and data-visiting over the Internet in a secure way 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 virtualisation/containerisation, 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/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 in order 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) [38], is based on the design principle that moving computation is cheaper than moving data [39]. In addition to this, further ViB development is planned to allow local customisations and other data types.

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 for 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 with regard to 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, in order to create greater flexibility in semantic language production, 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 the skills for the development of semantic machine-readable data will be obtained by data stewards in Africa [40].

Weaknesses included lack of flexibility in creating new templates and 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 in 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 localisation of user interfaces

and the semantic content and adoption to accommodate languages and culture into the system. Ease of use and flexibility as well as ease of deployment through dockerisation or virtualisation [41] for local deployment, with due consideration for infrastructure (such as bandwidth and storage), are also equally important. The system needs to also be customised for it to seamlessly integrate with FPDs and to be able to add new features beyond eCRFs.

Opportunity was identified, in response to the lack of flexibility in creating templates, in existing tools. The Center for Expanded data Annotation and Retrieval (CEDAR) Workbench [42] was approached as an alternative choice of implementation for data entry. CEDAR provides an easy to use and intuitive template creation feature.

It was further identified that both the CRF Wizard and CEDAR Workbench data entry tools lacked the possibility for importing bulk data, which was reported by a number of health facilities. To overcome the limitation on bulk data import within the CEDAR platform, a custom script was developed by the VODAN team for the two available solutions [43]. The script allows the upload of comma separated bulk 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.

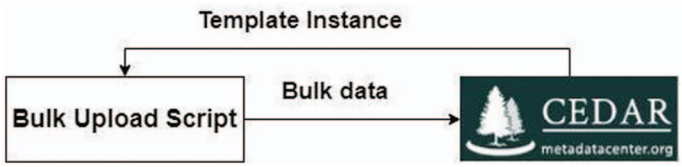

Figure 3. Bulk data upload [43].

Constructing metadata templates and filling in templates to generate high quality metadata seems to be versatile in CEDAR Workbench [43]. 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 localised 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 in premises. The team is currently working on a locally deployable editor for CEDAR.

Further opportunity was identified in the possibility to use and visualise data at point of care through data dashboards. An aggregate dashboard based on data-visiting of the locally-deployed data was also conceivable, strengthening the possibility to monitor the data for early health risks detection.

Threats identified pertained to the lack of ability to integrate requirements of the ministries of health to deploy data in HMISs. The lack of output component to the required systems, would mean that data clerks would have to input the data twice, which would undermine to objective intended: to edit data once for

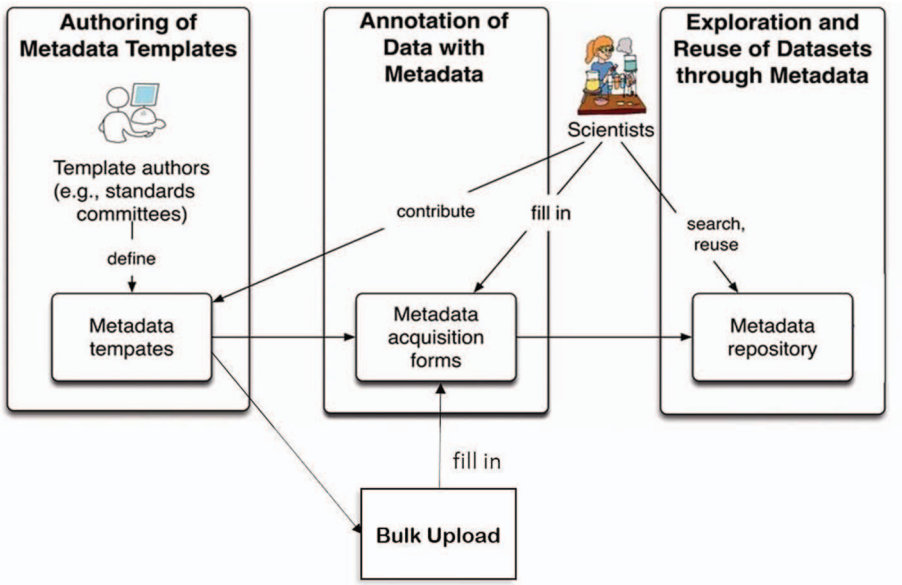

Figure 4. CEDAR ecosystem for metadata management and its communication with bulk upload tool [44].

interoperable and reuse across different needs. The lack of integration with other system was regarded as a serious threat that undermined the purpose of the innovation. It was further identified that training and capacity-building was critical to support the system with knowledgeable data stewards.

8. 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 are able to discover meaningful patterns in data from future epidemic outbreaks [33].

The ViB-FDP was designed to host WHO eCRFs on patients, which makes it possible to query the data while it remains local within the jurisdiction of the hospital or clinic that produces the data. Identifying the weaknesses, the data stewards concluded that the component of the WHO eCRF was limited with regard to different relevant data types, for example, the FAIR metadata production of research data 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 was a setback with regard 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 customised 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 has to be features that facilitate data utilisation, such as analytics tools, dashboards, and other reporting tools and formats.

Pandemics demand a swift response, both in terms of health services and data, for decision making and planning, among other things. The WHO eCRF is an adequate tool for collecting COVID-19 data, although lacking the additional HMIS elements that are demanded by regular health systems. Thus, a need arises to create a tool that can encompass other HMIS elements, which will need to be FAIR. 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 the duplication of effort. Hence, other data production tools should be explored and new integration tools created beyond the ViB, which is dependent on WHO's COVID-19 eCRF. This requires the enhancement, customisation, and localisation 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.

9. CONCLUSION

In the ongoing COVID-19 pandemic, very few data are publicly available in Africa. It is not easy to tell whether these data simply do not exist, or whether they do exist, but with limited access. The establishment of the FDPs, which support different data types, will help address 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 warehouses elsewhere); and enable Africa to be a huge resource of verified data and strengthen data-informed health systems for Africa and the world.

This research considered the deployment of a tool, called ViB, to facilitate the interoperability of COVID-19 clinical patient data held in residence and available as machine-readable semantic metadata. The proof of concept generated a first query, which was implemented across two continents. This is encouraging for the further development of ViB. However, a SWOT analysis identified a number of limitations of the ViB tool. The most critical is the need for flexibility in semantic data production. This could be solved by the localisation of existing platforms, such as CEDAR, with a locally deployable version. It was further identified that the development and deployment of such tools would require increased training of data stewards in semantic data science.

ACKNOWLEDGEMENTS

We would like to thank Misha Stocker for managing and coordinating this Special Issue (Volume 4) and Susan Sellars for copyediting and proofreading. We would also like to acknowledge VODAN-Africa, the Philips Foundation, the Dutch Development Bank FMO, CORDAID, and the GO FAIR Foundation for supporting this research. Finally, we acknowledge the district local government authorities, particularly the district health officers and the data management personnel who availed information that was crucial to the success of this study.

AUTHORS' CONTRIBUTIONS

Mariam Basajja (mariam.basajja@gmail.com, 0000-0001-7710-8843) contributed to the conception and design of the work, carried out the data collection in Uganda, conducted the data analysis and interpretation, contributed to the drafting of the original article, undertook critical revision of the article, and contributed to approval of the final version to be published. Marek Suchanek (marek.suchanek@fit.cvut.cz, 0000-0001-7525-9218) contributed to the drafting of the article, as well as critical revision of the article and approval of the final version to be published. Getu Tadelle Taye (getu.tadele@gmail.com; getu.tadle@mu.edu.et, 0000-0001-5146-6116) contributed to the conception and design of the work, carried out data collection in Ethiopia, conducted data analysis and interpretation, contributed to the drafting of the article, and contributed to approval of the final version to be published. Samson Yohannes Amare (samsonya@gmail.com; samson.yohannes@mu.edu.et, 0000-0002-5425-1126) contributed to the conception and design of the work, carried out data collection in Ethiopia, conducted data analysis and interpretation, contributed to the drafting of the article, and contributed to approval of the final version to be published. Sakinat Folorunso (sakinat.folorunso@oouagoiwoye.edu.ng, 0000-0002-7058-8618) contributed to critical revision of the article and approval of the final version to be published. Ruduan Plug (ruudplug@gmail.com, 0000-0001-5146-6116) contributed to the critical revision of the article and approval of the final version to be published. Francisca Oladipo (francisca.oladipo@kiu.ac.ug, 0000-0003-0584-9145) contributed to the critical revision of the article and the approval of the final version to be published. Mirjam Van Reisen (mirjamvanreisen@gmail.com, 0000-0003-0627-8014) contributed to the conception and design of the work, critical revision of the article and approval of the final version to be published.

CONFLICT OF INTEREST

All of the authors declare that they have no competing interests.

ETHICAL STATEMENT

The research was carried out using the ethical clearance obtained from Kampala International University Research Ethics Committee.

REFERENCES

[1] Freudenthal, E.: Ebola's lost blood: Row over samples flown out of Africa as 'big pharma' set to cash in. The Telegraph (6 February 2019). Available at: https://www.telegraph.co.uk/global-health/science-and-disease/ebolas-lost-blood-row-samples-flown-africa-big-pharma-set-cash/. Accessed 20 May 2021

[2] WHO: WHO coronavirus (COVID-19) dashboard [Online]. World Health Organization (2020). Available at: https://covid19.who.int/. Accessed 20 May 2021

[3] Mons, B.: The VODAN IN: Support of a FAIR-based infrastructure for COVID-19. European Journal of Human Genetics 28(6), 724–727 (2020)

[4] Oladipo, F.: Press release: COVID-19 computer-readable observational data installed at Kampala International University [Online]. Kampala International University, Uganda (22 July 2020). Available at: https://kiu.ac.ug/special-news-page.php?i=covid-19-computer-readable-observational-data-installed-at-kampala-international-university_1595432235. Accessed 20 May 2021

[5] Modjarrad, K., Moorthy, V.S., Millett, P., Gsell, P.-S., Roth, C., Kieny, M.-P.: Developing global norms for sharing data and results during public health emergencies. PLoS Medicine 13(1), e1001935 (2016)

[6] Van Reisen, M: International cooperation in the digital era. Universiteit Leiden, the Netherlands (2017)

[7] Research Data Alliance: Data together—RDA COVID-19 Working Group [Online]. Research Data Alliance (2020). Available at: https://www.rd-alliance.org/groups/rda-covid19. Accessed 20 May 2021

[8] Momtazmanesh, S., et al.: All together to fight COVID-19. American Journal of Tropical Medicine and Hygiene 102(6), 1181–1183 (2020). https://doi.org/10.4269/ajtmh.20-0281

[9] Ozili, P.: COVID-19 in Africa: Socio-economic impact, policy response and opportunities. International Journal of Sociology and Social Policy (2020). https://doi.org/10.1108/IJSSP-05-2020-0171

[10] Chitungo, I., Dzobo, M., Hlongwa, M., Dzinamarira, T.: COVID-19: Unpacking the low number of cases in Africa. Public Health in Practice 1, 100038 (2020). https://doi.org/10.1016/j.puhip.2020.100038

[11] WHO: Building on Ebola response to tackle COVID-19 in DRC [Online]. World Health Organization (25 June 2020). Available at: https://www.afro.who.int/news/building-ebola-response-tackle-covid-19-drc. Accessed 2 March 2021

[12] Owada, K., Eckmanns, T., Kamara, K.-B.O., Olu, O.O.: Epidemiological data management during an outbreak of Ebola virus disease: Key issues and observations from Sierra Leone. Frontiers in Public Health 4 (2016). https://doi.org/10.3389/fpubh.2016.00163

[13] Alliance for Accelerating Excellence in Science in Africa (AESA): Recommendations for data and biospecimen governance in Africa. ASP Policy Paper 3, African Academy of Sciences (AAS) (8 February 2021). Available at: https://www.aasciences.africa/sites/default/files/Publications/Recommendations%20for%20Data%20and%20Biospecimen%20Governance%20in%20Africa.pdf. Accessed 20 May 2021

[14] Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L.O.B., Mons, B., Schultes, E., Roos, M., Thompson, M.: A generic workflow for the data FAIRification process. Data Intelligence 2(1–2), 56–65 (2020). https://doi.org/10.1162/dint_a_00028

[15] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1), 1–9 (2016)

[16] VODAN Africa: About VODAN Africa [Online]. VODAN Africa and Asia (2020). Available at:

[17] Mons, B.: FAIR science for social machines: Let's share metadata knowlets in the Internet of FAIR Data and Services. Data Intelligence 1(1), 22–42 (2019). https://doi.org/10.1162/dint_a_00002

[18] GO FAIR: The Internet of FAIR Data & Services [Online]. GO FAIR (n.d.). Available at: https://www.go-fair.org/resources/internet-fair-data-services/. Accessed 3 March 2021

[19] Wise, J., de Barron, A.G., Splendiani, A., Balali-Mood, B., Vasant, D., Little, E., et al.: Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discovery Today 24(4), 933–938 (2019). https://doi.org/10.1016/j.drudis.2019.01.008
[20] Research Data Alliance: Data together COVID-19 Appeal and Actions. Committee on Data for Science and Technology (CODATA), Research Data Alliance (RDA), World Data System (WDS) and GO FAIR (2020). Available at: https://www.rd-alliance.org/sites/default/files/attachment/Data Together COVID-19 Statement FINAL.pdf. Accessed 21 May 2021
[21] Mons, B.: Data stewardship for open science: Implementing FAIR Principles. CRC Press (2018)
[22] WHO: The Access to COVID-19 Tools (ACT) accelerator [Online]. World Health Organization (2021). Available at: https://www.who.int/initiatives/act-accelerator. Accessed 20 May 2021
[23] GO FAIR: FAIR Principles [Online]. GO FAIR (n.d.). Available at: https://www.go-fair.org/fair-principles. Accessed 20 May 2021
[24] Van Reisen, M., Stokmans, M., Mawere, M., Basajja, M., Ong'ayo, A.O., Nakazibwe, P., Kirkpatrick, C., Chindoza, K.: FAIR practices in Africa. Data Intelligence 2(1–2), 246–256 (2020)
[25] Van Reisen, M., Stokmans, M., Basajja, M., Ong'ayo, A. O, Kirkpatrick, C., Mons, B.: Towards the tipping Point for FAIR implementation. Data Intelligence 2(1–2), 264–275 (2020)
[26] GO FAIR: Data Together [Online]. GO FAIR (n.d.). Available at: https://www.go-fair.org/go-fair-initiative/data-together. Accessed 20 May 2021
[27] Van Reisen, M., Oladipo, F., Stokmans, M., Mpezamihgo, M., Folorunso, S., Schultes, E., et al.: Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research. Advanced Genetics 2(2) (2021). doi: 10.1002/ggn2.10050
[28] Suchánek, M., Basajja, M.: VODAN in a Box: Proof of concept [Online]. Zenodo (27 November 2020). https://doi.org/http://doi.org/10.5281/zenodo.4321626
[29] GitHub: FAIR Data Point specification [Online]. FAIR Data Team, GitHub (2021). Available at: https://github.com/FAIRDataTeam/FAIRDataPoint-Spec. Accessed 2 April 2021
[30] GO FAIR: VODAN in a Box: The all in one solution for easy instalment of VODAN FAIR Data Points [Online]. GO FAIR (16 September 2020). Available at: https://www.gofairfoundation.org/vodan-in-a-box-the-all-in-one-solution-for-easy-instalment-of-vodan-fair-data-points/. Accessed 2 April 2021
[31] AllegroGraph: AllegroGraph: New FedShard Feature [Online]. AllegroGraph (2021). Available at: https://allegrograph.com/products/allegrograph/. Accessed 6 February 2021
[32] Pergl, R., Hooft, R., Suchánek, M., Knaisl, V., Slifka, J.: "Data Stewardship Wizard": A tool bringing together researchers, data stewards, and data experts around data management planning. Data Science Journal 18(1), 1–8 (2019). https://doi.org/10.5334/dsj-2019-059
[33] GO FAIR: Data together. Committee on Data for Science and Technology (CODATA), Research Data Alliance (RDA), World Data System (WDS) and GO FAIR (26 March 2020). Available at: https://www.go-fair.org/wp-content/uploads/2020/03/Data-Together_March-2020.pdf. Accessed 20 May 2021
[34] GitHub: GitHub repository [Online]. GitHub (2021). Available at: https://github.com/VODAN-Tech. Accessed 20 May 2021
[35] VODAN: VODAN in a Box documentation [Online]. VODAN (2020). Available at: https://docs.vodan.fairdatapoint.org/en/latest/. Accessed 20 May 2021
[36] FAIR Data Point: Advanced configuration [Online]. Dutch Techcentre for Life Sciences Revision (2020). Available at: https://fairdatapoint.readthedocs.io/en/latest/deployment/advanced-configuration.html. Accessed 20 May 2021
[37] Collins, S., et al.: Turning FAIR into reality. Final Report and Action Plan from the European Commission Expert Group on FAIR Data, European Commission (2018)

[38] Apache Hadoop: Apache Hadoop 3.2.1—HDFS Architecture [Online]. Apache Hadoop (2020). Available at: http://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html#a.E2.80.9CMoving_Computation_is_Cheaper_than_Moving_Data.E2.80.9D. Accessed 21 May 2021

[39] Sterling, T., Anderson, M., Brodowicz, M.: High performance computing: Modern systems and practices. Morgan Kaufmann, Cambridge, MA (2017)

[40] Oladipo, F., Folorunso, S., Ogundepo, E.A., Osigwe, E., Akindele, A.: Curriculum development for FAIR data stewardship. Data Intelligence 4(4), 991–1012 (2022)

[41] Kasireddy, P.: A beginner-friendly introduction to containers, VMs and Docker [Online]. Medium (4 March 2016). Available at: https://medium.com/free-code-camp/a-beginner-friendly-introduction-to-containers-vms-and-docker-79a9e3e119b. Accessed 7 February 2021

[42] CEDAR: Better metadata means better science [Online]. CEDAR (n.d.). Available at: https://metadatacenter.org/. Accessed 7 February 2021

[43] Gonçalves, R.S., O'Connor, M.J., Martínez-Romero, M., Egyedi, A.L., Willrett, D., Graybeal, J., Musen, M.A.: The CEDAR Workbench: An ontology-assisted environment for authoring metadata that describe scientific experiments. Lecture Notes in Computer Science, 10588 LNCS, pp. 103–110 (2019). https://doi.org/10.1007/978-3-319-68204-4_10

[44] Musen, M.A., et al.: The Center for Expanded Data Annotation and Retrieval. Journal of the American Medical Informatics Association 22(6), 1148–1152 (2015). https://doi.org/10.1093/jamia/ocv048