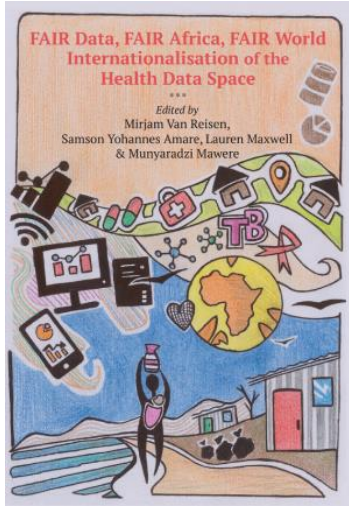
GO TRAIN: A Protocol for Metadata Creation for the FAIRification of Patient Data Health Records

Aliya Aktau, Samson Yohannes Amare, Mirjam van Reisen, Getu Tadele Taye, Tesfit Gebreslassie Gebremeskel, Putu Hadi Purnama Jati & Ruduan Plug

Chapter in:
Fair Data Fair Africa Fair World:
Internationalisation of the Health Data Space



Cite as: Aktau, A., Amare, S. Y., Van Reisen, M., Taye, G. T., Gebremeskel, T. G., Purnama Jati, P. H., & Plug, R. (2025). GO TRAIN: A protocol for metadata creation for the FAIRification of patient data health records. https://doi.org/10.5281/zenodo.15382836. In Van Reisen, M., Amare, S. Y., Maxwell, L. & Mawere, M. (Eds.), *FAIR data, FAIR Africa, FAIR world: Internationalisation of the Health Data Space.* (pp. 97–142). Bamenda: Langaa. URL: https://www.researchgate.net/publication/391750151_FAIR_Data_FAIR_Africa_FAIR_World_The_Internationalisation_of_the_Health_Data_Space

The About the Authors note can be found here: https://raee.eu/wp-content/uploads/2025/05/About-the-Authors-and-Editors.pdf
The list of figures and tables can be found here: https://raee.eu/wp-content/uploads/2025/05/List-of-Figures-and-Tables.pdf

Contents

Foreword…….....xi

Mouhammad Mpezamihigo

Acronyms……..……...xv

Chapter 1: FAIR Data, FAIR Africa, FAIR World: The Internationalisation of the Health Data Space…….…...1

Mirjam van Reisen, Samson Yohannes Amare, Lauren Maxwell & Munyaradzi Mawere

Chapter 2: Bridging Borders with FAIR Data: Transforming Digital Ecosystems for Maternal Health and Public Health Surveillance in Africa……25

Samson Yohannes Amare

Chapter 3: Introducing Data Sovereignty Over Patient Data: Patient Data Ownership in Residence of Health Facilities in Kenya……65

Reginald Nalugala, Putu Hadi Purnama Jati, Samson Yohannes Amare, Maxwell Omare, Jacinta Wairimu, Charles Kahiro, William Nandwa, Seth Okeyo, Dennis Kinoti, Aliya Aktau, Albert Mulingwa & Mirjam van Reisen

Chapter 4: GO TRAIN: A Protocol for Metadata Creation for the FAIRification of Patient Data Health Records……...97

Aliya Aktau, Samson Yohannes Amare, Mirjam van Reisen, Getu Tadele Taye, Tesfit Gebreslassie Gebremeskel, Putu Hadi Purnama Jati & Ruduan Plug

Chapter 5: De Novo FAIRification: A Literature Review……..…...143

Zhengyu Lin

Chapter 6: Federating Tools for FAIR Patient Data: Strengthening Maternal Health and Infectious Disease Surveillance from Clinics to Global Systems……..……..157

Samson Yohannes Amare, Getu Tadele Taye, Ruduan Plug, Araya Abrha Medhanyie & Mirjam van Reisen

Chapter 7: Adoption of FAIR-OLR Architectures to Support Insights from Patient Health Data Records in Africa……195

Putu Hadi Purnama Jati, Samson Yohannes Amare, Abdullahi Abubakar Kawu, William Nandwa, Getu Tadele Taye & Mirjam van Reisen

Chapter 8: A Critical Incident Assessment of a FAIR Implementation Study on Patient Health Data in Africa……..…241

Abdullahi Abubakar Kawu, Putu Hadi Purnama Jati, Lars Schrijver, Mirjam van Reisen, Getu Tadele Taye, Dympna O'Sullivan & Lucy Hederman

Chapter 9: The Potential of Adoption of FAIR Guidelines in Digital Healthcare in Kazakhstan……..…277

Aliya Aktau

Chapter 10: Harmonising Antenatal Care Records Across Four African Countries: A Comparative Analysis and Development of a Standardised Data Model……...301

Samson Yohannes Amare, Liya Mamo Weldu & Tesfit Gebremeskel

Chapter 11: Complexities of Recording and Reporting Perinatal Data: A Case Study from Namibia…….…….325

Beatrix Callard

Chapter 12: Enhancing Data Privacy and Availability in Indonesia's SATUSEHAT Platform through FAIR-OLR Principles…….……361

Putu Hadi Purnama Jati, Markus Sintong Tambah Lasroha & Mirjam van Reisen

Chapter 13: Creation of a FAIR Data Point for a Clinical Trial: The CoHSI2 Dataset..……...…….…397

Aliya Aktau & Mirjam van Reisen

Chapter 14: Implementation of De Novo FAIRification in Relational Legacy Systems: The Case of the Electronic Medical Record System for Maternal Health in Afya.ke…… ……...429

Zhengyu Lin

Chapter 15: Narratives in Public Agenda-Setting for FAIR Data and Health Data Management in Africa: Enhancing Maternal Health and Infectious Disease Outcomes..……..467

Putu Hadi Purnama Jati & Mirjam van Reisen

Chapter 16: Testing the Cross-Border Africa Health Data Space: Monitoring, Tracking and Prevention of Mother to Child Syphilis Transmission……..517

Samson Yohannes Amare, Liya Mamo Weldu, Araya Abrha Medhanyie & Mirjam van Reisen

Chapter 17: GO FAIR: Ontology Development of Health Semantics with Cultural Specificity: Traditional Health Practices using Tsebel in Conflict Zones..……...553

Bereket Godifay Kahsay, Mirjam van Reisen & Zhengyu Lin

Chapter 18: Satu Data Indonesia and FAIR Data: Advancing Coherent Data Management in Government Administration.……585

Intan K. Utami & Mirjam van Reisen

Chapter 19: Integrating the Personal Health Train Methodology into Healthcare Systems for Enhanced Elderly Care in the Netherlands…….625

Ria Landa-Figueroa & Lars Schrijver

Chapter 20: FAIR Data Implementation for Analysis of Research Data in Human Trafficking and Migration……..…….665

Kai Smits, Mehul Upase, Nimish Pandey, Senjuti Bala & Mirjam van Reisen

Chapter 21: A Higher Education Curriculum for Cultural Competence, Representation and Social Responsibility in AI and FAIR Data Practices……...685

Sakinat Folorunso, Francisca Oladipo, Mirjam van Reisen & Ibrahim Abdullahi

About the Editors……...721

Chapter 4

GO TRAIN:
A Protocol for Metadata Creation for the FAIRification of Patient Data Health Records

Aliya Aktau, Samson Yohannes Amare, Mirjam van Reisen, Getu Tadele Taye, Tesfit Gebreslassie Gebremeskel, Putu Hadi Purnama Jati & Ruduan Plug

Abstract

This research investigates the components required to create FAIR-compliant data architectures for patient health information systems. It highlights the interplay between terminology, ontologies, vocabularies and semantic data, which collectively enable structured and interoperable data environments. Among the ten assessed requirements for achieving FAIR data, eight were deemed both feasible and relevant, including FAIR patient data production, multi-functional access and control services, data visualisation, and enhancements in FAIRness through repositories and bulk uploads. DHIS2 compliance and CEDAR workbench federation were found unsuitable due to limited data quality and software usability challenges. The GO TRAIN pillar is pivotal in operationalising the requirements by equipping stakeholders with essential skills through capacity building, technical training, and knowledge dissemination. GO TRAIN ensures the implementation of FAIR-compliant systems by focusing on access-control processes, decision-making tools, and advanced data visualisation techniques. Training led to a culture of continuous improvement, embedding FAIR principles into organisational workflows.

Keywords: FAIR guidelines, health information systems, FAIR data stewardship training, electronic health records, CEDAR, DHIS2

Introduction

Operationalisation of data curation (Wilkinson et al., 2016) as Findable, Accessible (under well-defined conditions), Interoperable, and Reusable (FAIR) is a non-trivial but important problem to resolve the need to compute large volumes of data for Machine Learning (ML) and Artificial Intelligence (AI). However, Strawn (2021) observes that FAIR guidelines provide a policy prescription and not an implementation. Stocker, et al. (2022) point out that there has been a remarkably fast and widespread policy uptake on the FAIR proposition in the European Union (EU), pertaining to the curation of scientific data as an asset of the EU investments in research. The digital data assets represent value. Stocker et al. (2022) also find that the operational practices of FAIR data curation stayed behind the political recognition of the relevance of it. While the policy that was put in place, to require recipients of research grants to deploy their data as FAIR in repositories where the data would be findable, was put in effect in 2020, the understanding of how to carry out the curation in a practical protocol was not available (Stocker et al., 2022).

While Jacobsen et al. (2020) discuss a theoretical protocol which serves as a data curation plan, they also do not propose an operational set of steps for the process of the creation of FAIR data as semantic machine-actionable instances. Semantic data refers to data that is enriched with meaning, using ontologies and vocabularies. It adheres to principles that allow machines to understand and interpret the data meaningfully. Semantic data enables data integration, querying, and reasoning across heterogeneous datasets. Semantic data uses Linked Data, the technique of connecting pieces of data through shared semantics, enabling integration and discovery represented in a. Resource Description Framework (RDF) which is a standard model for representing semantic data. While the semantic web is not new, the FAIR guidelines are inherently Federated and Federated AI Ready, which adds a new layer of complexity to semantic web solutions (Taye et al., 2024).

The reference protocol for this process of FAIR data curation, can be referred as the FAIRification process (Jacobsen et al., 2020). The problem of the lack of operational clarity on the FAIRification process was particularly pressing in efforts that were seeking to

expand adoption of FAIR-data in new regions, hitherto being left behind (Lin et al., 2022; Van Reisen, Stokmans, Basajja et al., 2020), suppressing the ambition underlying the idea of FAIR guidelines, that it would help connect data globally. The Go FAIR Foundation proposed three actions for obtaining global uptake and progress, namely GO BUILD, GO CHANGE and GO TRAIN, progress in all three domains depend on clarity about the operational FAIRification steps that provided the basic building blocks for building, training and changing FAIR-based architectures (Taye et al., 2024).

Looking at the scenario of potential building blocks, the lack of available scenarios for concretising the FAIRification process was even more pressing in granular data curation in FAIR format when such data was held in federated containers, included personal and sensitive data, and was subject to different regulatory frameworks relevant to different jurisdictions of provenance of the data (Van Reisen et al., 2022). A proof of concept was carried out in 2020 to demonstrate information retrieval from separately held data containers in hospitals located in two different continents through SPARQL (Basajja, Van Reisen & Oladipo, 2022b). The results showed the need for a more rigorous set of specifications and requirements for the FAIRification process (Van Reisen et al., 2022). This was particularly needed to concretise the skills for data stewards to be trained to carry out the FAIRification process (Oladipo et al., 2022) in line with the GO TRAIN workflow, that is necessary to accompany GO BUILD and GO CHANGE to stimulate the expansion of FAIR data uptake (Van Reisen, Stokmans, Mawere et al., 2020).

Wilkinson et al. (2016) recognise the FAIR-guidelines as community specific, and Mons (2018) builds on this by emphasising the need for community specific decision-making in the FAIRification process. Following Jacobsen et al. (2020), Van Reisen et al. (2022) acknowledge the community specific experiences of Value-driven Ownership of Data and Accessibility Network (VODAN) and document these as an example of how decisions are made, adapted and assessed within a particular context of purpose of a research group. They point to the relevance of situational awareness and

recognition of design choices and dilemmas, relating to the academic or societal challenges that should be solved.

The 15 facets of a mature FAIR architecture create the digital objects which are FAIRified as semantically enriched instances which are readable by humans and machines (Mons, 2018). The VODAN research group has attempted to establish FAIR datasets in health facilities, containing patient records linked to an ontology for disease classifications, enabling automated reasoning about health trends. This chapter describes the results on establishing a protocol of steps for FAIR metadata creation. The research question is: How are the GO TRAIN, GO BUILD and GO CHANGE modes of the FAIRification process represented in the design of a De Novo FAIRification workflow for data handling by the VODAN research community?

Methodology

The objective of the study focuses on introducing FAIRification in a FAIR agnostic situation. In this case, the success of the FAIRification relies on instructions that clarify the steps that translate the FAIR principles into a concrete set of operational actions that clarify what the FAIRification process is, what it results in and why it would be carried out. The purpose of this study was to establish a protocol for the data FAIRification process.

Location of the study

The study was carried out in Africa, in Uganda, Nigeria, Kenya, Ethiopia, Zimbabwe and Tanzania. Somalia was later added, but not part of this study report. The study was implemented in the Research group of the Virus Outbreak Data Network (VODAN), which was in 2024 rebranded as the Value and Ownership of Data Access network (VODAN). The VODAN research network was established in 2020 as a response to the lack of data from the African region during the Covid-19 pandemic (Van Reisen et al., 2022). VODAN is a network of researchers and research stakeholders, investigating the global application of FAIR data, with a particular focus on implementation in non-Western regions.

The study was undertaken with a view to using the FAIRification process for the creation of interoperability and reusability in patient data, which were monitored for the surveillance of the spread of the novel coronavirus (COVID-19) pandemic. This was important because COVID-19 data from Africa was missing, even though patient data existed. However, using and sharing this data was challenging due to its sensitive nature as personal and health information. This environment provided a good case for studying the deployability of FAIR data as secure and privacy preserving digital assets that could render relevant information to help monitor the COVID-19 pandemic in Africa. The overall research that was carried out on the potential to analyse patient data federated across health facilities and reposited in containers in health facilities, was carried out by VODAN. The design of a FAIR-data architecture of VODAN was set in a real-life natural environment.

Study design

The study is an ethnographic case study design study. An ethnographic design study is useful in case of a wicked problem (Buchanan, 1992), which is a problem that is new, has not been researched and needs thorough exploration. Sanders (2008) provides a scheme of different types of design studies; this study can be seen as a participatory design study, with expertise provided by the lead researchers. The participation of key stakeholders was important to steer priorities of the design to make sure the design would fit the use of it, as intended by the stakeholders.

Reasoning for the focus on patient data for FAIRification

The study took place as stand-alone research in the context of the research by VODAN. The decision to focus on patient data as the foundation for the FAIRification process in this study is based on the combination of its academic and societal relevance.

Patient data provides a compelling academic case for FAIRification due to its intrinsic characteristics. It represents routine data that is inherently well-structured, with a widely shared understanding of its semantics and schema across diverse geographic and cultural contexts. Unlike many other datasets, patient data exhibits minimal

interpretative variability, making it a highly suitable candidate for testing the application of FAIR principles in diverse settings.

Furthermore, the societal relevance of this focus cannot be overstated. Patient data from resource-constrained regions in Africa and Asia has historically been underrepresented in global health surveillance efforts (Van Reisen et al., 2021). This lack of inclusion represents a significant gap, particularly during critical periods such as the COVID-19 pandemic, when comprehensive and equitable data sharing is crucial for effective response strategies. This is particularly the case for pregnant women and the maternal health domain. The urgency to address this gap provided a unique and timely impetus for attempting FAIRification in these contexts.

The sense of urgency of the need of FAIRification in health was compounded by the fact that the East Africa Health Commission has already adopted FAIR as the guidelines of choice in 2017 (EAHRC, 2017), for the creation of an interoperable health information system (Van Reisen et al., 2021) whereas the study of the equivalence of FAIR guidelines and digital health policies in countries (Uganda, Kenya, Zimbabwe, Nigeria, Kazakhstan, Indonesia) has all showed positive trends to adoption of FAIR-relevant principles (Van Reisen et al., 2022) during research carried out in 2019/2020. The lack of interoperability and the proliferation of competing vertical digital health pipelines; the silos of health data, the unavailability of (quality) health data in Ministries of Health and in health facilities at point of care, all contributed to a sense that FAIR interoperable and reusable workflows would help integrate the patient data in the health system and that this was overdue and much needed (Van Reisen et al., 2021).

The global COVID-19 pandemic further underscored the critical need for interoperable and reusable patient data to inform surveillance, policy, and intervention strategies for prevention and treatment. FAIRification in health during the pandemic was expected to provide a stronger engagement from researchers and stakeholders operating within the ecosystem. It was expected that the societal relevance would help in motivating them to engage with and support the challenging and innovative process of FAIRification. The urgency of the pandemic created a conducive environment for experimenting with the development and operationalisation of a FAIRification

workflow, particularly in FAIR-agnostic situations where established practices or tools were absent.

Timeline of the study

The study was carried out from 2020 to 2022. The study was carried out in the following steps:

1. Exploration of the problem
2. Design of a first protocol
3. Evaluation
4. Agreement on a set of principles and requirements
5. Design of a new protocol
6. Training of the protocol
7. Establishment of a FAIR data-curated structure
8. Test of the FAIR data production
9. Evaluation of the FAIR data production and
10. Deduction of key principles of the final protocol

All these phases were carried out within a collaborative assessment of the wider research group. The first author was the lead researcher on ontology creation and was in charge of integrating interoperability in the ten steps. The second first author was the technical lead to investigate how the interoperable system could be deployed in a federated fashion. The third author was the PI of the research programme and was responsible for ensuring the alignment of the decisions regarding the data-curation in relation to the overall objectives of the research program. Taye was responsible for translating the research into a training manual (2024). All other authors contributed to the decisions regarding the design of an operational implementation strategy.

Input by experts

To create a workflow for VODAN, on the operational steps to create FAIR data and metadata, expertise was obtained from groups of experts available in 2020. The following groups were identified as relevant as a resource on data and metadata creation.

GO FAIR Foundation

The GO FAIR initiative was launched in 2016, and the GO FAIR Foundation was officially established in 2018. The GO FAIR Foundation was created to accelerate the implementation of the

FAIR principles by fostering collaboration and providing practical tools, standards, and frameworks for FAIR data stewardship. It acts as a global movement to support the realisation of the European Open Science Cloud (EOSC) and other open science infrastructures. Its Head Quarter is established in Leiden and associated with Leiden University and Leiden University Medical Centre (LUMC).

Data Stewardship Wizard (DSW)

The Data Stewardship Wizard (DSW) was developed starting in 2017 and was made publicly available in 2018. DSW was created to support researchers and data stewards in creating data management plans (DMPs) and implementing FAIR data practices. It provides a guided, interactive platform for users to ensure their data management aligns with FAIR principles. The DSW is primarily developed and maintained in the Czech Republic. DSW originated as a collaborative project between the Czech National Infrastructure for Biological Data (ELIXIR CZ) Czech Republic node and other partners, including the University of West Bohemia (UWB) and the ELIXIR CZ. The development team is based at the University of West Bohemia in Plzeň.

CEDAR (Center for Expanded Data Annotation and Retrieval)

The CEDAR project was launched in 2014 with funding from the U.S. National Institutes of Health (NIH). CEDAR was created to enhance the creation and retrieval of richly annotated metadata. The goal was to improve metadata quality and accessibility, particularly for biomedical research, by offering tools for designing and managing structured metadata templates. CEDAR was established in 2014 to create a computational ecosystem for development, evaluation, use, and refinement of biomedical metadata. CEDAR is based in Stanford University.

Comparing Go FAIR Foundation, DSW and CEDAR

GO FAIR, DSW and CEDAR provided the expert resources that served as input for the creation of an operational workflow for FAIRification in VODAN. Together, they contributed to the initial ecosystem required for FAIR data stewardship, focusing on different but complementary aspects: policy and advocacy (GO FAIR),

planning and guidance (DSW), and metadata creation and management (CEDAR).

The shared common element of the technical resources is their focus on implementing and promoting the FAIR principles:

- GO FAIR Foundation provided advocacy, guidelines, and support to integrate FAIR principles globally.
- DSW facilitated the creation of FAIR-aligned data management plans, enabling researchers to make their data FAIR.
- CEDAR ensures that metadata, a crucial component of FAIR data, is rich, standardised, and reusable.

Table 1. Differences between GO FAIR Foundation, DSW and CEDAR

Feature	GO FAIR Foundation	DSW	CEDAR
Nature	Global initiative and foundation	Interactive tool/platform	Research project and software platform
Primary goal	Advocacy and support for FAIR data	Supporting data management planning and FAIRification	Metadata creation and retrieval
Focus	Broad FAIR implementation and education	Data management planning	Metadata quality for biomedical data
Audience	Policymakers, organisations, researchers	Researchers and data stewards	Biomedical researchers and curators
Funding origin	European and international funding	EU and local academic funding	NIH (U.S.-based funding)

The contributions from the three resources were integrated in a chronological sequence, starting with general exploration provided by the GO FAIR Foundation, progressing to the development of a DMP, and culminating in the creation of FAIRified data and a FAIR

ecosystem for data generation by VODAN in CEDAR. Each organisation contributed experts who guided the research team. These experts supported the ten steps of the research process, with their input aligned to the timeline and delineated according to the respective organisation's role.

GO FAIR Foundation

1. Exploration of the problem

DSW

2. Design of a first protocol
3. Evaluation
4. Agreement on a set of principles and requirements

CEDAR

5. Design of a new protocol
6. Training of the protocol
7. Establishment of a FAIR data-curated structure
8. Test of the FAIR data production
9. Evaluation of the FAIR data production and
10. Deduction of key principles of the final protocol

The following timeline is associated with the support provided by the resource experts:

2020 - Training Metadata for Machines. The instructor was Erik Schultes, lead FAIR Implementation and trainer for the GO FAIR foundation, assigned to VODAN. The Metadata for Machines training with VODAN was run in 2020 and was registered and posted on YouTube. The function of this training was to demonstrate the possibility of FAIR guidelines for human and machine operations.

2020 - Material was provided by the experts of the Data Stewardship Wizard (DSW) for metadata creation, in the second half of 2020, which resulted in a Proof of Concept of metadata retrieval across the Netherlands (Europe) and Uganda (Africa), demonstrating the possibility of federated data retrieval based on an electronic Case Report Form (eCRF). The Proof of Concept was evaluated, and a detailed set of Requirements and Specifications were discussed in VODAN and approved (VODAN, 2021).

2021 - Material was provided by the CEDAR, which was selected for expert input, based on the Requirements and Specifications developed at the end of 2020 (VODAN, 2021). Throughout 2021 trainers of CEDAR worked with the first author to develop a workflow that would correspond with the Requirements and Specifications (VODAN, 2021). During the 2021 provision of expertise, data stewards in VODAN were trained by the first author, with support of CEDAR, to test whether the workflow could be integrated within the work of the data stewards.

Data collection and analysis

The data used for this research is the elements of work that lay down discussions, considerations, decisions, workflows, tools, assessments and evaluations carried out during the ten steps. These included video recordings, notes, formal documentation, workflows, and write ups of assessments carried out during the whole of the design process.

The researchers created a detailed archive of the material collected and developed during the study. All the material was archived in digital folders. Detailed material was archived by the first researcher. Assets shared in the VODAN research group, were archived in the VODAN community folder in Open Science Framework (OSF). Some of the material was also stored in publicly available format, such as on YouTube. In preparation of the analysis carried out for this study, the material in the digital folders was revisited and retrieved to generate a chronological order. The analysis of the data focuses on the chronology of the FAIRification process developed in the 2020-2022 period. The material was labelled based on the addition that the material demonstrates in function of the FAIRification process that was designed. In a subsequent step, the key materials were selected for discussion in this study. The materials are discussed in chronological order.

Purpose of the analysis

This analysis aims to systematically document the development of a FAIRification process in the absence of established operational steps and tools for FAIRification, tailored to the agnostic context of the VODAN researchers at the start. The research undertaken followed a dynamic and iterative methodology. It is acknowledged that this

documentation does not claim the operational workflow utilised was optimal, nor does it remain the optimal approach given the emergence of new tools and workflows since its implementation. Instead, the purpose of this documentation is to capture the initial phase of FAIRification, contributing to the foundational efforts of engineering a FAIR-based ecosystem across Africa and Asia.

Theoretical framework: FAIRification of data

The FAIRification process is the operational workflow of realising the 15 facets of the FAIR guiding principles, which ensure that data is Findable, Accessible, Interoperable, and Reusable. The FAIRification process relies on the creation of metadata that provides meaning to the data (Mons, 2018). These principles or facets are organised into four categories, and certain elements are particularly relevant to the FAIRification process, as indicated in table 2.

Table 2. Categories of FAIR principles

Principle	Description	Type
Findable		
F1	Data and metadata are assigned a globally unique and persistent identifier (e.g., DOI, UUID).	Relevant
F2	Data is described with rich metadata.	Critical
F3	Metadata clearly and explicitly include the identifier of the data they describe.	Critical
F4	Data and metadata are registered or indexed in a searchable resource.	Critical
Accessible		
A1	Data and metadata are retrievable by their identifier using a standardised communications protocol. A1.1: The protocol is open, free, and universally implementable.	Critical
	A1.2: The protocol allows for authentication and authorisation where necessary.	Critical
A2	Metadata is accessible, even when the data is no longer available.	Critical
Interoperable		
I1	Data and metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.	Critical
I2	Data and metadata use vocabularies that follow FAIR principles.	Critical

Principle	Description	Type
I3	Data and metadata include qualified references to other (meta)data.	Critical
Reusable		
R1	Data and metadata are richly described with accurate and relevant attributes.	Relevant
R2	Metadata is released with a clear and accessible data usage license.	Relevant
R3	Metadata is associated with detailed provenance.	Critical
R4	Metadata meet domain-relevant community standards.	Critical

FAIRification focuses heavily on metadata quality, accessibility protocols, and adherence to community standards (I2, R1.3) to ensure data usability across domains. The metadata creation is the critical element defined in the FAIRification workflow by Jacobsen et al. (2020), which focuses on the semantic metadata creation as the critical step to creating interoperability and reusability.

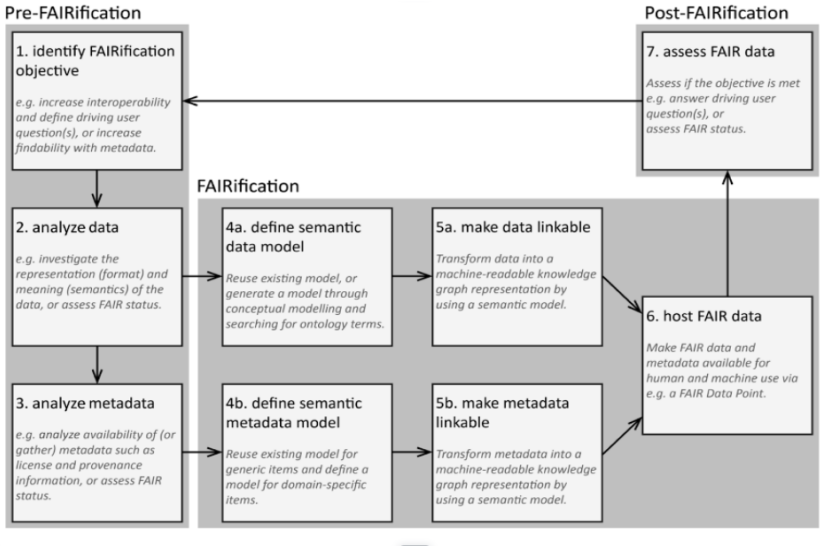


Figure 1. FAIRification process

Source: Jacobsen et al., 2020

Jacobsen et al. (2020) provide a foundational academic framework outlining a structured methodology for transforming existing datasets into FAIRified resources. The authors emphasise the iterative and context-dependent nature of FAIRification. They highlight that while the principles are universal, their implementation often requires

tailored approaches for specific data types, domains, and infrastructures.

The proposition by Groenen et al. (2020) includes the notion of De Novo FAIRification. This can be defined as the process of creating digital data assets as FAIR from their point of creation. De Novo FAIRification provides the strongest relationship between a digital asset and its provenance, since metadata of the data asset is added at the point of creation. This enhances the tracing to the origin of the digital asset, since provenance information is maintained, and this helps the trustworthiness of the data (and the ability to investigate the trustworthiness). Another term that can be used for De Novo FAIRification is 'FAIR by Design'.

The following definitions are introduced by Plug et al. (2022) for ingredients of the FAIRification process:

Terminology: offers a clear set of domain-specific terms to standardise communication among healthcare providers and researchers.

Vocabulary: provides consistent labels for annotating data, such as symptom names or test results.

Ontology: used to define and model the relationships between healthcare entities like 'Patient', 'Disease', and 'Treatment'.

Semantic data: integrates data as triples following the object-predicate-subject structure that is machine-readable and interoperable across systems, leveraging linked data principles.

Table 3 compares the tools to help define data instances, and their function in a FAIR environment.

Table 3. Comparing terminology, vocabulary, ontology and semantic data

Feature	Terminology	Vocabulary	Ontology	Semantic data
Structure	List of terms and definitions	Informal, controlled list	Formal, logical	Data enriched with semantics

Feature	Terminology	Vocabulary	Ontology	Semantic data
Purpose	Ensure linguistic precision	Standardise annotation	Enable reasoning	Integrate and interpret data
Complexity	Low	Moderate	High	Varies
Example	Medical terminology guide	Symptom vocabulary	OWL-based health ontology	RDF-linked patient records
Role in FAIR	Accessibility, Reusability	Findability, Accessibility	Interoperability, Reusability	Interoperability

Jacobsen et al. (2020) stress the importance of collaboration between domain experts, data stewards, and technical teams to overcome challenges related to semantic alignment, metadata quality, and technical interoperability. The article serves as a guide for researchers and organisations aiming to adopt FAIR principles, providing both theoretical insights and practical steps for operationalising FAIR data practices.

Findings

The findings with regards to the data-handling in VODAN community is discussed in six themes: (i) the initial input received on the FAIRification process by Metadata for Machines (M4M) workshops, leading to a Proof of Concept; (ii) the evaluation of the Proof of Concept and articulation of new requirements for the data-handling process; (iii) the articulation of a reference architecture, as the abstract plan for data-handling, dictating both the data-handling process and the requirements for underlying supporting soft- and hardware; (iv) the process of creation of tools for data-handling – the creation of a data-handling set up of a De Novo FAIR protocol through CEDAR; (v) training for data-handling; (vi) assessment of the tested data-handling protocol.

The findings reflect elements of GO TRAIN (i and v), GO BUILD (iii and iv) and GO CHANGE (ii and vi) (Van Reisen, Stokmans, Mawere et al., 2020).

(i) FAIRification for data federation: Proof of concept

The first step of the VODAN FAIRification process was to engage with the Metadata for Machines (M4M) workshops which relates to the GO TRAIN track. The purpose of the workshops is to introduce participants to the ideas underlying the creation and enhancement of semantic metadata that is specifically designed to be machine-readable and machine-actionable, aligning with the FAIR principles.

The M4M workshops helped the VODAN-team to understand how metadata could be processed, interpreted, and reused by computational systems without direct human intervention. The M4M workshops serve to bridge the gap between data producers and technical experts. The M4M workshops promoted collaboration among the diverse participants in the VODAN community. This was important as it ensured that metadata not only meets domain-specific standards but also adheres to broader, community-endorsed ontologies and semantic frameworks.

The M4M workshops advanced the idea of how development of FAIR-compliant data ecosystems could be operationalised and achieved in concrete terms. The M4M workshops did not by themselves deliver any concrete operational steps. They served to build confidence that operational translation of the FAIR principles was achievable. It also served to create understanding that data stewards or data science experts alone could not achieve a meaningful FAIRification process. It generated a realisation that all stakeholders in the community were important to achieve a common understanding of the process, the ecosystem and the interpretation of what would be meaningful FAIR assets, relating to the goals set by the community. The M4M's contribution was the creation of a common understanding on how machine-actionable metadata reflects a key priority of the FAIR paradigm: empowering computational systems to automate data discovery, retrieval, and analysis, and creating a common understanding of the potential to create FAIR data that is held in the location of provenance, in federated format and owned by the creators of the digital assets.

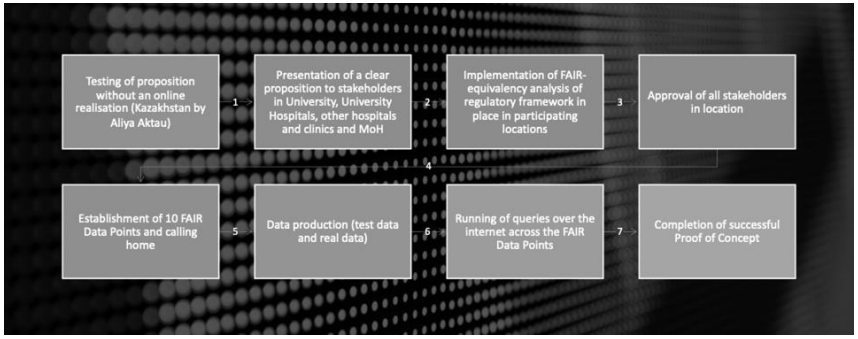


Figure 2. Steps towards Proof of Concept of patient data in federated data-stores

Source: Van Reisen & Oladipo, 2020

The M4M workshops led to the successful Proof of Concept of a cross-continental querying of data between Leiden (the Netherlands) and Kampala (Uganda) with ten simple FAIR Data Points (Van Reisen & Oladipo, 2020).

Design of a first protocol to produce FAIR granular patient data

The first protocol developed in collaboration with DSW has been described by Basajja, Suchanek et al. (2022). The objective of the exercise was to produce patient data on COVID-19 diagnoses and treatment. DSW produced the vocabulary based on FAIR-guidelines, with unique identifiers, standardised ontologies and linked data triples, in a format for COVID-19 registrations provided by the World Health Organization (WHO) eCRF. This process of creating data as new instances through metadata at the source is referred to as De Novo FAIRification (Groenen et al., 2020) or FAIRification by Design.

The use of the form for data registration proved functional; the protocol demonstrated that it was possible to produce granular patient data that remained in the place where it was produced, with a multipronged functionality (the data could be used for different use cases), with queries run over the data through a data-visiting procedure. Plug et al. (2022) defined the terminology describing the underlying process. The process was accompanied by a detailed investigation of the regulatory frameworks to ensure that the

procedure aligned with the policy provisions available in the different countries where data production was carried out (Van Reisen et al., 2022). The DSW COVID-19 eCRF was used to demonstrate the possibility of exchange of data through data-visiting in federated format. The assessment of the Proof of Concept relates to the GO CHANGE track.

(ii) Evaluation: Lacking ownership over data handling

The production of data through the DSW produced form was evaluated at the end of 2020 by the VODAN research group. While the DSW form with its FAIR-based back end was technically correct and created according to required parameters, the team identified several shortcomings of the procedure that had been followed.

In a series of weekly online meetings of the research team, the procedure followed with the assistance of DSW was assessed (November 2020 – February 2021). The evaluation was pertinent as a problematic situation had arisen, demonstrating the inflexibility of the eCRF based data production. A team of researchers in Tunisia, working with out of clinic patients (mostly migrants) to investigate the prevalence of COVID-19, was unable to record its data in the eCRF format. The data had different characteristics, which could not be captured in the form. This led to a range of fundamental questions being discussed on the interoperability of the tools being proposed through the procedure that had been followed.

The meetings to assess the procedure ended in a consensus within the research group on the shortcomings of the process that had been followed. The points that were raised in the meetings can be summarised as follows:

- Lack of flexibility

The form was pre-made and therefore was not adaptable to variations in information to be recorded in health facilities

- Lack of clear basis for standardisation of vocabularies

It was unclear how the vocabularies were standardised with internationally recognised thesauri and terminology services and common data models such as for instance Open Biomedical

Ontologies, Observational Medical Outcomes Partnership Common Data Model (OMOP CDM), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), International Statistical Classification of Diseases and Related Health Problems 10th Revision (IDC10/11) and how vocabularies were to be maintained and updated.

- Lack of training of data stewards

The service came with a package to engage with the data-stewards, but did not include transfer of knowledge. The data stewards associated with the process lacked insight in the procedure and could not handle the preparation for data production tools independently using DSW. Consequently, the possibilities to maintain, change or expand the use of a FAIR data infrastructure for patient data handling was extremely limited, capping the research in the COVID-19 domain, based on the prepared e-form i.e., the WHO eCRF.

- Lack of adaptability of the supporting tools

The FAIR based COVID e-form was not supported by software that allowed the e-form to run integrated in other digital tools, which was undermining the overall purpose to investigate data integration through FAIR data models. The information to allow adaptation was not available among the community that should use, adopt, introduce and adapt the tools.

- Lack of ownership over vocabulary creation, tools or purposes

Due to the dependency on DSW service provision, the creation of the FAIR-based COVID e-form did not enhance the capability of expansion of vocabulary creation, of FAIR tools and enhancement of purposes for data integration, given that the data stewards did not own the skills or the tools of the creation and production process.

- Creation of trust in data production and data handling

It became clear that the process of data creation and handling should be accompanied by strong and well-defined measures of trust building. The realisation was that this required an approach of co-creation with stakeholders. This approach would strengthen the inclusion of critical elements for trust-building, specifically (i) the

adherence to regulatory requirements in the place of data production (ii) providing insights in the data in the place of data production (iii) teaching data clerks and data stewards on the entire process of data handling (Van Reisen et al., 2021; Van Reisen et al., 2022).

The assessment of the Proof of Concept is related to the GO CHANGE track, since it defined what changes were needed in the engineering process to enhance adoption, support and uptake.

Positive requirements for FAIRification of data-handling

Inversing the negative judgements, the following positive requirements were identified:

- Flexibility

The capability to make a form fitting the data-handling.

- Standardisation of vocabularies

The capability to relate metadata to recognised thesauri and terminology tools, such as for instance Open Biological and Biomedical Ontologies (OBO), OMOP, SNOMED, ICD10/11, with community standards and protocols on the maintenance and update of vocabularies and Common Data Models (CDM).

- Training of data stewards

Provision of training of data clerks and data stewards and anyone related to the data-handling process to ensure data-handling is related to understanding of the process of the data-handling.

- Adaptability of the supporting tools

Support of the data-handling processes with software that allowed the integration with other digital tools.

- Ownership over vocabulary creation, tools or purposes

The capability of expansion of vocabulary creation, of FAIR tools and enhancement of purposes for data integration.

- Creation of trust in data production and data handling

The process of data creation and handling must be carried out with a pace and information provision that allows trust-building. This requires an approach of co-creation. The assessment of the positive

requirements for a VODAN community standard relates to the GO BUILD track as it sets a clear set of specifications and requirements for the engineering effort that can enhance buy-in by key stakeholders.

Towards a standard for health data handling: 'Ownership'

The assessment of the Proof of Concept is described as a turning point leading to a defined set of Requirements and Specifications in (Van Reisen et al., 2021; VODAN, 2021). The assessment by the VODAN team was influenced by the introduction to the CEDAR workbench. The CEDAR workbench gave the research team the expectation that it was possible to use a practical tool for training data stewards to understand, use, change and adapt the terminology behind the forms that would be used for data production. The CEDAR workbench also provided several micro-services that could align the deployment and maintenance of FAIR-based data production tools. The possibility of this alternative, which was expected to create greater operational freedom, had influenced the decision-making on the way forward. Van Reisen et al. (2021) summarised the Requirements and Specification adopted in February 2021 (VODAN, 2021) as:

> *Change 1: Design needs to be radically adapted in order to fit realities of places deployed.*
>
> *Change 2: CEDAR as a workbench to produce machine readable vocabularies.*
>
> *Requirement 1: Flexible data production (based on VODAN controlled vocabulary).*
>
> *Requirement 2: Localisation of the CEDAR Metadata System in order to achieve:*
>
> *Convergence between CEDAR localised formats*
>
> *Localised availability of CEDAR templates for premise installation in 70 hospitals each in Uganda, Ethiopia, Kenya, Liberia, Somalia, Tanzania, Nigeria, and Zimbabwe (Figure 2)*
>
> *CEDAR templates based on the Health Management Information System (HMIS) (including District Health Information System 2 (DHIS2) forms in use in the hospitals and with a VODAN agreed vocabulary* (Van Reisen et al. (2021, based on VODAN, 2021)

The Requirements and Specifications strongly reflected the need expressed in the VODAN research group that FAIR should lead to

increased ownership over data, the production and storage of it, the use and reuse of it, the tools required for its handling and that the data handling process should be based on regulatory governance in the place of provenance. The Requirements and Specifications also identified the need that data produced and handled in the place of provenance should add insights and value at this location and for the people that the data is concerned with. This is especially important for sensitive patient data, as it provides information to improve the health and wellbeing of the individuals it concerns.

The Requirements and Specifications therefore strongly identified that 'ownership' over the data-handling process, should be located in the place of provenance (or under its direct control) and that the purpose of the data-handling must be with consent by and benefit the health facility and the interests of the data-subject need to be given specific consideration. The Requirements and Specifications established a clear regulatory and ethical standard, to be the basis for all processes associated with the patient and health data handling (Van Reisen et al., 2022). The assessment of the requirements and specifications for the VODAN community reflects the GO BUILD track and sets the parameter for it.

(iii) A decision on a first reference architecture of VODAN

The VODAN reference architecture approved in February 2021 translated the Requirements and Specifications (Van Reisen et al., 2021) in an information flow chart. The reference architecture for the VODAN community reflects the GO BUILD track, establishing the concrete engineering plan.

The Information flow chart identified key features discussed below.

R1: One-time FAIR patient data production

The purpose of the VODAN Health Information System is to achieve a one-time patient data production as a digital asset. This asset must be enriched with semantic and standardised metadata. This asset must also be expressed in machine readable and actionable RDF format which can be serialised in formats like JavaScript Object Notation for Linked Data (JSON-LD), Turtle, eXtensible Markup

Language (XML), and N-Triples and loaded into a triple store for data visiting.

R2: Multi-pronged functionality

The FAIR data production will facilitate multiple functionalities in parallel. The data will be handled at source, for data visiting processes, based on explicit access and control procedures, to run algorithms over the data.

R3: Functionality for patient's health and wellbeing

The FAIR patient data must be handled first and foremost in the interest of the patient, aligned with the purpose for which this data is produced, and digitised. The process in which the data is generated, namely during a visit of the patient to the clinic for medical advice, must dictate the principal use of the data, which is to contribute to the health and wellbeing of the patient. Any further use of the data must have specific consent by the health facility, with a clear reasoning for the reuse of data for this additional purpose and considering the patients' interests.

R4: Visualisation of the data in the health facility

As the digital patient data remains an abstract asset for patients and health workers, many of whom have never seen a digital instance of patient data, or a visualisation of digital patient data, visualisation is imperative to ensure human understanding of what information and insights is generated by the data. Even if the dashboard has limited functionality, it increases the involvement of the principal stakeholders of the process in the clinic, the patients and the clinical health workers, in understanding that data produces insights (Plug et al., 2022).

R5: Visualisation of the data in the VODAN community

The digital data of patients collected by VODAN are interoperable in all the health facilities served by VODAN, independent of their location or the country in which the health facility resides. VODAN offers a cross-border set up for surveillance, while conditioning that the data handling is fully compliant with regulations available in each place. The VODAN architecture has the power to demonstrate the relevance of interoperability in a community, to advance surveillance

across different places and to visualise different realities across places. For this purpose, the reference architecture includes a dashboard of insights generated across the community on key information points agreed by the community and with signed permission by the health facilities on the indicators visualised. The community dashboard is also visible in the health facilities to compare their own situation on the indicators with other situations.

R6: Repository in a triple store

The health facilities are invited to handle the patients' data internally in the location of the health facility or under data handling control by the health facility. For this purpose, a triple store is included in the architecture for storage of the data as triples. Running queries over the triples produces knowledge and insights. Such graph-based knowledge refers to information or conclusions derived from existing data, observations, or premises through logical reasoning, rather than being explicitly stated or directly observed. It is generated through deductive or inductive reasoning, computational algorithms, or semantic reasoning techniques.

R7: FAIR Data Point

The Data is findable and accessible through a FAIR Data Point that provides access to not only the FAIR data but also a clear metadata which allows reuse. The VODAN FAIR Data Point is publicly exposed. The Health Facilities FAIR Data Points are exposed within the VODAN community only for security purposes.

R8: Bulk upload

To facilitate the data production, a bulk upload is included in the architecture. This allows FAIRification by increment with legacy data.

R9: DHIS2 compliance

DHIS2 is an open-source, web-based platform designed to collect, manage, analyse, and visualise health-related data in aggregated format for reporting purposes. It is widely used in public health to support decision-making, monitor programs, and improve healthcare service delivery, particularly in low- and middle-income countries. In 2020/2021 many health facilities were paper based, except for producing aggregate monthly reports for DHIS2 in digital format. To

not duplicate the digital input process, the DHIS2 forms were used to define the terminologies in the templates for data entry. Such a procedure would avoid duplication of effort to produce the data instances and ensure the data instance was produced only once, with automatic output from the Minimum Viable Product (MVP) to DHIS2, ensuring the compliance with R1, R2 and R8.

R10: Federation of the CEDAR Workbench

The CEDAR Workbench is used as the tool of preference, requiring the workbench to be federated to align with the needs in an African internet setting, the requirements for patient data to be handled and reposited in place of jurisdiction and on the African continent, and to ensure that offline facilities remain included in the health information system.

The following is the architecture of the data handling process:

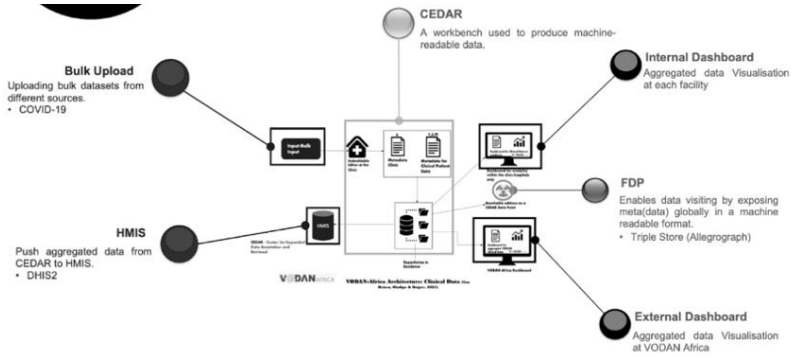


Figure 3. VODAN Architecture of Patient data Handling in the Health Facilities (2021)

Handling of multiple and parallel data catalogues in a health facility

The VODAN research group decided to focus on DHIS2 reporting forms for the OutPatient Department (OPD) data registration and the Antenatal Care (ANC) registration.

The following figure presents the different patient data catalogues (OPD and ANC) to be handled in the next phase of the research.

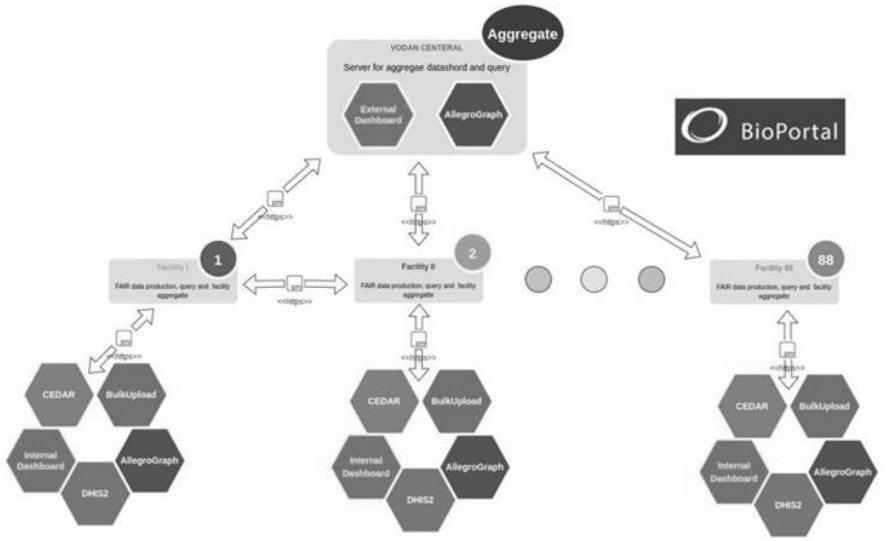


Figure 4. FAIR data production of different catalogues of patient data in selected health facilities

The figure distinguishes the software availability, with a federation of software to support the data handling in location, to support an edge-based infrastructure (Van Reisen et al., 2023).

(iv) Design of a De Novo FAIR protocol through CEDAR

There are four critical elements distinguished in the preparation of data production:

- Creation of a template for data input structuring
- Selection of available ontologies and vocabularies from a terminology server
- Creation of ontologies and vocabularies if these do not exist according to requirement
- The storage of the data in JSON-LD and RDF format

This phase of the engineering of a data-handling design is part of the GO BUILD Track.

Creation of a template for data input structuring

Figure 4 sets out the overall process from template creation to vocabulary selection, with data stored in Json LD and RDF, within a triple store (outside CEDAR), allowing the querying of the data for multipronged data handling processes (R2).

The template is important for the creation of the terminology to be incorporated. The importance of the template is that it ensures consistent and precise communication in a domain, often aligned with controlled vocabularies but broader in scope. The template is defined by a domain-specific focus: it aims at linguistic clarity rather than computational use. The terminology of the template has a flexible use. It can be used in documentation, datasets, or communication without requiring computational formalism.

The template identifies controlled vocabularies to limit terms, ensure consistency, and avoid ambiguity. Vocabularies are not formal and do not typically include logical relationships. A vocabulary is a curated set of terms, and their definitions used within a domain. It often serves as a controlled list of terms for consistent annotation of data. To provide a standardised set of labels or terms that data stewards can use vocabulary for tagging and organising information. The critical element of a vocabulary is that it provides a controlled language: this is that it limits the terms to ensure consistency and avoid ambiguity.

An important resource used to standardise metadata in Electronic Health Records (EHR) is International Classification of Disease 10 (ICD-10) and ICD-11, which is a medical classification system, maintained by the WHO and are used for coding and classifying diseases and health conditions globally.

SNOMED CT provides Ontology/Clinical Terminology. The resource provides a detailed, hierarchical vocabulary for diseases, procedures, findings, body structures, and more. It can support semantic interoperability in EHRs. SNOMED-CT is one of the most comprehensive, multilingual healthcare terminologies in the world. It is designed to encode, organise, and standardise clinical terms for use in EHRs and other healthcare information systems. SNOMED CT is organised into three core components: It links concepts to Unique, machine-readable identifiers for clinical ideas or entities. (For instance: '22298006' is the concept ID for 'Myocardial Infarction'.) It provides human-readable terms associated with concepts. Each concept may have multiple descriptions, including a fully specified name (e.g., 'Myocardial Infarction') and synonyms (e.g., 'Heart Attack'). SNOMED CT defines how concepts are related. (For

instance: 'Myocardial Infarction' is related to 'Cardiovascular Disease' through an 'is-a' relationship.) ICD 10/11 and SNOMED CT can together be used for mapping to other ontologies and standards.

An ontology is a formal, structured representation of knowledge within a specific domain. It defines the entities (concepts) and the relationships between them, using logical frameworks like Web Ontology Language (OWL). An ontology enables machines to reason about data by establishing a common framework of meaning and relationships. The key feature of an ontology is that it establishes a formal logic. Ontologies include axioms and constraints for reasoning. They establish rich semantic meaning-giving relationships, defining not only hierarchical relationships (e.g., 'is-a') but also properties and complex relationships between entities.

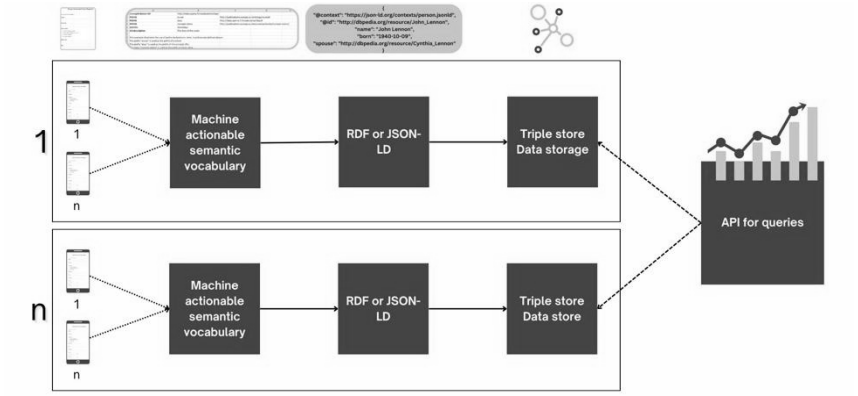


Figure 5. De NOVO FAIRification process supported by the CEDAR Workbench

Figure 5 represents the distinct stages that form the VODAN FAIRification process through CEDAR:

- Identification of terminology, controlled vocabularies, based on existing classifications, such as ICD 10 and 11 and SNOMED CT and creation of a data model.
- Creation of a template providing Machine-Readable Context: by supplementing additional metadata to describe relationships and context of the data catalogues.
- Supplementing the template with semantic vocabularies which is machine actionable (coding) language (i.e., Uniform

Resource Identifier [URI], Uniform Resource Locator [URL]).

- Using Linked Data: connect pieces of data through shared semantics, enabling integration and discovery (e.g., Json LD, RDF triples).
- Repositing the data in a triple data store.
- Creation of a Dashboard using an API for the creation of data insights (querying of the data).

The design of the template(s) allowed a finite number (n) of data sources to be queried for insights based on the common decisions regarding the engineering of FAIR data production and processing. The process improved the Findability and Accessibility of data by ensuring that data is described using consistent, well-defined terms. The use of metadata was essential for achieving Interoperability in that data carries machine-readable context. The template and curation with semantic data also supported Reusability and Accessibility by ensuring clarity in human and machine interactions.

Examples used in VODAN include a terminology guide for healthcare providers defining terms like 'Primary Care', 'Referral', 'Outbreak', etc., to ensure shared understanding. VODAN templates include vocabulary for symptoms including for instance terms of symptoms such as 'Fever', 'Cough', 'Headache', etc., with clear definitions. The template creates a dataset containing patient records linked to an ontology for disease classifications, enabling automated reasoning about health trends. An ontology in VODAN in data-handling of patient records may define entities such as 'Patient', 'Diagnosis', 'Treatment', and their relationships (e.g., 'Patient is diagnosed With Disease').

Creation of semantic data and selection of available ontologies

Semantic data is based on a data map, linking ontologies and vocabularies. An ontology represents a class structure with certain relations and hierarchy. Semantic data refers to data that is enriched with meaning, often through the use of ontologies and controlled vocabularies. It adheres to principles that allow machines to understand and interpret the data meaningfully. This enables data integration, querying, and reasoning across heterogeneous datasets.

The CEDAR workbench allows for the inclusion of available ontologies and controlled vocabularies in the back end of a template, which are linked on the workbench and were stored in BioPortal. The role of this process in designing a FAIR architecture is that it enhances the interoperability and reusability by enabling shared understanding of complex datasets. BioPortal provides tools and services for users to discover, browse, and query ontologies in the life sciences and healthcare fields. BioPortal is a comprehensive repository for biomedical ontologies.

With BioPortal, CEDAR provides the following features:

- Ontology repository: this hosts thousands of biomedical ontologies, allowing researchers to access and use them for semantic interoperability.
- Search and navigation: the capability facilitates keyword-based searches and hierarchical navigation through ontology structures.
- Annotation: the feature enables users to annotate granular datasets, catalogues and documents using ontology terms.
- APIs: the architecture offers APIs for programmatic access to ontologies, allowing integration into external tools or workflows.

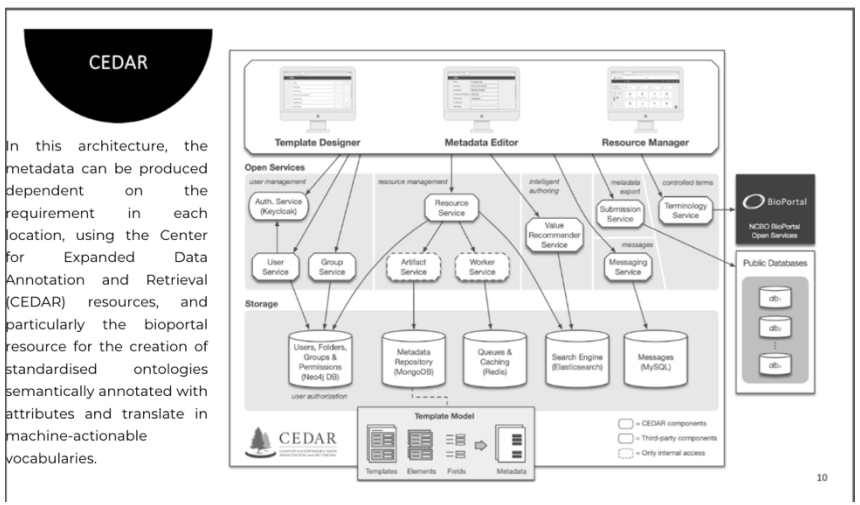


Figure 6. BioPortal mini-service in the metadata back-end functionality of FAIR data handing template creation

In CEDAR, BioPortal allows for a detailed mapping of data and metadata. This is important to create a process of collaboration in a data-handling community that helps to identify the key information elements.

The BioPortal capability of CEDAR was used by VODAN for the process of creating a community back-end of metadata. The BioPortal is one of the mini-services available within the CEDAR workbench.

Creation of new metadata

CEDAR offers the device of OntoPortal for the creation of new metadata that is needed to reflect metadata relevant for a particular community. OntoPortal is a software platform for creating custom ontology repositories on which BioPortal was created. It is the framework behind BioPortal and is designed for organisations or domains that need their own ontology repository services.

The functions provided by OntoPortal are the following:

- Ontology hosting: provides storage, curation, and access to ontologies for specialised fields.
- Interoperability and Application Programming Interface (API): facilitates integration with other systems and services through its APIs.
- Ontology services: includes annotation, search, and term mapping for hosted ontologies.
- User and access management: supports role-based access control to manage contributors and users.
- Custom repository setup: allows organisations to deploy a localised ontology repository tailored to their specific domain.

Initially, the training by the BioPortal expert of CEDAR involved working through an excel sheet to determine the elements and relations of the metadata to be entered on a new ontology, uploaded in Simple Knowledge Organization System (SKOS).

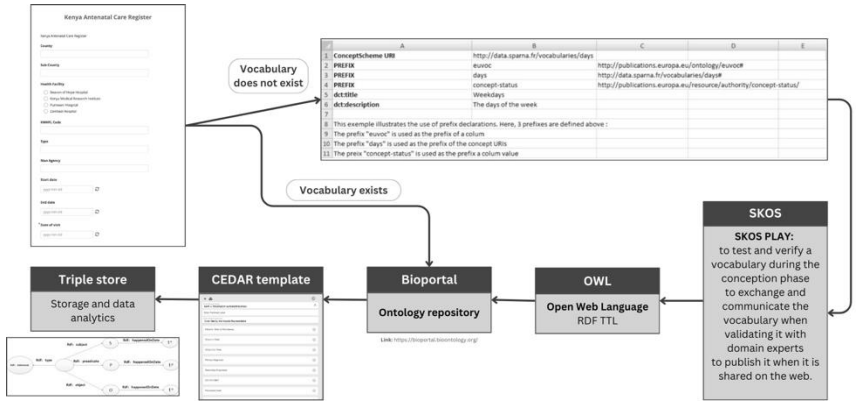


Figure 7. Creation of new ontologies

The VodanaPortal was designed to control the creation of linked data ontologies for the VODAN research group. By leveraging ontologies to standardise healthcare data for interoperability and enhanced analysis, the VodanaPortal supports integration of ontologies. The portal supports federated data sharing by health facilities. The VodanaPortal assists in tracking and analysing health-related data to support decision-making of infectious diseases and maternal health. Overall, the VodanaPortal supports FAIR data principles for managing maternal health and infectious disease data as an ontology repository, similar to BioPortal, community and domain specific to the VODAN research group.

(v) Training of the protocol and certification requirements

The VodanaPortal supports collaboration and training among students, researchers, health practitioners, and policymakers and provides training on FAIR principles and data sharing. During the scope of this research, one group of data stewards was trained with established certification requirements.

(vi) Evaluation of the FAIR data production

Based on the procedure applied, VODAN produced in 2023 a quality data pipeline production, amounting to 20.000 patient data and 300.000 triples from 12 health facilities in four countries (Ethiopia, Kenya, Nigeria and Uganda). An evaluation was carried out over the quality and (re)useability of the data, involving an inspection of the quality of terminology, vocabulary, ontologies and semantic data.

Assessing the design of the terminology creation for FAIR data handling by VODAN, a few observations are made.

1. Integration with existing resources

In healthcare and life sciences communities, such as the VODAN research group, ensuring compliance with the OMOP CDM and other established standards is essential. Where applicable, adherence to frameworks such as OBO should be prioritised. Additionally, effective mapping and integration of thesauri and terminologies are critical for semantic interoperability and data standardisation.

This integration will help increase the interoperability across communities operating in health and life sciences.

Collaboration with the Observational Health Data Sciences and Informatics (OHDSI) as a framework/community with standardised vocabularies would help the standardisation of terminology, vocabulary and ontologies. The OHDSI Common Data Model OMOP harmonises observational health data to a standard format. This includes mappings to various terminologies, e.g., SNOMED CT, and other standards such as LOINC[1], RxNorm[2], ICD, relevant for different elements of the healthcare continuum.

2. Integration of FHIR

Terminology mapping could include FHIR standards. FHIR is designed to be interoperable with other healthcare standards, such as HL7 v2, HL7 v3, and Clinical Document Architecture (CDA), by leveraging modern web technologies, including RESTful APIs, JSON, and XML. FHIR promotes data exchange between disparate systems, such as EHRs, labs, and apps.

[1] LOINC stands for Logical Observation Identifiers Names and Codes. LOINC is a standardized system for identifying laboratory and clinical observations. It provides a universal code for tests, measurements, and observations, enabling seamless data exchange across healthcare systems.

[2] RxNorm stands for Prescription Normalized (commonly abbreviated as RxNorm). RxNorm provides standardized names for clinical drugs and their relationships. It is designed to facilitate the interoperability of drug information across healthcare systems, pharmacies, and software applications. For instance: 'Atorvastatin 20 mg tablet' has a unique RxNorm code that links it to other drug-related databases.

3. Lack of relevance of the DHIS2 reporting format

The analysis identifies at least two critical challenges; the incompleteness of DHIS2 data and the lack of consistency concerning variables in the DHIS2 reporting forms, which significantly compromises data quality and reduces its utility for risk stratification (Amare et al., 2025). The assessment showed great variability of variables included in the DHIS2 forms across countries, limiting the value for a cross-country analysis of the data. The researchers found that data quality was found to be mediocre and often incomplete. In mapping the data, the experience was that often data terminology is not well related to clear definitions and controlled vocabularies, nor to ontologies and the terms are not available in linked data structures based on existing standards (Amare et al., 2025). The quality of the data collected through DHIS2 requirements was questioned by the health workers, who were asked to provide quality data for the VODAN FAIRification process, on the ground that DHIS2 is not linked to the patients not the EMR systems, and that DHIS2 only provides aggregate data that is difficult to check, and the provenance of the data is not traceable. Additionally, it only provides aggregate data rather than patient-level detail, limiting its utility for FAIR data applications. The detail of the mismatch is further elaborated by Jati et al. (2025), Amare et al. (2025a), Amare et al. (2025b).

Feasibility and relevance of identified requirements

The feasibility and relevance of the original requirements and specifications (VODAN, 2021), referred under section (iii) for developing a data handling architecture aligned with the FAIR principles for patient health information systems, was assessed as follows:

R1: One-time FAIR patient data production

This is a feasible feature, but requires strong integration with existing resources, a strong access-control capability (a process in GO BUILD which is outside the scope of this paper), and requires integration with EMR systems where feasible.

R2: Multi-pronged functionality

This is feasible, it requires the creation of services related to access and control developed as services carried over the data visiting process (a process in GO BUILD which is outside the scope of this paper). It also requires further clarification on the duty for consent by health facilities and patients (data-subjects).

R3: Functionality for patient's health and wellbeing

This is possible but requires stronger use of computer-based decision-making services (a process in GO BUILD which is outside the scope of this paper. See: Amare et al. 2025).

R4: Visualisation of the data in the health facility

This is possible but requires stronger development of dashboards for computer-based decision-making services (a process in GO BUILD and a clarification of the procedure for dynamic querying processes (which is outside the scope of this paper).

R5: Visualisation of the data in the VODAN community

This is possible but requires stronger development of dashboards for computer-based decision-making services (a process in GO BUILD and a clarification of the procedure for dynamic querying processes (which is outside the scope of this paper).

R6: Repository in a triple store

This is possible and enhances the FAIR-maturity of the architecture.

R7: FAIR Data Point

This is feasible, relevant and enhances the FAIR-maturity of the architecture.

R8: Bulk upload

This is possible and enhances the FAIR-maturity of the architecture. This affects the De Novo set up of the workflow, and requires attention is given to the provenance quality of the data instances.

R9: DHIS2 compliance

This requirement should be dropped. The templates lack structure, standards, quality and relevance, especially in cross-country settings.

R10: Federation of the CEDAR Workbench

This requirement should be dropped. The implementation of the federated CEDAR Workbench showed problems which affected the quality of the data entry. While the VODAN research team had access to the source code, modifications were not feasible within the required timeframe and demanded significant expertise. The problems included for instance the generation of a false error message, which was confusing for data clerks and data stewards and demotivated data entry. A poor user experience was consistently reported, demotivating data entry, and jeopardising the functionality of the system (Kawu et al., 2025)

Discussion

Following Plug et al. (2022), this research establishes that in a FAIR data handling setup, the terms ontology, vocabulary, semantic data, and terminology have distinct but interrelated roles, particularly in the context of a project like VODAN. The terminology, vocabulary, ontology and semantic data components work together to ensure a FAIR-compliant data environment, with ontologies offering structure, vocabularies providing standardised labels, semantic data enabling integration, and terminology ensuring precise communication.

The BioPortal, OntoPortal and VodanaPortal platforms are useful for ontology definition and FAIR compliance. The portals target different aspects of ontology management, customisation, and data sharing for specific communities and needs. While BioPortal is specifically focused on the GO BUILD track, the OntoPortal allows for GO CHANGE, to adopt ontology management for different communities. The VodanaPortal is a community GO BUILD tool that reflects the changes made to develop community relevant ontologies.

The assessment of the requirements provides the following summary:

Relevant and feasible requirements (R1–R8):

R1: One-time FAIR patient data production: Feasible but requires strong integration with existing resources and robust access-control processes (GO BUILD)

R2: Multi-pronged functionality: Achievable through services for access and control developed around the data visiting process (GO BUILD)

R3: Patient health and wellbeing functionality: Possible with advanced computer-based decision-making services (GO BUILD)

R4 & R5: Data visualisation at health facilities (R4) and within the VODAN community (R5): Possible with stronger dashboard development and dynamic querying capabilities (GO BUILD).

R6, R7, R8: Enhancing FAIR maturity: Triple Store Repository (R6), FAIR Data Point (R7), and Bulk Upload (R8) are feasible and improve the architecture's FAIR maturity (GO BUILD)

Requirements Recommended for Dropping (R9 & R10):

R9: DHIS2 Compliance:

Found to be unsuitable due to variability in forms, poor data quality, lack of linkage to controlled vocabularies and ontologies, and absence of traceable provenance. Health workers questioned its utility for FAIR data processing as it provides aggregate data rather than patient-level details (GO CHANGE).

R10: Federation of CEDAR Workbench:

Identified as problematic due to persistent quality issues in data entry processes, false error messages, and overall poor user experience. These issues demotivated data clerks and data stewards (GO CHANGE).

R1-R8 are relevant and feasible for enhancing the FAIR architecture (GO BUILD).

R9 and R10 should be dropped due to significant quality and usability concerns (GO CHANGE).

The study emphasises the need for stronger integration and dashboard development to achieve the outlined goals while excluding requirements that undermine data quality and usability.

Relevance of GO TRAIN in Achieving R1–R8

The GO TRAIN phase plays a critical role in operationalising and sustaining the implementation of the relevant and feasible

requirements (R1–R8) for a FAIR patient health information system. Its focus on capacity building, training, and knowledge dissemination ensures that the necessary technical skills and operational know-how are embedded within the system's users and stakeholders. This phase complements the GO BUILD phase by addressing human and institutional factors essential for the system's success.

Key contributions of GO TRAIN to each requirement

R1: One-time FAIR patient data production

Ensures that healthcare professionals and data stewards are well-trained in implementing robust access-control processes and integrating existing resources. Training programs, such as developed by Taye et al. (2024) and Folorunso et al. (2025) focus on standard operating procedures for data collection, FAIRification, and security.

R2: Multi-pronged functionality

Provides training for stakeholders on utilising and managing access-control services and the data visiting process. By educating users on the technical and ethical aspects of multi-pronged data functionalities, it ensures effective system use, such as proposed by Folorunso et al., 2025).

R3: Patient health and wellbeing functionality

Focuses on empowering healthcare workers with the skills to use advanced computer-based decision-making tools. Training modules ensure that staff can interpret and apply system outputs to enhance patient care and wellbeing.

R4 & R5: Data visualisation at health facilities and within the VODAN community

Facilitates capacity building on dashboard use and dynamic querying processes. Through practical training, users learn to generate and interpret visual data for local health facility needs (R4) and community-wide insights (R5).

R6, R7, R8: Enhancing FAIR maturity

R6 (Triple Store Repository): Teaches technical teams how to maintain and query the triple store for optimal performance.

R7 (FAIR Data Point): Ensures that data stewards understand how to manage FAIR Data Points to uphold the system's FAIR maturity.

R8 (Bulk Upload): Focuses on training for efficient and error-free bulk data upload processes.

Sustainability through GO TRAIN

The relevance of GO TRAIN lies not only in the technical upskilling it provides but also in the creation of a culture of continuous improvement and innovation in an interdisciplinary set-up. By embedding the principles of FAIR data management into organisational workflows, it ensures that the system remains effective and adaptable to future needs.

Contrast with GO CHANGE for R9 and R10

For requirements R9 (DHIS2 Compliance) and R10 (Federation of CEDAR Workbench), the challenges identified indicate that GO TRAIN efforts would not resolve fundamental issues related to system design and data quality. Instead, these requirements necessitate structural changes (GO CHANGE) to address their unsuitability.

This study found that GO TRAIN is indispensable for ensuring the successful implementation and sustained operation of R1–R8, providing the necessary knowledge and skills to maximise the system's potential while supporting the principles of FAIR data.

This section positions GO TRAIN as a vital enabler for meeting the feasible requirements while distinguishing its role from the structural changes required for unsuitable requirements (R9 and R10).

The results underscore the conclusions of earlier studies that lack of knowledge, understanding and technical training on FAIR data stewardship are detrimental to the uptake of it in both policy and technical propositions (Zhang, 2022).

Conclusion

This research examines the critical components required for creating a data architecture compliant with FAIR (Findable, Accessible, Interoperable, and Reusable) principles, focusing on patient health information systems within projects like VODAN. It identifies the

interrelated roles of terminology, ontologies, vocabularies and semantic data, in enabling structured and interoperable data environments. The study assessed ten requirements for achieving FAIR maturity within patient health information systems. Eight requirements were identified as both feasible and relevant. These include one-time FAIR patient data production, multi-functional access and control services, functionalities for improving patient health and wellbeing, data visualisation at health facilities and within the VODAN community, and enhancements in FAIR maturity through triple-store repositories, FAIR Data Points, and bulk upload processes.

Two requirements, DHIS2 compliance and federation of the CEDAR workbench, were found to be unsuitable. due to inherent respective data quality and usability challenges, requiring structural adjustments through GO CHANGE. DHIS2 compliance was deemed unfit due to variability in data forms, poor data quality, lack of connection to controlled vocabularies and ontologies, and the absence of traceable provenance. Similarly, the CEDAR workbench federation requirement was identified as problematic due to persistent data entry issues, usability challenges, and a poor user experience that demotivated data clerks and stewards. These challenges necessitate structural changes within GO CHANGE to resolve fundamental system design flaws and ensure alignment with FAIR principles.

This study affirms the indispensability of a systematic, multi-phased approach, which incorporates training as a vital component, to building sustainable FAIR data architectures for patient health information systems. By addressing both societal needs and academic feasibility, the proposed approach demonstrates the practical application of FAIR principles to patient data while establishing a scalable framework for FAIRification across other domains. This alignment between academic rigor and societal urgency exemplifies how FAIR principles can address real-world challenges, advancing data stewardship practices and improving the FAIR maturity of data in healthcare and life sciences.

The research emphasises the critical role of the GO TRAIN phase in operationalising the identified feasible requirements. By focusing on capacity building, technical training, and knowledge dissemination,

GO TRAIN ensures that stakeholders are equipped with the necessary skills to implement and sustain FAIR-compliant systems. Key contributions of GO TRAIN include training healthcare professionals on access-control processes, empowering stakeholders to use advanced decision-making tools, and building expertise in data visualisation, querying, and bulk data uploads. GO TRAIN also fosters a culture of continuous improvement and innovation, embedding FAIR principles into multidisciplinary organisational workflows to ensure long-term effectiveness and adaptability of the system. It is concluded that the GO TRAIN task is an essential cornerstone for translating theoretical FAIR principles into practical implementation by ensuring the technical expertise, institutional capacity, and continuous improvement needed to operationalise and sustain FAIR-compliant data systems in healthcare and beyond.

Acknowledgements

The authors would like to express our sincere gratitude and appreciation to our supervisors, particularly Professor Mirjam van Reisen, for her outstanding leadership and inspiring mentorship throughout this study. Furthermore, we acknowledge the invaluable support provided by VODAN-Africa, the Philips Foundation, the Dutch Development Bank (FMO), CORDAID. Their contributions have been instrumental in facilitating this research. We deeply appreciate the collective efforts and collaboration that have made this study possible.

Authors' Contributions

Aliya Aktau and **Samson Yohannes Amare** are both first authors of this chapter and conducted the research for this chapter as part of their PhD studies. They prepared the manuscript and reviewed its content. Putu Hadi Purnama **Jati** was responsible for the documentation. **Getu Tadelle Taye** developed the training manual. **Tesfit Gebreslassie Gebremeskel** and **Ruduan Plug** assisted in the design and implementation of the research. **Mirjam van Reisen** reviewed and edited the chapter.

Ethical Considerations

Tilburg University, Research Ethics and Data Management Committee of Tilburg School of Humanities and Digital Sciences REDC#2020/013, June 1, 2020-May 31, 2024, on Social Dynamics of Digital Innovation in remote non-western communities. Uganda National Council for Science and Technology, Reference IS18ES, July 23, 2019-July 23, 2023.

References

Amare, S. Y., Taye, G. T., Gebreselassie, T. G., Plug, R., & Van Reisen, M. (2023). Realizing health data interoperability in low connectivity settings: The case of VODAN-Africa. In: FAIR Connect, 1(1), 55-61. IOS Press. https://doi.org/10.3233/FC-221510

Amare, S.Y., Medhanyie, A. A. & Mirjam Van Reisen. (2024). Data Visiting in Digital Black Holes: FAIR Based Digital Health Innovation during War. In: Van Reisen, M. & Mawere, M. (Eds.), *Tigray. War in a Digital Black Hole*, Volume 1. Langaa, Bamenda. Pp. 477-507. Chapter URL: https://www.researchgate.net/publication/385554817_Data_Visiting_in_Digital_Black_Holes_FAIR_Based_Digital_Health_Innovation_during_War

Amare, S.Y., Weldu, L.M., Gebremeskel, T. (2025). Harmonizing Antenatal Care Records Across Four African Countries: A Comparative Analysis and Development of a Standardized Data Model. In M. Van Reisen, M., Amare, S. Y., Maxwell, L. & Mawere, M. (Eds.), *FAIR data, FAIR Africa, FAIR world: Internationalisation of the Health Data Space*. Bamenda: Langaa

Amare, S. Y., Taye, G. T., Plug, R. B. F., Medhanyie, A. A., & Van Reisen, M. (2025). Federating tools for FAIR patient data: Strengthening maternal health and infectious disease surveillance from clinics to global systems. In Van Reisen, M., Amare, S. Y., Maxwell, L. & Mawere, M. (Eds.), *FAIR data, FAIR Africa, FAIR world: Internationalisation of the Health Data Space*. Bamenda: Langaa

Basajja, M., Suchanek, M., Taye, G.T., Amare, S.Y., Nambobi, M., Folorunso, S., Plug, R., Oladipo, F.O., Van Reisen, M. (2022). Proof of concept and horizons on deployment of FAIR Data Points in the COVID-19 pandemic. *Data Intelligence, 4*(4) (2022). https://doi.org/10.1162/dint_a_00179

Basajja, M., Van Reisen, M., Oladipo, F. (2022). FAIR Equivalency with regulatory framework for digital health in Uganda. *Data Intelligence, 4*(4) (2022). https://doi.org/10.1162/dint_a_00170

Buchanan R. (1992). Wicked problems thinking in design. *Design Issues*. 1992;8 (2):5-21.

East African Health Research Commission (EAHRC). (2017). *Research for Health and Prosperity*. Digital Reach Initiative Roadmap.

Folorunso, S. O., Oladipo, F. O., & Van Reisen, M. (2025). A higher education curriculum for cultural competence, representation, and social responsibility in AI and FAIR data practices. In Van Reisen, M., Amare, S. Y., Maxwell, L. & Mawere, M. (Eds.), FAIR data, FAIR Africa, FAIR world: Internationalisation of the Health Data Space. Bamenda: Langaa

Gebreslassie, T. G., Kidanu, G., Mamo, L., Amare, S. Y. & Van Reisen, M. (2024). Life in Darkness: The Communication Blockade during the Tigray Siege. In: Van Reisen, M. & Mawere, M. (Eds.), *Tigray. War in a*

Digital Black Hole, Volume 1. Langaa, Bamenda. Pp. 103-144. Chapter URL: https://www.researchgate.net/publication/385452677_Life_in_Darkness_The_Communication_Blockade_during_the_Tigray_Siege

Groenen, K. H. J., Jacobsen, A., Kersloot, M. G., Vieira, B., van Enckevort, E., Kaliyaperumal, R., Arts, D. L., 't Hoen, P. A. C., Cornet, R., Roos, M., & Schultze Kool, L. (2020). The de novo FAIRification process of a registry for vascular anomalies. *medRxiv*. https://doi.org/10.1101/2020.12.12.20245951

Jacobsen, A., Kaliyaperumal, R., Olavo Bonino da Silva Santos, L., Mons, B., Schultes, E., Roos, M., Thompson, M. (2020). A Generic Workflow for the Data FAIRification Process. *Data Intelligence, 2*(1-2): 56–65. https://doi.org/10.1162/dint_a_00028

Jati, P. H., Amare, S. Y., Kawu, A. A., Nandwa, W., Taye, G. T., & Van Reisen, M. (2025). GO CHANGE: Adoption of FAIR-OLR architectures to support insights from patient health data records in Africa. In Van Reisen, M., Amare, S. Y., Maxwell, L. & Mawere, M. (Eds.), *FAIR data, FAIR Africa, FAIR world: Internationalisation of the Health Data Space.* Bamenda: Langaa

Kawu, A. A., Purnama Jati, P. H., Schrijver, L., Van Reisen, M., Tadele, G., O'Sullivan, D., & Hederman, L. (2025). A critical incidence assessment of a FAIR implementation study on patient health data in Africa. In Van Reisen, M., Amare, S. Y., Maxwell, L. & Mawere, M. (Eds.), *FAIR data, FAIR Africa, FAIR world: Internationalisation of the Health Data Space.* Bamenda: Langaa

Lin, Y., Purnama Jati, P.H., Aktau, A., Ghardallou, M., Nodehi, S., Van Reisen, M. (2022). Implementation of FAIR Guidelines in selected non-Western geographies. *Data Intelligence, 4*(4) (2022). https://doi.org/10.1162/dint_a_00169

Mons, B. (2018). *Data Stewardship for Open Science: Implementing FAIR Principles.* CRC Press.

Oladipo, F., Folorunso, S., Ogundepo, E., Osigwe, O., Akindele, A. (2022). Curriculum Development for FAIR Data Stewardship. *Data Intelligence, 4*(4): 991–1012. https://doi.org/10.1162/dint_a_00183

Plug, R., Liang, Y., Aktau, A., Basajja, M., Oladipo, F., Van Reisen, M. (2022). Terminology for a FAIR framework for the Virus Outbreak Data Network-Africa. *Data Intelligence, 4*(4). https://doi.org/10.1162/dint_a_00167

Sanders L. (2008). On modeling: an evolving map of design practice and design research. *Interactions*, 2008; 15 (6):13-17. https://doi.org/10.1145/1409040.1409043

Stocker, M., Stokmans, M., Van Reisen, M. (2022). Agenda setting on FAIR Guidelines in the European Union and the role of expert committees. *Data Intelligence, 4*(4) (2022). https://doi.org/10.1162/dint_a_00168

Strawn, G. (2021). 75 years of astonishing evolution of IT: 1946–2021. *IEEE IT Professional, 23*(4), 6–11. https://doi.org/10.1109/MITP.2021.3070963

Taye, G. T., Amare, S. Y., Gebreslassie, T. G., Jati, P. H. P., Kawu, A. A., Kievit, R., Plug, R., Smits, K., Stocker, J., & Van Reisen, M. (2024). *Training Manual: A practical guide to FAIR Data Stewardship.* Brussels: EEPA. VODAN. https://aun.mu.edu.et/ahds/training/

Taye, M., Medhanyie, A. A. & Van Reisen, M. (2024). War-related Destruction of the Digital Health Data Infrastructure: Discovering Features for a Resilient Digital Health Information System. In: M. Van Reisen & M. Mawere (Eds.), *Tigray. War in a Digital Black Hole*, Vol 1, pp. 439-476. Langaa, Bamenda. Pp. 439-476. Chapter URL: https://www.researchgate.net/publication/385554308_War-related_Destruction_of_the_Digital_Health_Data_Infrastructure_Discovering_Features_for_a_Resilient_Digital_Health_Information_System

Van Reisen, M. & Oladipo, F. (2020) Proof of Concept developed by VODAN Africa and Asia. *GO FAIR*. 27 October 2020. Available at: https://www.go-fair.org/2020/10/27/proof-of-concept-developed-by-vodan-africa-and-asia/

Van Reisen, M., Amare, S.Y., Nalugala, R., Taye, G.T., Gebreselassie, T.G., Medhanyie, A.A., Schultes, E. & Mpezamihigo, M. (2023). Federated FAIR principles: Ownership, localisation and regulatory compliance (OLR). In: *FAIR Connect*, 1(1), 1-7. IOS Press. https://doi.org/10.3233/FC-230506

Van Reisen, M., Oladipo, F., Mpezamihigo, M., Plug, R., Basajja, M., Aktau, A., Purnama Jati, P.H., Nalugala, R., Folorunso, S., Amare, Y.S., Abdulahi, I., Afolabi, O.O., Mwesigwa, E., Taye, G.T., Kawu, A., Ghardallou, M., Liang, Y., Osigwe, O., Medhanyie, A.A., Mawere, M. (2022). Incomplete COVID-19 data: The curation of medical health data by the Virus Outbreak Data Network-Africa. *Data Intelligence, 4*(4) (2022). https://doi.org/10.1162/dint_e_00166

Van Reisen, M., Oladipo, F., Stokmans, M., Mpezamihigo, M., Folorunso, S., Schultes, E., Basajja, M., Aktau, A., Amare, S. Y., Taye, G. T., Purnama Jati, P. H., Chindoza, K., Wirtz, M., Ghardallou, M., van Stam, G., Ayele, W., Nalugala, R., Abdullahi, I., Osigwe, O., Graybeal, J., Medhanyie, A. A., Kawu, A. A., Liu, F., Wolstencroft, K., Flikkenschild, E., Lin, Y., Stocker, J., & Musen, M. A. (2021). Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research. *Advanced Genetics, 2*(2), e10050. https://doi.org/10.1002/ggn2.10050

Van Reisen, M., Stokmans, M., Basajja, M., Ong'ayo, A., Kirkpatrick, K., & Mons, B. (2020). Towards the tipping point of FAIR implementation. In B. Mons, A. Jacobsen, & E. Schultes (Eds.), *Data Intelligence, 2*(1-2), 264–275. MIT Press. https://doi.org/10.1162/dint_a_00049

Van Reisen, M., Stokmans, M., Mawere, M., Basajja, M., Ong'ayo, A., Nakazibwe, P., Kirkpatrick, K., & Chindoza, K. (2020). FAIR practices in Africa. In B. Mons, A. Jacobsen, & E. Schultes (Eds.), *Data Intelligence, 2*(1-2), 246–256. MIT Press. https://doi.org/10.1162/dint_a_00047

VODAN. (2021, January). *Technical requirements document [Version 1]: Phase II implementation project*. VODAN Africa & Asia.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3(1), 1–9.https://doi.org/10.1038/sdata.2016.18

Zhang, S., Dos Santos Vieira, B., Bernabé, C. H., Abaza, H., Benis, N., Cámara, A., Cornet, R., Le Cornec, C. M. A., 't Hoen, P. A. C., Schaefer, F., van der Velde, K. J., Swertz, M. A., Wilkinson, M. D., Jacobsen, A., & Roos, M. (2022). Towards FAIRification of sensitive and fragmented rare disease patient data: Challenges and solutions in European Reference Network registries. *Orphanet Journal of Rare Diseases, 17*, 436. https://doi.org/10.1186/s13023-022-02558-5