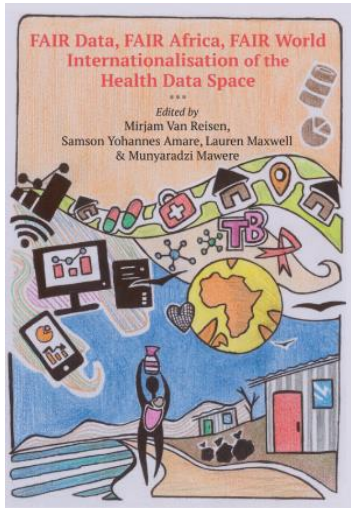
Bridging Borders with FAIR Data: Transforming Digital Ecosystems for Maternal Health and Public Health Surveillance in Africa

Samson Yohannes Amare

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



Cite as: Amare, S. Y. (2025). Bridging borders with FAIR data: Strengthening maternal health and public health surveillance in Africa. https://doi.org/10.5281/zenodo.15382805. In M. Van Reisen, S. Y. Amare, & M. Mawere (Eds.), *FAIR data, FAIR Africa, FAIR world: Internationalisation of the Health Data Space.* (pp. 25–64). 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

Chapter 2

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

Samson Yohannes Amare

Abstract

This research documents an implementation study of a federated data architecture for managing patient data across eight African countries. The digital platform, deployed as a minimal viable product in 74 out of 88 signed-up health facilities, enabled localised data management with distinct entry forms prepared for one-time data entry. This process ensured the creation of FAIR (Findable, Accessible, Interoperable, and Reusable) data enriched with semantic and machine-readable features. Adaptations were made to accommodate country-specific resource constraints and regulatory frameworks. These choices were documented in a FAIR Implementation Profile and shared to make them globally accessible as practices for implementers along with specifications, which were made public. This process was supported by locally deployed tools as micro-services, supporting FAIRification. Analytics derived from patient data contributed to more informed clinical decision-making and enhanced patient care. The availability of high-quality, FAIR-compliant data enhanced research outcomes, supporting evidence-based medical advancements. We found that the De Novo FAIRification of routine patient data through data visiting could be supported by workflows that are responsive to resource-limited African settings. The study found that the FAIR framework, which supports Ownership, Localisation, and Regulatory Compliance (OLR) of patient data, presents a viable new digital ecosystem that addresses challenges related to reusing sensitive, individual-level health data in low-resource settings.

Keywords: De Novo FAIRification, FAIR by design, FAIR data, HMIS, patient data, digital health system, data visiting

Introduction

Jochems et al. (2016) demonstrate that distributed learning facilitates the development of predictive models across multiple hospitals, while addressing data-sharing barriers. This approach enables the extraction and utilisation of routine patient data, while ensuring compliance with national and European data protection regulations. Using real-time digital patient data recorded by health workers in Kenya, Aksünger et al. (2021) identified key determinants of the continuum of maternal care, highlighting critical preventive actions for improved antenatal health outcomes. Dos Santos Vieira et al. (2022) found that rare disease patient data, which are highly sensitive, distributed across multiple registries, and managed by different custodians, often lack interoperability. Ensuring that data is Findable, Accessible, Interoperable, and Reusable (FAIR) at the source, both for humans and machines, enabled federated discovery and analysis across disparate databases, supporting accurate diagnosis, optimised clinical management, and personalised treatment.

To realise the benefits of a federated approach to participant-level data reuse, Strawn (2021) stated that "it is necessary to understand the data before it can be used as training data" (p. 26), but concluded that "we cannot do this efficiently yet" (p. 27). He emphasised the central importance of the FAIR principles in enabling the reuse of federated data. The FAIR principles provide minimal standards and working implementations, which may eventually evolve into a fully integrated Internet of FAIR Data and Services (IFDS) (Schultes & Wittenburg, 2019). The principles emphasise machine actionability, acknowledging that digital objects exist along a continuum of possible states and are utilised by computational agents (Wilkinson et al., 2016). Ensuring effective digital object management that supports machine actionability requires the curation of detailed metadata, facilitating autonomous and computational data exploration and analysis.

Despite acknowledging that the FAIRification of patient data can strengthen health systems by facilitating data reuse for improved insights, the role of semantics in machine interpretation is poorly understood (Strawn, 2021), and there is a lack of operational tools to support the implementation of FAIR principles. Stocker et al. (2022)

highlighted the lack of tools tailored to diverse technological and social contexts, which hinders the scalability and efficiency of data curation as machine-actionable assets. Developing a comprehensive digital platform architecture that enables machine-actionable FAIR data requires advancements in both technical and social domains.

The lack of actionable tooling for FAIR implementation is critical in the African context, where few architectures have been tested (Van Reisen, Stokmans, Basajja et al., 2020; Lin et al., 2022). The health data reuse landscape in Africa is characterised by fragmented efforts that lack interoperability (Neumark & Prince, 2021). Data findability for care and research is limited, with little to no access control mechanisms, which limits trust in emerging or existing data reuse efforts. Van Reisen et al. (2021) and Van Reisen et al. (2022) outlined how FAIR data can be used to address fragmented data reuse efforts, thereby improving health services and research.

Pointing to the relevance of a FAIR data ecosystem, Gregurick (2020) outlined a strategy in which FAIR resources would enhance treatments for affected newborns, develop new and improved prevention and treatment strategies, and optimise effective treatments. Building up the ecosystem requires FAIR Supporting Resources (FSR) (FAIRConnect, n.d.). This is "a resource that supports the FAIRification or FAIR Orchestration of data and metadata" (Gregurick, 2020).

By systematically applying the FAIR principles and leveraging FSRs, addressing tooling gaps presents an opportunity to expand FSRs and develop context-specific solutions. This approach facilitates the development of tools and semantic frameworks that enhance the speed and scale of machine-actionable data curation. Ensuring that the tools themselves adhere to FAIR principles will contribute to the sustainability of IFDS (Amare, 2023).

The FAIR principles outline a progressive pathway toward achieving machine-actionability, culminating in an optimal state in which machines can fully understand, utilise, and reuse digital objects. As machines navigate the data ecosystem, they should be able to read data and act autonomously on it by understanding a range of data types and formats, as well as various access mechanisms and

protocols (Wilkinson et al., 2016). An optimal state is defined by a machine's ability to make informed decisions autonomously when encountering new data. Machine actionability applies to both the contextual metadata and the content of the digital object itself, with each existing along its own continuum of actionability (Wilkinson, 2016).

In scenarios where data is sensitive, personally identifiable, or pertains to non-data research objects, FAIRness can still be achieved by providing rich metadata that fully describes the digital object without requiring the publication of the data itself. This approach ensures compliance with FAIR principles, while maintaining ethical and regulatory standards (Wilkinson et al., 2016).

The FAIR guiding principles outline 15 facets aimed at enhancing the process of making data FAIR for both human and machine agents, enabling efficient and accurate analysis of data from diverse sources (Wilkinson et al., 2016). FAIRification refers to the process of aligning meta/data with these guiding principles (Jacobsen et al., 2020). Notably, the principles are technology-agnostic and do not prescribe specific standards, tools, or implementation solutions. Instead, they precede implementation choices, allowing flexibility in their application.

As the FAIR principles do not dictate a specific sequence of steps for the FAIRification process, various experiences and recommendations have emerged regarding the best approach to achieve FAIR compliance. This variability reflects the principles' adaptability to different contexts and use cases, while also highlighting the need for shared practices to enhance consistency, interoperability, and eventual convergence. This ensures that data is curated as FAIR and Federated, and AI-Ready (Strawn, 2021), which will accelerate the availability of inclusive, quality data pipelines that can serve the health system in creating use cases that can generate insights from the data by building quality models (Strawn, 2021; Amare, 2023).

Jacobsen et al. (2020) developed a generic multiple-step workflow to help data FAIRification. Jacobson et al. (2020) and Groenen et al. (2021) explicitly position the FAIRification process at the beginning of data creation. The implementation study by Groenen et al. (2021)

follows those FAIRification steps in a process of FAIRifying data from an electronic data capture (EDC) system. Groenen et al. (2021) curated data on rare diseases—on which there is generally scarce data and can therefore benefit from the interoperability of data across multiple datasets—and made it "as open as possible and as closed as necessary".

This research investigated whether or not and how the De Novo FAIRification workflow could be effectively implemented for patient data curation in African settings, considering the need for relevance, specificity, and convergence to maximise the benefits of FAIR participant-level health data, with a focus on maternal and child health-related data. Specifically, the study investigated the potential for implementing FAIR-by-design data principles to enable data visiting and interoperability across diverse communities while maintaining context-sensitive and sustainable practices.

Methods

This research was carried out within the Value-driven Ownership of Data and Accessibility Network (VODAN) Africa research group. The VODAN research group supports a platform for the federated reuse of participant-level data from health facilities in Africa, representing a collective effort supported by universities, national and regional health ministries, and offices.

We employed an ethnographic approach to address the 'wicked' problem of implementing the FAIR principles in practice and transitioning from a centralised data reuse approach to a federated platform for data reuse (Van Reisen et al., 2021). The research was conducted in select African health facilities, following principles of implementation science for adoption.

This research facilitated the discovery of factors related to the success or failure of software development in a setting distinct from the mainstream implementation contexts in Europe and the US (Van Reisen, Stokmans, Basajja et al., 2020; Van Reisen, Stokmans, Mawere et al., 2020). Operating in research-limited contexts in diverse African countries allowed for an in-depth investigation into the design challenges and tooling requirements for implementing cutting-edge technology in low-digital-resource settings. Throughout the design

and conduct of the study, the research group actively engaged with health workers who are the primary users of the system and the data stewards who support them.

Study location

This study was conducted in eight African countries: Ethiopia, Kenya, Nigeria, Somalia, Tanzania, Tunisia, Uganda, and Zimbabwe. The initial deployment was prepared for 88 health facilities across 8 countries. The actual deployment of the complete tool was realised in 74 health facilities. The uptake of health facilities was more extensive than initially expected. In the original study design, a deployment was planned in 30 health facilities across 3 countries. The expanded uptake demonstrated health system stakeholders' interest in the potential of FAIR curation of patient data, as no financial compensation was available to the health facilities, which had to bear the costs of implementation.

Study timeline

The study was carried out from January 2021 to December 2022. This study built on a previous study undertaken in 2020 and documented in a Special Issue of Data Intelligence (Van Reisen et al., 2022).

Co-design approach

The research employed a community co-creation design approach to develop the platform. The VODAN research group, which includes the VODAN-Africa chapter, is organised into a community of practice ecosystem. During the study period, VODAN Africa consisted of groups from the eight participating countries, which met virtually weekly, along with researchers from Europe and Asia. Community members representing the participating organisations in the eight countries were nominated in part because of their data stewardship competency and their experience working closely with health facilities in their respective countries to facilitate cross-country learning. The virtual, cross-regional meetings provided VODAN chapters with the opportunity to share progress and experiences and to collectively resolve challenges experienced during the implementation of the FAIRification and federated data reuse pipelines. Additionally, country chapters met regularly, both in person and online. In WhatsApp community groups, participants regularly

posted about their work, progress, challenges, and opportunities, serving as a record of the work. The weekly meetings were video-recorded and available to all community members, serving as further records of the implementation process. The documentation from the implementation process was archived as a resource for the research community.

Teams established in each country represented in the VODAN Africa research group had one or more country coordinators. The coordinator was responsible for building the network in each country, keeping the ministries and bureaus of health informed, implementing introductions and follow-up activities, ensuring the deployment was implemented, and monitoring the FAIR data production. The country coordinator was responsible for ensuring regulatory compliance at both the national and regional levels, as well as for creating conditions that facilitated the execution of activities in their respective health facilities and research centres. To translate principles into practice, community members discussed implementation decisions to reach a consensus with their respective communities on the FSRs to be created or reused, as well as on the data pipeline to be established. The country coordinators were responsible for documenting and following up on the implementation research. In countries with multiple coordinators covering different regions, the coordinators worked together.

The community was organised with specialised technical teams focusing on (i) semantic data modelling and linkage, (ii) software tooling, (iii) training/data stewardship, and (iv) deployment.

Reusing existing tooling for open science

Our approach to building cross-facility interoperability and federated data reuse (Sanders, 2008; Sanders & Stappers, 2014) focused on creating a high-quality data pipeline that adhered to the FAIR guiding principles, while translating these principles into the use and adaptation of products and services that help produce FAIR data. We used FAIR Implementation Profiles (FIPs) to provide a structured approach for documenting the choices made to implement each of the 15 principles and to offer guidance for implementing the

FAIRification workflow using FSRs, which encompass software systems that facilitate the production of FAIR data.

The co-design of a minimum viable product (MVP) platform followed a phase of testing a proof of concept and the deployment of a data production tool developed by the Data Stewardship Wizard (DSW) (Van Reisen et al., 2022). Following the assessment of this phase, the VODAN Research group conducted an evaluation that led to the development of requirements and specifications (VODAN, 2021). The requirements and specifications included the use of the Center for Expanded Data Annotation and Retrieval (CEDAR) Workbench for metadata template creation and AllegroGraph as a triple store. The choice reflected the situation at the start of this research when just a handful of open-source tools were available for FAIRification, including DSW and CEDAR. The selection of CEDAR was based on the requirement that data would be entered in the health facility only once for multiple functional operations, as well as for multiple functional operations in a De Novo FAIRification workflow (VODAN, 2021).

In the engineering of the platform for patient data curation, the modus operandi was to identify available tools and test them for use in curating patient health data within health facilities. The practicality of deploying tools in situational settings for patient data curation was assessed by identifying available tools and testing them for use in the curation of patient health data in healthcare facilities, as well as evaluating their practicality in situational settings. Beyond the deployment of such tools, making them work in harmony requires building an interoperability framework at the data, application, and process levels.

Following the analysis of the FSR landscape and the needs of the health facilities, adaptations were engineered. In cases where no tooling was available to respond to users' needs or existing tools were required to perform beyond their initial design, new tooling was programmed according to specifications produced for the specific tool.

All engineering efforts were carried out using Free and Open Source (FOS) platforms and freely available assets. This corresponds with the

general philosophy of Open-Source (FOS) platforms and freely available assets, in line with the values of FAIR Open Science. The use of non-proprietary, free, and open-source software is an enabler in the deployment of extensive automation systems. This approach enables the deployment of extensive automation systems while avoiding large payments to suppliers and ensuring sustainability (May et al., 2006). The engineering choices recognised the digital reality in different places as well as the social and regulatory situation.

Data selection

The patient data selected for FAIRification included data from antenatal care (ANC) service registries and outpatient departments (OPDs). Most health facilities in Africa are required by their respective ministries to prepare a routine health management information system (HMIS) report. This report is prepared in many countries by encoding aggregate data into a software system known as the District Health Information System 2 (DHIS2). The HMIS registries were identified as suitable for data entry, particularly given that in the majority of health facilities, the first digitisation of data occurred in a DHIS2-compatible format and that all countries had some DHIS2-related registration responsibilities.

Regulatory and privacy preservation concerns

To ensure regulatory compliance, the VODAN research team conducted a study to measure the equivalency of the country's regulatory and policy framework with the FAIR data principles (Van Reisen et al., 2022). In addition to the FAIR Equivalency analysis, the ministries of health and the relevant bureaus of health were informed, and permission was requested for the study, which is typically granted through an exchange of letters. A data use agreement was signed with the administrator at each health facility, the relevant ministry or bureau of health, the Country Coordinator of VODAN, and VODAN's Executive Coordinator. The data use agreement outlined in detail how the data were curated and reposted, as well as the processes run over the data. The agreement referred to the European Union's General Data Protection Regulation (GDPR) as a binding reference document on personal data protection, ensuring data privacy and security.

Relevance of the study

Public health crises, such as Ebola, have made it evident that having a robust federated data reuse system in place enables the timely detection and response to local and global health challenges, thereby substantiating the need for federated data (Van Reisen et al., 2021). The COVID-19 pandemic highlighted the urgent need for innovative global interventions, robust data pipelines, e-learning platforms, and other digital solutions to mitigate the impact of cross-border epidemics and pandemics in Africa and beyond.

The FAIR principles have been widely adopted in policies and practice (Stocker et al., 2022) and are regarded as enhancing the management, utilisation, and reusability of health data (Lin et al., 2022). Although the European Union has adopted the FAIR principles as the gold standard for data sharing and other forms of reuse (Guillot et al., 2023), these principles lack comprehensive technical guidance for their practical implementation (Stocker et al., 2022).

The concept of the IFDS builds on the FAIR principles by emphasising the integration of FAIR-aligned data and services within a unified digital ecosystem (Van Reisen et al., 2021). This approach extends the application of FAIR principles beyond scientific data assets to include the curation and integration of any data relevant to achieving FAIR operability. As such, the IFDS introduces the potential to enhance the quality and utility of patient data through FAIR-aligned practices. The implementation of the IFDS as part of the COVID-19 response leveraged the work of Babcock et al. (2021) on creating a linked ontology for infectious diseases.

Making large healthcare datasets available for federated reuse in research related to analytics and discovery is a novel challenge. Research environments that prioritise privacy-by-design for health data reuse, such as the Netherlands' Personal Health Train (PHT), are gaining popularity to support federated learning, aided by tools for data exchange, including the Fast Healthcare Interoperability Resources (FHIR) (Gebreslassie et al., 2023). Software tools that take data protection and privacy requirements into account from the start have the potential to address issues of access to sensitive patient health data (Zhang & Kamel Boulos, 2022).

Analytical framework

De Novo FAIRification, also known as FAIR-by-design, is a workflow where data is enhanced with semantic properties and made into machine-readable assets from the outset, during the creation of the digital data instance. This approach is proposed and followed by Jacobson et al. (2020) and Groenen et al. (2021). In contrast, other work follows FAIR by increment, implementing the FAIR principles retrospectively for legacy data. An example of FAIR by increment is the work by Smits et al. (2025), published in this volume.

Mature FSRs support FAIRification processes and facilitate convergence among implementers. That said, the documentation and exploration of FAIRification tools, techniques, and practices is underdeveloped. This study addresses the critical need for software tools that operationalise the FAIR principles, either by design or increment. While Jacobson et al. (2020) and Groenen et al. (2021) describe a FAIRification workflow which can provide the basis for a structured and generalisable approach to data management on patient data generated in clinical settings, their work only supports a FAIR-by-design approach and their proposed workflows have not been tested in the African context (Van Reisen, Stokmans, Basajja et al., 2020). The absence of a solution tested in various contexts and with legacy data collection processes in health facilities underscores the necessity of context-specific adaptations to achieve scalable and sustainable FAIR-compliant data systems, which are more cognisant of a situation where data has a legacy (Jacobsen et al., 2020; Basajja & Nambobi, 2022)

To enable FAIRification in various locations and countries across Africa, a new level of FAIRness, referred to as FAIR-Ownership of Data in Locale with Regulatory Compliance (OLR), was developed (Van Reisen et al., 2023). This approach identified the federated aspect of a FAIR architecture as fundamental. FAIR-OLR enables data to be owned by the entity responsible for its production, held locally, and maintained in compliance with regulatory requirements in the location where the data resides. The OLR principles strengthen the component of FAIR data sovereignty, acknowledging that FAIR data is fundamentally federated and AI-ready.

Workflow specifications and requirements

Data-driven platforms can facilitate making health data FAIR. The FAIR Guiding Principles serve as a framework for the data FAIRification process (Wilkinson et al., 2016). The workflows for their realisation by (Jacobsen et al., 2020) include the following steps: (1) identify the FAIRification objective, (2) analyse the data, (3) analyse the metadata, (4) define a semantic model for data (4a) and metadata (4b), (5) make data (5a) and metadata (5b) linkable, (6) host FAIR data, and (7) assess FAIR data. Jacobson et al. (2020) and Groenen et al. (2021) both recommended five phases of De Novo FAIRification: (1) the pre-FAIRification analysis, (2) facilitating FAIRification, (3) data collection, (4) FAIR data generation and (5) FAIR data use in the FAIRification of data from a registry.

The parameters for implementing the FAIRification workflow in this study were informed by several contextually relevant challenges identified by the VODAN research community, following the analysis of the first proof of concept for federated data reuse (Aktau et al., 2025; Van Reisen et al., 2021). The outcome of the assessment was translated into a set of parameters dictating the engineering process, referred to as specifications and requirements. These included:

- *One-time data input:* The data is curated at the point of creation.
- *Multi-pronged functionalities:* The data-interoperability and reuse algorithms are run over FAIR-curated data, enriched with semantic meaning and machine-actionable capabilities, allowing for multiple functionalities to be handled across FAIR data assets.
- *Federated storage*: The data is handled in the repository of the entity responsible for data quality, data control, and data processing.
- *Cross-country FAIRification:* The community applies FAIRification with curation processes that maximise FAIRness, including interoperability, across countries.
- *Data curation* to enhance value in the place where the data is produced and handled: Enhancing the value of the data for

stakeholders at the entity where the data is produced and managed is a primary goal of these investments.

- *Data provenance for data quality:* Clear data provenance is assigned in the curation process, which enhances the quality of the data and trust in the overall system.
- *Data visualisation in the place of data production:* Dashboards displaying insights from patient data in health clinics are crucial for the usability of the data in primary clinical processes.
- *Data value is retained in the place handling the data*: The VODAN research group found that ensuring data quality is managed at the point of creation helps retain its value at the original location (VODAN, 2021).

Platform development

The requirements and specifications formulated by VODAN (2021) served as the basis for the parameters used in engineering the platform's development and deployment. The development was carried out in three stages. In the first stage, the requirements and specifications (VODAN, 2021) were translated into the architectural and strategic development of an integrated environment that enabled the collection, processing, analysis, and sharing of data. The resulting scalable, interoperable, and secure infrastructure supported various data science workflows, tools, and stakeholders.

Key components of the platform included:

- *Data ingestion and integration:* Facilitating the production of data from multiple sources, in this case, 88 health facilities across 8 countries
- *Storage and management:* Establishing distributed data repositories for each health facility, controlled locally by the health facility, while ensuring accessibility
- *Processing and computation:* Supporting data transformation to metadata with clear data-mappings
- *Interoperability and FAIR principles:* Ensuring that data is FAIR across different systems and organisations

- *Security and compliance:* Implementation of governance policies, access control, encryption, and compliance with relevant regulations, including the GDPR and national data protection laws
- *Analytics and artificial intelligence (AI) deployment:* Providing tools for exploratory data analysis, visualisation, and machine learning (ML) deployment, as well as automated pipelines
- *User interface and collaboration:* Providing dashboards, application programming interfaces (APIs), and collaborative tools to support data scientists, analysts, and policymakers in implementing queries on the metadata

In the second phase, to the extent possible, available resources were identified to support the development of the platform. We tested various components and deployment approaches, with a focus on FSR, to support the specific challenges faced by health facilities in low-resource settings, as was the case for the initial VODAN Africa deployment.

In the third phase, the modus operandi for facility-specific deployment of federated data reuse was prepared. We analysed the implementation decisions that led to the specific digital setup at each facility, including how the deployment varied across facility characteristics.

Findings

In this section, we discuss: the design of the overall platform, the engineering of tools to make the platform operational with different elements connecting the components of the platform, and the facility-level adaptations for final deployment in varying contexts.

Facility-specific platform development

Based on requirements identified through structured conversations with different stakeholders at each health facility, FSRs were bundled to ensure availability at health facility locations. Each facility had a localised deployment of essential software systems, including the CEDAR Workbench, a bulk upload tool, an internal dashboard, and access to an external dashboard. To support knowledge graph storage

and capabilities, some facilities also installed AllegroGraph, a triple store, depending on service availability.

At the central level, an external dashboard was developed to which facilities periodically synced aggregate statistics. This central server hosted publicly available, aggregated data and additional platforms, including AllegroGraph, which enabled remote queries on de-identified, shared data triples using the SPARQL Query Language and the Resource Description Framework (RDF). The central infrastructure also supported federated analytics and federated learning. The main components of the platform are illustrated in Figure 1.

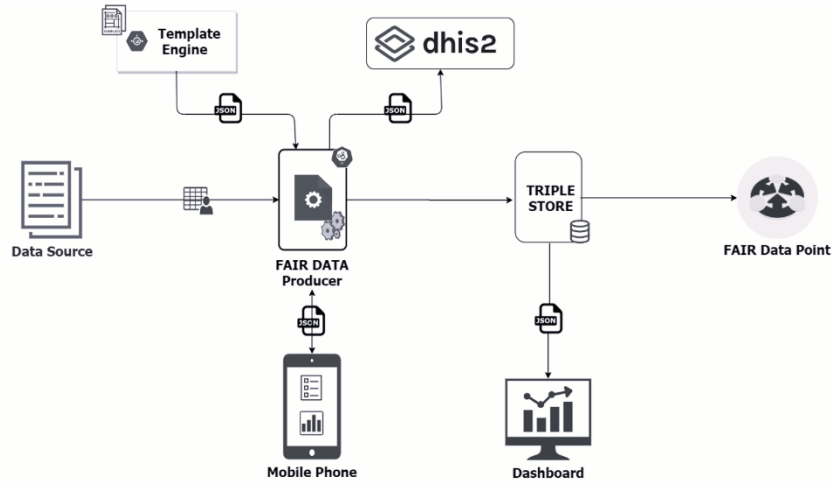


Figure 1. VODAN Africa architecture for FAIR infrastructure in health facilities

Source: Amare et al., 2023

The main components included a metadata processing component, a data curation component, a data visualisation component (internal clinic and VODAN community dashboards), query functionality using a triple store, and reporting components. The platform employed a modular approach, allowing each facility to deploy different components of the architecture.

Building on the experience of the VODAN research network, the Ubuntu operating system was selected as the platform host. Its open architecture facilitated the seamless integration of various systems,

enabling them to be packaged as a unified, easily installable, and user-friendly solution for data stewards.

To ensure the data were FAIR, a metadata template engine was implemented using the CEDAR Workbench. Once created, metadata templates were hosted in the FAIR data producer to enable FAIR data production. The data source provided input to the data producer. Data available in hardcopy were encoded into the FAIR data producer. A bulk upload tool was used when legacy data were only available in soft copy. Aggregate reports were computed and synced with DHIS2 to support required HMIS reporting, making the systems interoperable. Identified aggregates, based on data collected through the system, were displayed on a dashboard. The data and metadata were curated in the form of triples and stored in a triple store. Metadata were made available through a FAIR Data Point to foster findability and accessibility.

Templates for controlled vocabulary, semantic data model and linkage

To initiate the FAIRification process, data instances must be processed to link them to structured terminologies, allowing for the determination of relevant metadata labels. A vocabulary development process was carried out by undertaking a detailed analysis of HMIS paper-based forms. In most instances, this paper-based registry was digitised by the clinic's data clerk. The digitised instances serve as a source of data for the DHIS2 system, the production of which is, in most countries, a legal requirement for facilities by the relevant ministry of health.

The FAIRification of DHIS2 templates focused on OPD and ANC forms. In some countries, including Uganda and Ethiopia, these HMIS abstract registers were standardised across all clinics. In others, like Nigeria, variations existed within the country despite the ministry's efforts to standardise them. HMIS reporting forms also differed between countries. The DHIS2 template forms were compared on all the variables in the fields. Subsequently, for each form, each variable, along with its corresponding questions and responses, was meticulously examined to identify the relevant vocabulary that could be used for designing the metadata template in CEDAR.

CEDAR enables the creation of templates through an intuitive drag-and-drop, machine-actionable (meta)data collection template designer. The platform uses the CEDAR Workbench as a data and metadata template engine. For each unique form, a separate template was created in CEDAR. Each template was set up to create the metadata by selecting vocabularies for each field. A list of controlled and preferred vocabularies was established during the vocabulary identification and selection process to represent each field and enhance interoperability. The CEDAR system is linked to BioPortal, which allows it to utilise semantic ontologies in the design process of the data curation templates. BioPortal enables selection from other Common Data Model (CDM) standards and well-established clinical vocabularies, such as Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT). Within a given template, each field was linked to preferred semantic ontologies identified in BioPortal. CEDAR then automatically assigns a unique Uniform Resource Identifier (URI) to each field. The list of controlled vocabularies and preferred ontologies was compiled in an Excel spreadsheet, and the ontologies were subsequently reposted in the VODAN Africa portal for easy selection by other groups, encouraging FAIR convergence.

As shown in Figure 2, when new terms needed to be selected, they were created using an Excel sheet format based on the Simple Knowledge Organization System (SKOS) framework and subsequently transformed into RDF format via SKOS Play. RDF identifies a data-instance as a triple, an object-predicate-subject structure. The transformation of the data instance through SKOS resulted in an RDF Terse Triple Language (TTL), also known as Turtle, file. This results in a serialised description of the RDF triple graph of the instance, which identifies URIs and International Resource Identifiers (IRIs). This process links URIs and IRIs to value properties, which gives semantic meaning to the novel data instance. The novel triple was then uploaded to BioPortal to enable open ontology searches. It was further used for template creation in platforms such as CEDAR Workbench, Research Electronic Data Capture (REDCap), and other systems. The VODAN Africa portal, created based on OntoPortal, assembled the new ontologies. The new ontologies were also published on GitHub.

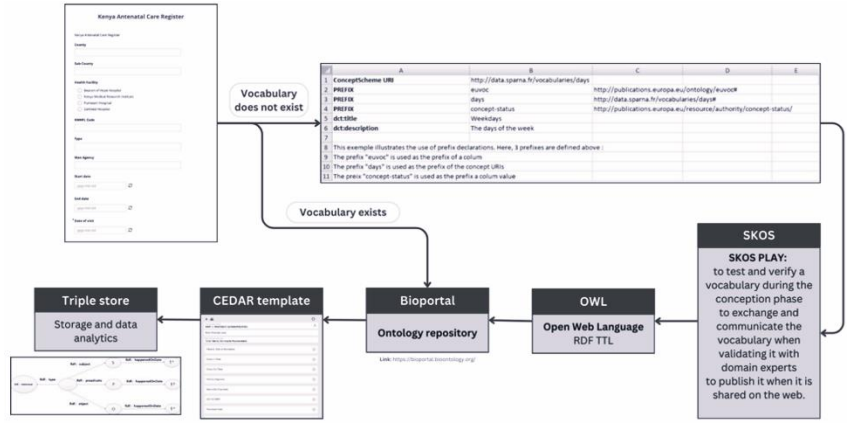


Figure 2. New vocabulary creation through CEDAR

Source: Haixia Li and Li Yan, 2021, used without modification

The template designer provided functionalities, such as auto-completion and drop-down menus, which allowed users to select from controlled vocabularies created earlier during data production. The template was used to populate data based on controlled vocabularies and semantic ontologies, and the output was encoded in JavaScript Object Notation for Linked Data (JSON-LD) as well as RDF formats.

The data were stored in a massive database, MongoDB, a document-based NoSQL (AKA 'not only SQL') database. Unlike traditional relational databases that use structured, rule-based tables, NoSQL databases store data in a non-tabular format. This approach supports high-volume data applications, content management systems, real-time analytics, Internet of Things (IoT) applications, and distributed applications that require high data availability and reliability. NoSQL database approaches offer more flexible and efficient query functionality compared to relational databases.

The platform prepared for each health facility was preloaded with templates tailored to the facility's specific requirements. This allowed for a controlled and standardised manner of collecting patient data reporting formats from the health facilities that was responsive to the health facility's specific needs while fostering interoperability in participant-level data and metadata at the cross-hospital level.

Repository as knowledge graphs

Through the semantic model and linkage, the linked data presented as knowledge graphs were placed in residence. For better provenance and ownership, the data were stored and made available for remote query in triple stores hosted in the health facilities. The data was stored in the form of RDF triples in a graph database named AllegroGraph. These data were made available for local and remote queries through a SPARQL endpoint. SPARQL is a query language and protocol used for retrieving and manipulating data stored in RDF format. It enables users to extract information from databases or any data sources that can be mapped to RDF. SPARQL is accessible via an API, allowing external or internal systems to integrate directly into the knowledge graph. Health facilities contribute to data and knowledge sharing by establishing appropriate linkages within their datasets using a well-defined semantic model, which serves as the foundation for the knowledge graph of the data available within and across facilities.

Access and reuse of FAIR data

Data access and control mechanisms were defined to enable authorised users and other systems to access the system through an API or by direct login. Access levels were defined at various levels, including the repository level and the single triple level.

The platform prepared for each health facility was preloaded with templates tailored to the facility's specific requirements. This allowed for a controlled and standardised manner of collecting patient data reporting formats from the health facilities.

A semantic data model was developed based on a systematic review of HMIS registers across various countries, considering local reporting requirements and data collection needs. Figure 3 provides a high-level overview of the development of the federated data reuse platform, including the FAIRification process.

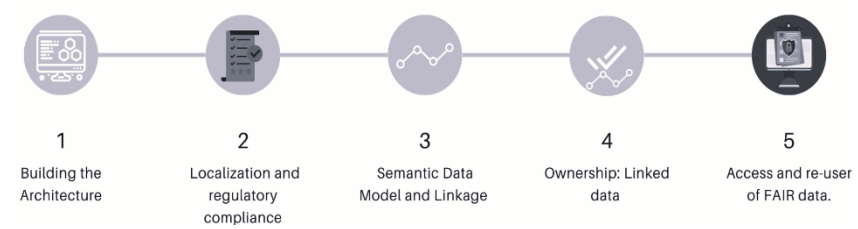


Figure 3. Platform development and FAIRification approaches followed in creating the VODAN Africa platform

Semantic data modelling plays a crucial role in enhancing data interoperability, enabling remote querying, and supporting comprehensive data analytics both within individual healthcare facilities and across multiple institutions.

To ensure data control and sovereignty as well as promote localised usage, data was required to be stored in residence at the point of production and primarily used within the originating facility. For this purpose, data was stored in the selected triple store tool, AllegroGraph. The process ensured that knowledge graph-based remote querying and analytics were supported, while the data remained within its original location. Access and control mechanisms were implemented at the port of the store in each facility, through data use agreements, and at the granular level, allowing permissions to be assigned at the level of individual triples. Additionally, the data was stored in other databases, including MongoDB, and was accessible through an internal dashboard. The internal dashboard was designed to enhance data ownership and usability for local stakeholders, providing them with direct access to data and actionable insights tailored to their specific needs.

Integration and development of platform components

To curate the FAIR (meta)data, CEDAR was used in the health clinics by installing it in the residence where the data was produced and curated. The registers were converted into CEDAR templates, which used controlled vocabularies during the process.

Bulk upload

To expedite the data curation process, it was necessary to upload the backlog data in bulk. Although the CEDAR system lacked the

functionality to upload data in bulk, its robust API allowed the research team to develop a tool that enabled bulk input. This tool enabled FAIR-by-increment, also known as post-hoc FAIRification. The bulk upload tool has integration points with CEDAR, AllegroGraph, and DHIS2. It utilises an API key from each of the applications and sends GET and POST requests. The bulk upload tool retrieves data from CEDAR, where the one-time data entry occurs, and posts it into AllegroGraph. The bulk upload tool uses the triple store API to convert Comma-Separated Values (CSV) legacy data into the knowledge graph.

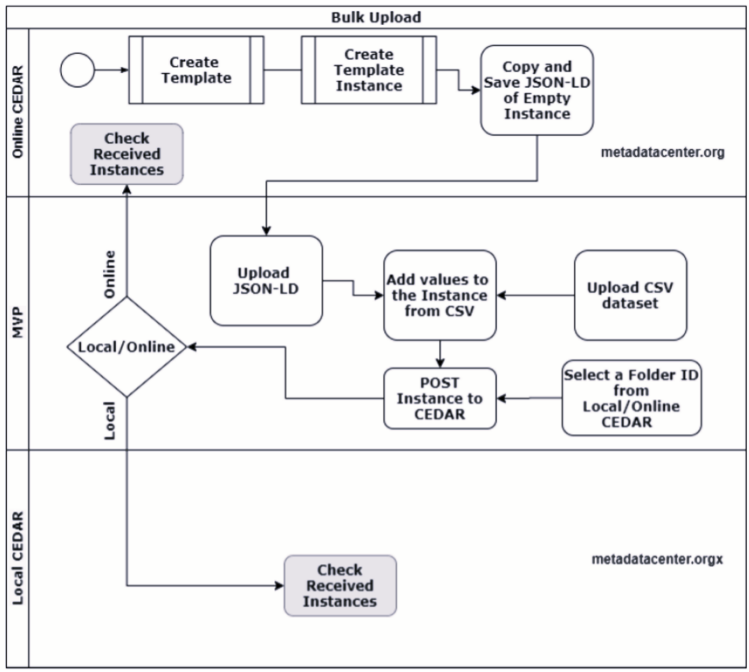


Figure 4. CSV file BulkUpload process of VODAN platform

Figure 4 shows how the VODAN platform can bulk upload data in CSV format to the local and remote deployments of CEDAR using its API.

The tool converts the data it reads from the database into triples. In cases where it reads from relational databases and CSV files, the conversion is performed by making the row ID/Object ID the 'subject', the column names the 'predicate', and the cell value the 'object'. In each of the subject-predicate-object triplets, the system

retrieves their equivalent URI or IRI by searching through the controlled vocabulary file stored in BioPortal or an offline file.

The system also receives requests from CEDAR and retrieves data for reporting purposes. The system does a computation to create an aggregate report. Through a POST request using the DHIS2 API, the report is posted in DHIS2. A JSON template used for reporting in DHIS2 was used to map the aggregate reports to the HMIS.

The bulk upload system also enables the direct posting of RDF triples created in CEDAR to AllegroGraph. To accommodate previously entered data in CEDAR, the system utilises Allegro Graph's API and makes a POST request to create the knowledge graph based on the data in CEDAR. The bulk uploaded data can also be converted into triples, similar to how data can be extracted from relational databases and then loaded into a triple store in real time.

The tool for bulk FAIRification was then enhanced to add more features, including the ability to bulk upload to a triple store such as AllegroGraph and backup and restore functionality. Other features requested by the health facilities and the data stewardship team, such as integration with DHIS2, were added. As shown in Figure 5, the bulk upload tool also features the ability to reuse templates created in the remote CEDAR (online version) for use in a locally deployed and fully offline CEDAR system. The CEDAR system, hosted in the cloud, enables users to create data and metadata templates that can then be shared with other cloud users.

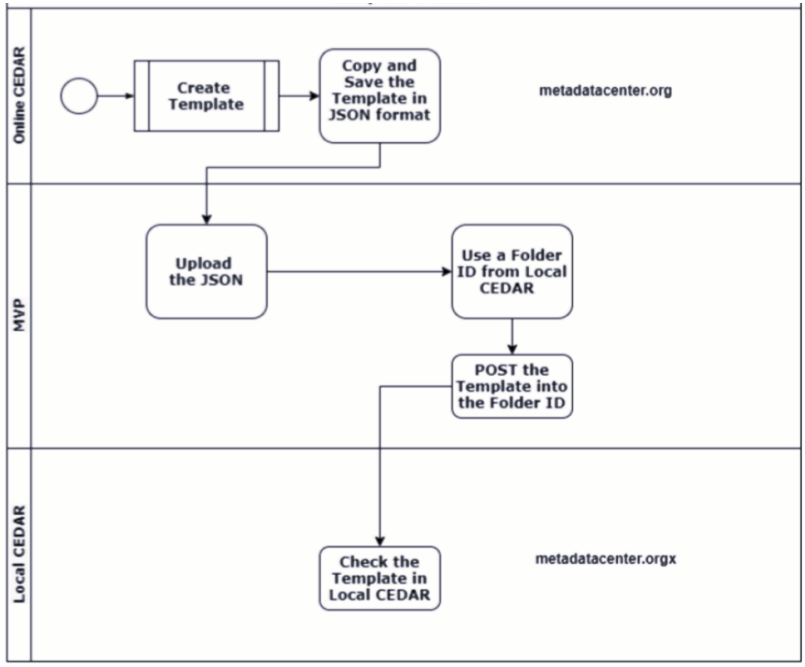


Figure 5. Tool to automatically load prepared templates from the central CEDAR platform onto the local installation

Initially, the localised version of CEDAR did not allow for the export of templates. The templates created by the data stewardship team had to be deployed on the localised systems, which meant recreating them. As there were many forms, creating a script in the system was more straightforward than migrating the templates to the local version. Using the bulk upload tool, a practical workaround was created. The template creation was done at metadatacenter.org and stored in JSON-LD format. We downloaded the templates from the remote CEDAR deployment and recreated them locally using an automated script that leveraged both APIs. This approach enabled a rapid rollout, which was practical given the large number of health facilities and templates needed. Alternatively, the JSON can be copied manually, saved locally, and synced with the installed tools.

Different countries had different needs based on their infrastructure. Tigray in Ethiopia, one of the implementation sites, had low connectivity and, in addition to having limited resources, experienced a total communication blackout and war during the implementation study. The FAIRification of existing research and clinical data was

carried out by preparing the data in a CSV file and uploading them using the bulk upload tool to CEDAR. The data were then visualised on the dashboards and could be queried using SPARQL.

CEDAR stores the data created via the templates in MongoDB in the form of JSON-LD. This data can be read directly by the system by adding the MongoDB extension to Laravel. The system was also made available as open-source software, providing a reusable FSR. MongoDB, a NoSQL database that utilises a document-based format and serves as a data source across various platforms, thereby enhancing interoperability through its schema-less design and widespread adoption for healthcare data.

Data visualisation component

In addition to data reuse for reporting requirements, healthcare workers wanted to utilise the data produced by facilities for clinical decision-making. VODAN Africa worked with local stakeholders to develop internal dashboards that reflected the specific clinical decision-making priorities for that health centre. The internal dashboard provided frequencies based on FAIR data generated by data stewards and health professionals, ensuring instant visual feedback on their work performance, including daily visits and other statistics chosen per discussions and agreements made through the data use agreement. In contrast to data curation activities prior to the initiation of the VODAN Africa federated data reuse platform, where data were produced for and reported to external groups, like ministries of health, through facilitating the reuse of their data for patient care, these dashboards posited clinics and hospitals as both data generators and the primary beneficiaries of data reuse, prompting them to produce more data of higher quality. Data were reused where they were initially produced, ensuring the appropriate distribution of benefits to data providers.

Aggregate statistics were shared and displayed on the external community dashboard, shown in Figure 6. The community dashboard refers to the external dashboard of VODAN, presenting aggregate statistics aggregated from data across participating hospitals to facilitate cross-facility inference at the regional, national, or cross-national level,

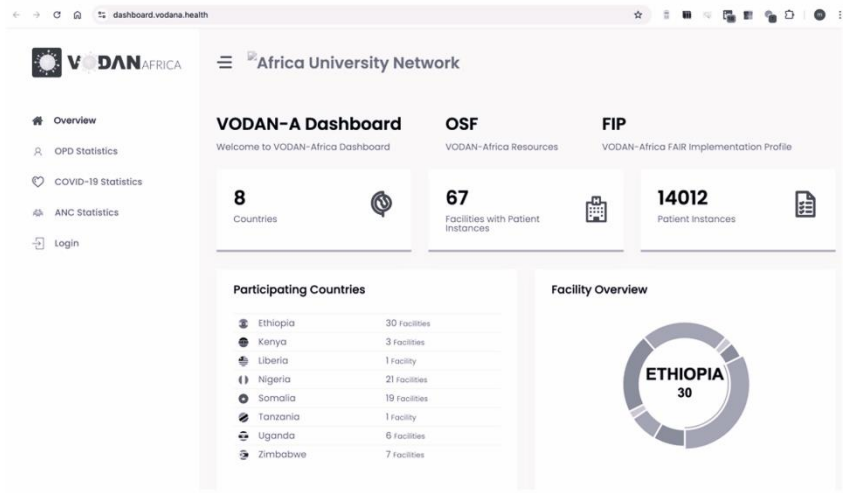


Figure 6. Screenshot from the external dashboard – the VODAN community dashboard

Source: https://vodana.health/ screenshot 23 March 2024 9:45

The VODAN community dashboard demonstrated the potential for sharing relevant information within the network and conducting cross-border showcases, highlighting the possibility of sharing information and performing surveillance, with data stored and in and under the control of the health facility. The system also features an API that allows interested parties to access and connect with the dashboard and the data system in VODAN Africa.

Remote queries component

The federated architecture ensures data production and use in a federated manner without compromising data provenance or its residency. The data is exposed as metadata in a linked data format, stored as subject-predicate-object triples within a knowledge graph. Remote queries can be performed on the RDF data hosted in health facilities using SPARQL. AllegroGraph was used to store the knowledge graph through which the query was being processed. AllegroGraph also allowed visually created queries and visualisation of the resulting knowledge graph and was used for the initial exploration of the clinical data that was also made available for research.

Terminology service

The use of ontologies facilitated the creation of templates and the data production process. When there were no standardised ontologies, the VODAN platform integrated with the BioPortal terminology server. Controlled vocabularies were publicly hosted in the portal and were accessed using an API based on the associated API key. Depending on the context, the vocabularies were either fetched directly through the API or stored for offline use in the facilities when Internet connections became an issue.

The integration code was written as part of the bulk upload tool. It utilises the API key associated with BioPortal to fetch data and metadata that would otherwise be inaccessible. The terminology service in CEDAR was natively integrated with vocabularies on BioPortal. The lightweight application and bulk upload system used the terminology server, which required a script to access the terms using a GET request from BioPortal, leveraging users' API keys.

Extract transform load component

Different scenarios can use the data extract, transform, and load (ETL) service to migrate legacy data into the knowledge graph. Data stored as files or from other legacy systems were transformed into triples and stored in the knowledge graph. The data extraction process for the My Structured Query Language (MySQL) server data followed the platform's interaction to read from the database, transform the data, and load it into both MongoDB in CEDAR and AllegroGraph.

Federated learning infrastructure

To enable federated learning, also known as remote data science, the required software infrastructure needed to be set up. Python libraries were installed in the computers used for FAIR data production in the facilities. To facilitate the installation process and accommodate new deployments, a remote installation script was developed. The PySyFt library was used for federated data reuse due to its free and open-source nature and since it has a strong, active, and responsive support community on Slack and other platforms.

HMIS linkage component

The data curated in the VODAN platform primarily came from abstract registers, which serve as inputs for HMIS reports. Health technology professionals, nurses, and other HMIS-related health workers use tally sheets to count and compile HMIS monthly and periodic reports. To facilitate the acceptance of FAIR data curation, the system was made interoperable with DHIS2, one of the most widely adopted reporting tools. VODAN programmes developed an automated extraction script that aggregated data into DHIS2 to automatically derive the required reports for health facilities. Figure 7 illustrates the interoperability of the VODAN platform with the standard HMIS system, DHIS2. The DHIS2 JSON template data from the VODAN platform was aggregated and submitted using its API.

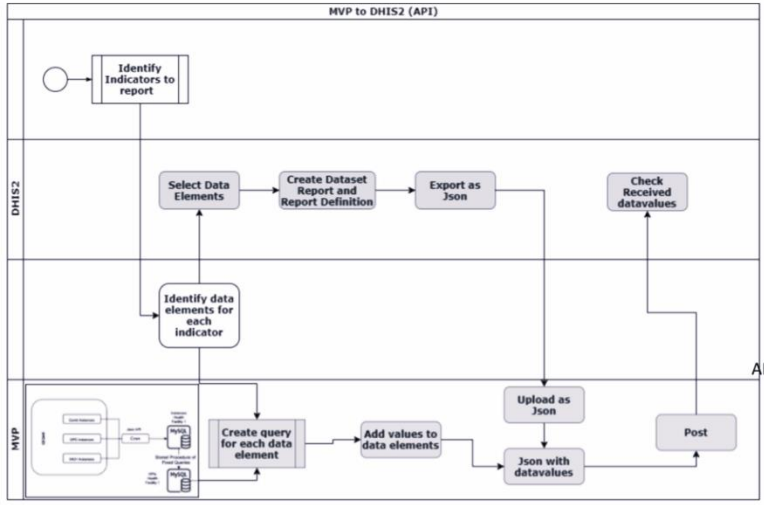


Figure 7. VODAN platform to DHIS2 interoperability workflow

DHIS2 JSON template data from the data storage component: Document database and triple store

The data storage component manages various data types and stores data locally. Data was stored in the form of CSV files and legacy relational database systems. Those datasets and newly created data were transformed and stored in the MongoDB document database as JSON-LD and as a knowledge graph in a triple store using AllegroGraph. Data storage localisation was one of the key principles

in FAIR data production and was implemented per FAIR-OLR principles.

Varying modus operandi for deployment

Different deployment and operational options were employed in various countries and sites within countries, depending on the availability of infrastructure and data stewardship expertise. In areas where capacity was limited and deployment needed to be accelerated, the system was created as a virtual image that included all necessary components. Although deployment was facilitated through the virtual image approach, performance was a challenge because the system only used the system portion allocated to the virtual machine.

We also tested the system's installation on a central server, following a client-server architecture. Large hospitals with numerous data collection points, where making the system available in each unit was not feasible, used the central server approach. Computers with low specifications and tablets were used to collect and store data on a central server at the health facilities. In low-resource settings with limited infrastructure and network accessibility, a lightweight system was developed and integrated into bulk upload tools. This system enabled the curation of data in the form of RDF triples, which were efficiently transferred to the triple store and MongoDB. Curated data were accessible through a dashboard, facilitating visualisation and real-time analyses.

Discussion

Through this implementation research, we explored the dimensions of applying FAIR principles to patient data for both healthcare-related surveillance and personalised care decisions in the context of antenatal care at the health facility, across health facilities, and at the single data entry level. Our work demonstrates the potential of federated health data reuse to address disparities and inequalities in health in dynamic, resource-limited contexts while preserving data sovereignty. We validated the VODAN Africa approach for regions under a digital blackout due to war. We demonstrated the utility of federated data reuse for cross-border infectious disease surveillance by performing complex queries on participant-level data securely retained by health facilities. This demonstrates the benefits of

federated reuse for regulatory compliance, data sovereignty, and real-time, patient-level, cross-border surveillance.

The implementation of the VODAN Africa approach across countries and facilities enabled an in-depth analysis of the design challenges and tooling requirements necessary to adapt cutting-edge technologies to low-digital-resource environments in Africa, at both large and small health facilities, in various regulatory environments. The findings contribute to the broader discourse on data sovereignty and provide valuable insights for future research on the equitable application of FAIR principles within under-resourced communities. The proposed architecture presented in this study, along with the practices followed to implement it within different data siloes in a health facility and across health facilities, offers another perspective on how to make FAIRification possible in low-resource settings.

Development of FAIR Supporting Resources

FSRs are rapidly emerging as critical components of data management and interoperability. Identifying and understanding FSRs requires a systematic investigation to document their current landscape and practical applications effectively.

Touré et al. (2023) used semantic web technologies and a fit-for-purpose FAIRification strategy within the Swiss Personalized Health Network to enhance data accessibility and usability. Their approach involved the development of federated infrastructures that enable permissioned access to health data as secondary sources for research. By leveraging RDF, they facilitated data aggregation and exploration, thereby improving the interoperability and utility of health datasets.

While ideally following the FAIR principles should make interoperability easier, there is still a need for harmonising FAIRification methods and best practices to maximise the value of FAIRification (Dos Santos Vieira et al., 2022). At the same time, there are no clear guidelines yet on how to measure how 'FAIR' levels of data are, also referred to as levels of FAIRness, and how to prioritise what to FAIRify and at what depth. Alharbi (2022) has usefully identified best-practice approaches to support the decision-making process.

Engineering ethnography

The development of complex systems with multiple interactive components necessitates a well-structured architectural framework to meet key functional and operational requirements (Garlan, 2000). Addressing wicked, complex problems requires a co-creative engineering approach that actively involves diverse stakeholders. In this regard, engineering ethnography provides an effective method for gathering stakeholder input to ensure that system design aligns with real-world needs (Van Reisen et al., 2021). The development of the FAIR-OLR-based patient data platform was guided by community input and expertise from health facility professionals, ensuring that the system aligned with local healthcare needs and regulatory requirements.

To implement research in real-world, diverse environments, it is essential to consider the ecological diversity of the problem (Van Reisen et al., 2021). This understanding facilitates a more targeted approach to problem identification and helps define the core components of an effective solution.

Earlier research by the VODAN Africa research group, which explored the core of the patient data handling problem, was informed by lessons from the Ebola crisis in Liberia. When the crisis was under control, all the patient data had been removed from the country and was no longer under the government's control. The analysis of this experience emphasised the necessity of localising data and systems to enhance efficiency, accessibility, and data sovereignty (Van Reisen et al., 2021).

The retention of data within health facilities while allowing access for analytical purposes has facilitated the availability of medical data for research. Traditionally, such data remained confined to patient care within healthcare institutions, typically in paper-based formats. Aggregate data was compiled for DHIS2 reporting requirements, but this digitalisation of DHIS2 data had no impact on the decision-making for patient care in the facilities; the insights from the DHIS2 data were generally unavailable in the health facilities and the data was not sufficiently granular to produce insights that were relevant to support health workers decisions. The stakeholders engaged with the

study reported that DHIS2 tooling itself lacked local sovereignty and ownership.

For ease of management, cost efficiency, and the aggregation needs of data system providers, many implementations continue to rely on relatively centralised computing models. In contrast to these assumptions, this research demonstrates that federated data infrastructures have significant potential for tracking, monitoring, and addressing clinical issues, while enabling data utilisation for research both during and after patient care. More research is needed to ensure the usability of the systems at the local level. The successful implementation of a federated data system requires consideration of enabling supporting structures, including substantial data stewardship programmes, infrastructure development, community engagement, and governance frameworks to ensure security, scalability, and ethical data use, as well as tools to handle access control and permissions for data access.

Theoretical insights for further research

The findings from the current study suggest that control over data must be complemented by control over the tools used to manage and process it, thereby ensuring greater data sovereignty and technological autonomy. In other words, this study found that control over data requires control over the tools to handle the data.

This part of the discussion addresses: the design of the overall platform; the engineering of tools to make the platform operational with different elements connecting the components of the platform; the blending of FAIR by design and FAIR by increment; and adaptation for final deployment in varying context situations.

Design of the overall platform based on data visiting

Having established the requirements and specifications set by the community, following its analysis of the core of the problem, the next phase of the study involved the technical engineering of localised tools and in-residence data management, reinforcing the principles of data ownership, interoperability, and regulatory compliance within the regional healthcare ecosystem. This architecture relies on the principle of data visiting.

In contrast to data sharing, which requires the transfer of data to a centralised location outside of health facilities, data visiting aligns with the FAIR and FAIR-OLR principles by ensuring that data remains accessible without compromising ownership, security, or regulatory compliance. Remote querying techniques such as SPARQL and federated data analytics are commonly used to enable data visiting in healthcare research, epidemiology, and cross-border collaborations.

Table 1. A comparison of data sharing and data visiting

Concept	Definition	Key characteristics
Data sharing	The process of transferring data from one entity to another often involves duplicating or moving datasets.	Data is physically copied, transferred, or downloaded. It requires agreements on data ownership, security, and compliance. Data sharing increases risks related to privacy, security breaches, and unauthorised use.
Data visiting	A model in which data remains at its original location, and external users send queries to access and analyse it remotely.	Data never moves; only results of queries are returned. Data visiting supports federated learning and remote analytics. It enhances data sovereignty, privacy, and regulatory compliance. It requires interoperable and secure infrastructures.

The key distinction between data sharing and data visiting lies in how data is accessed and controlled. Data visiting refers to a privacy-preserving approach in which data remains at its source and, instead of being transferred, researchers or algorithms visit the data through remote querying or federated analytics. This method is beneficial in sensitive domains such as healthcare, where data protection principles restrict direct data sharing.

Data visiting enhances data sovereignty and ensures that data remains under the control of the entity that owns them. It enhances security and privacy preservation by complying with data privacy regulations and mitigating the risks associated with data movement. It also supports federated learning, enabling collaborative research and AI model training without exposing sensitive, participant-level data.

Finally, data visiting reduces duplication and storage costs, as it eliminates the need for multiple copies of large datasets.

The implementation of a data-visiting framework and semantic interoperability across healthcare facilities has demonstrated the feasibility of integrating data-driven insights to enhance clinical decision-making for health workers at the point of care. Simultaneously, these frameworks facilitate collaboration at regional, national, and international levels, enabling the generation of broader health insights. We found that, in a data-visiting, federated framework, smaller healthcare facilities can leverage high-quality machine learning diagnostic models that are securely trained on large-scale datasets from major hospitals, thereby enhancing diagnostic accuracy and clinical decision-making in settings with limited data. These findings present a significant opportunity for future research, particularly in exploring how larger hospitals can provide relevant, data-driven insights to support decision-making in smaller clinics. Further investigation is required to fully realise the potential for these collaborative, AI-driven healthcare models.

Engineering of tools and standardisation of common data models

The federated reuse of participant-level data from disparate systems necessitates standardisation, which requires a common data model for data exchange, such as FHIR. In observational health research, standardisation enhances data quality and the efficiency of cross-database comparisons (Voss et al., 2015). Cross-standard collaboration, as between FHIR and OMOP/SNOMED, can further enhance such standardisation to strengthen interoperability within and across communities. This study demonstrated that the use of controlled vocabularies and semantic ontologies in triple-reading machine-readable RDF syntax facilitates the computation of sensitive data stored in a federated manner across facilities and countries. The study found that the standardisation workflow could be deployed in the African digital health data context.

Blending FAIR by design and FAIR by increment

While current FAIRification efforts often focus on the retrospective implementation of FAIR, this work focuses on De Novo FAIRification, also known as FAIR-by-design, blended with

techniques from FAIR-by-increment workflows. Existing datasets were FAIRified through bulk upload, demonstrating the possibility of integrating De Novo and incremental FAIRification workflows.

The combination of De Novo FAIRification workflows and FAIR by increment workflows developed in this project offers insights into how high-quality interoperable data can be delivered along with robust data provenance can be achieved. The blending demonstrates the potential for data handling in one place, where a single data entry can be used to achieve this. Integrating rich metadata templates during data production ensured structured and standardised data management. Data was stored in a triple store, enabling remote SPARQL queries, execution of analytics tasks, and implementation of multiple functional data processing workflows.

Federated data and systems can achieve interoperability by sharing structured knowledge through a knowledge graph. In this research, knowledge graphs were constructed from FAIR-compliant patient data, which included both De Novo FAIRified datasets and legacy data that had been incrementally FAIRified.

Conditions for deployment in varying context situations

Graph databases serve as a foundational technology for storing knowledge graphs, offering more efficient data merging and mapping compared to relational databases, which require rigorous referential integrity management. In relational databases, data integration relies on the use of primary and foreign keys, as well as an understanding of the cardinality of relationships between tables. In contrast, triple stores store data in the form of triples, which can be easily appended and combined with minimal effort, enhancing scalability and flexibility in data integration.

Document databases such as MongoDB were employed during the FAIRification process due to their real-time data storage and retrieval capabilities. MongoDB's schema-flexible structure facilitates efficient data storage, access, and scalability, making it well-suited for handling heterogeneous datasets (Chauhan & Bansal, 2017).

The platform was designed to facilitate interoperability and support the generation of new knowledge through inference and the utilisation of relationships in linked data using various data analytics

tools. Remote access via distributed queries and federated analytics enabled data utilisation while adhering to data protection regulations, although further refinement of these mechanisms is required. To fully leverage these capabilities, it is essential to implement robust security measures and comprehensive data privacy protections to ensure compliance and safeguard sensitive information.

Conclusion

Grounded in the FAIR principles and the need for Localization, Ownership, and Regulatory Compliance (FAIR-OLR), the VODAN-Africa research group developed and implemented a system capable of curating and managing FAIR data. This system was deployed across 74 health facilities in Africa, enabling the production and use of FAIR data at the point of care and service delivery.

This study allowed for an examination of the successes and challenges of software development in environments that differ significantly from mainstream implementation settings, such as those in Europe and the United States, where FAIR principles are widely promoted. Best practices for FAIR biomedical data remain disproportionately skewed toward Western nations, with low- and middle-income countries (LMICs) underrepresented in the FAIR implementation landscape (Bezuidenhout, 2020).

The study highlighted the successful deployment of a federated architecture for managing sensitive patient data, localised and retained in-residence, across eight African countries. The research explored the development of an overarching platform, the integration and enhancement of its components, and its subsequent deployment to realise De Novo FAIRification. Each of the 74 health facilities in which the system was implemented had its own distinct data entry forms prepared for one-time data input. This approach ensured the production of FAIR data enriched with semantic and machine-readable features.

Conducted as an ethnographic study of engineering implementation, the research objective was to test the feasibility of using the FAIR Guidelines for patient data curation. The implementation focused on data curation for ANC, OPD, and COVID-19 by the data clerks at participating health facilities. A dashboard was provided to each

health facility to generate insights from the data curated in the health facility. A cross-country dashboard showed the surveillance insights generated through federated data reuse from participant repositories. The surveillance was conducted across borders, in full compliance with specific regulations in each location and under the control of the health facility.

This experience provided valuable insights into the tools and methodologies developed during the implementation process. Importantly, the success highlighted the potential of federated architectures, supported by lean microservices, suitable for low-resource settings. The result of this research offers documented practices that can be adopted globally and locally by other implementers. It laid the groundwork for federated data curation, infrastructure, and data pipelines.

Various adaptations were documented to address the unique resource constraints and regulatory frameworks of the different countries. This is the first step towards establishing a health data environment for a Personal Health Train setup, supported by federated AI and machine learning in Africa. This research has the potential to inform the development of FAIRification processes in various settings, adapted to local contextual conditions.

Acknowledgments

Sincere thanks to the administrators and medical staff, health workers and data stewards in the health facilities that assisted in conducting this study. The study was implemented in low-resource conditions and the research team is so grateful for accepting to participate in this study.

The researcher express their gratitude to **Prof Mirjam van Reisen** for supervising the overall research and editing the different drafts and provided comments on all drafts and for her mentorship and guidance across the whole journey. The researcher also is grateful to **Dr Araya Medhanyie** as he assisted in the public health domain underlying and suggestions on the approach and analysis of the study. He also provided comments to the various drafts of the publication. The researcher also expresses their gratitude to **Getu Tadelle, Tesfit Gebremeskil, Ruduan Benjamin Franklin Plug**, and **Abdulahi Kawu** for the technical engineering contributions during implementation.

The researcher would also like to thank the Country Coordinators in Nigeria, Uganda and Kenya for their support to the implementation of the study.

Authors' Contributions

Samson Yohannes Amare is the main researcher of this study who conceptualised the research, wrote the research and implementation plan, set up the approach and conducted all main elements of this study.

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

Aksünger, N., De Sanctis, T., Waiyaiya, E., van Doeveren, R., van der Graaf, M., & Janssens, W. (2022). What prevents pregnant women from adhering to the continuum of maternal care? Evidence on interrelated mechanisms from a cohort study in Kenya. *BMJ Open, 12*(1), e050670. https://doi.org/10.1136/bmjopen-2021-050670

Aktau, A., Amare, S. Y., Van Reisen, M., Taye, G. T., Gebremeskel, T. G., Jati, P. H., & Plug, R. (2025). GO TRAIN: A protocol for metadata creation for the FAIRification of patient data health records. 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. xx–xx). Bamenda: Langaa

Alharbi, E. A. (2022). *The practices, costs and benefits of FAIR implementation in pharmaceutical research and development*. The University of Manchester (United Kingdom).

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

Babcock, S., Beverley, J., Cowell, L. G. & Smith, B. (2021). The Infectious Disease Ontology in the age of COVID-19. *Journal of Biomedical Semantics, 12*(13). https://doi.org/10.1186/s13326-021-00245-1

Basajja, M. & Nambobi, M. (2022). Information Streams in Health Facilities: The Case of Uganda. Data Intelligence 2022; 4 (4): 882–898. doi: https://doi.org/10.1162/dint_a_00177

Bezuidenhout, L. (2020). Being fair about the design of FAIR data standards. *Digital Government: Research and Practice, 1*(3), 1-7.

Chauhan, D., & Bansal, K. L. (2017). Using the advantages of NOSQL: a case study on MongoDB. *International Journal on Recent and Innovation Trends in Computing and Communication, 5*(2), 90-93.

Dos Santos Vieira, B., Bernabé, C.H., Zhang, S. et al. (2022). Towards FAIRification of sensitive and fragmented rare disease patient data: challenges and solutions in European reference network registries. *Orphanet J Rare Dis 17,* 436. https://doi.org/10.1186/s13023-022-02558-5

FAIRConnect. (No date). FAIR Supporting Resources. *FAIRConnect*. Retrieved March 20, 2025, from https://fairconnect.pro/fair-supportingresources/#:~:text=A%20FAIR%20Supporting%20Resource%20(FSR,which%20includes%20metadata%20about%20it.

Garlan, D. (2000, May). Software architecture: a roadmap. In *Proceedings of the Conference on the Future of Software Engineering* (pp. 91-101).

Gebreslassie, T. G., Van Reisen, M., Amare, S. Y., Taye, G. T., & Plug, R. (2023). FHIR4FAIR: Leveraging FHIR in health data FAIRfication process: In the case of VODAN-A. In: FAIR Connect, 1(1), 49-54. IOS Press. DOI: https://doi.org/10.3233/FC-230504

Gregurick, S. (2020, January 16). *NIH's strategic vision for data science: Enabling a FAIR-data ecosystem for HEAL* [Presentation]. National Institutes of Health.

Groenen, K. H. J., Jacobsen, A., Kersloot, M. G., dos Santos Vieira, B., van Enckevort, E., Kaliyaperumal, R., Arts, D. L., 't Hoen, P. A. C., Cornet, R., Roos, M., & Schultze Kool, L. (2021). The de novo FAIRification process of a registry for vascular anomalies. *Orphanet Journal of Rare Diseases, 16,* Article 376. https://doi.org/10.1186/s13023-021-02004-y

Guillot, P., Bøgsted, M., & Vesteghem, C. (2023). FAIR sharing of health data: A systematic review of applicable solutions. *Health and Technology, 13*(6), 869-882.

Haixia Li and Li Yan. (2021). A Temporal RDF Model for Multi-grained Time Information Modeling. In 2021 4th *International Conference on Data Science and Information Technology (DSIT 2021), July 23-25, 2021, Shanghai, China.*

Jochems, A., Deist, T. M., van Soest, J., Eble, M., Bulens, P., Coucke, P., Dries, W., Lambin, P., & Dekker, A. (2016). Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. Radiotherapy and Oncology, 121(3), 459–467. https://doi.org/10.1016/j.radonc.2016.10.002

Jacobsen, A., Kaliyaperumal, R., Bonino da Silva Santos, L. O., 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

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). https://doi.org/10.1162/dint_a_00169

May, C. (2006). Escaping the TRIPs' trap: The political economy of free and open source software in Africa. *Political studies, 54*(1), 123–146.

Neumark, T., & Prince, R. J. (2021). Digital health in East Africa: innovation, experimentation and the market. *Global Policy, 12,* 65-74. https://doi.org/10.1111/1758-5899.12990

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

Sanders, E. & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *CoDesign*, 4(1), 5-18. https://doi.org/10.1080/15710880701875068

Schultes, E., & Wittenburg, P. (2019). FAIR Principles and Digital Objects: Accelerating convergence on a data infrastructure. In *Data Analytics and Management in Data Intensive Domains: 20th International Conference, DAMDID/RCDL 2018, Moscow, Russia, October 9–12, 2018,*

Revised Selected Papers (Vol. 20, pp. 3–16). Springer International Publishing.

Smits, K., Upase, M., Pandey, N., Bala, S., & Van Reisen, M. (2025). FAIR-data implementation for analysis of research data in human trafficking and migration. 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

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)*. https://doi.org/10.1162/dint_a_00168

Strawn, G. O. (2021). 75 years of astonishing evolution of IT: 1946–2021. IT Professional, 23(3), 21–27. https://doi.org/10.1109/MITP.2021.3061997

Touré, V., Krauss, P., Gnodtke, K. et al. (2023). FAIRification of health-related data using semantic web technologies in the Swiss Personalized Health Network. *Sci Data 10, 127 (2023)*. https://doi.org/10.1038/s41597-023-02028-y

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). *FAIR Connect, 1*(1), 1–7. 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). 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., Jati, P. H. P., 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). 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: *Data Intelligence 2(2020), 264–275*. doi: https://doi.org/10.1162/dint_a_00049 (eds. Mons, B., Jacobsen, A. & Schultes, E.). MIT Press, http://www.mitpressjournals.org/dint

Van Reisen, M., Stokmans, M., Mawere, M., Basajja, M., Ong'ayo, A. O., Nakazibwe, P., Kirkpatrick, C., & Chindoza, K. (2020). FAIR

practices in Africa. *Data Intelligence, 2*(1–2), 246–256. https://doi.org/10.1162/dint_a_00047

Van Reisen, M., Yohannes, S., Plug, R., Tadele, G., Gebremeskel, T., Kawu, A. A., Smits, K., Woldu, L. M., Stocker, J., Heddema, F., Folorunso, S., Kievit, R., & Medhanyie, A. A. (2024). Curation of federated patient data: A proposed landscape for the Africa health data space. In A. L. Imoize, M. S. Obaidat, & H. Song (Eds.), *Federated learning for digital healthcare systems* (Intelligent Data-Centric Systems series). Elsevier. https://doi.org/10.1016/B978-0-443-13897-3.00013-8

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

Voss, E. A., Makadia, R., Matcho, A., Ma, Q., Knoll, C., Schuemie, M., DeFalco, F. J., Londhe, A., Zhu, V., & Ryan, P. B. (2015). Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association, 22*(3), 553–564. https://doi.org/10.1093/jamia/ocu023

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, P., & Kamel Boulos, M. N. (2022). Privacy-by-design environments for large-scale health research and federated learning from data. *International Journal of Environmental Research and Public Health, 19(19), 11876*.