Secure distributed machine learning in healthcare: a study on FAIR, compliance and cybersecurity for federated learning

Plug, R.B.F.

Part I

Article 1

Note on Prior Publication

This chapter is equivalent to the following published journal article for which I am the first author:

Plug, R., Liang, Y., Aktau, A., Basajja, M., Oladipo, F. O., van Reisen, M. (2022). *Terminology for a FAIR Framework for the Virus Outbreak Data Network-Africa*. Data Intelligence 4(4), 698–723. MIT Press. `doi: 10.1162/ dint_a_00167`

I contributed to the conception, main writing, terminology development, FAIR knowledge base, and regulatory framing in the VODAN-Africa healthcare context. My co-authors supported domain-specific terminology refinement, contextual validation across African clinical settings and editing.

Terminology for a FAIR Framework for the Virus Outbreak Data Network-Africa

Ruduan Plug[1], Yan Liang[1], Aliya Aktau[1], Mariam Basajja[1],
Francisca Oladipo[2], Mirjam van Reisen[1,3]

[1] Leiden University, 2311 EZ Leiden, the Netherlands

[2] Kampala International University, 260101 Kampala, Uganda

[3] Leiden University Medical Centre (LUMC), 2333 ZG Leiden, the Netherlands

Abstract

The field of health data management poses unique challenges in relation to data ownership, the privacy of data subjects, and the reusability of data. The FAIR Data Principles have been developed to address these challenges. To apply the FAIR Principles in Africa, an underlying problem is identified in that data extraction without consent by the person or community to whom the data pertains, may lead to mistrust regarding data use by third parties. The Virus Outbreak Data Network (VODAN)-Africa architecture builds on the FAIR principles, using the General Data Protection Regulation (GDPR) framework to ensure compliance with local data regulations, while using information knowledge management concepts to further improve data provenance and interoperability. The essence of the initiative is to combine Data Ownership and Data Visiting as lead requirements for an interoperable health data system. This article discusses the terminology used in the field of FAIR data management, with a specific focus on FAIR-compliant health information management, as implemented in the VODAN-Africa architecture.

Keywords: Data Management, Federated Data, Data Governance, FAIR Guidelines, FAIR Data and Services, FAIR Data Point, FAIR Framework

Acronyms

CEDAR	Center for Expanded Data Annotation and Retrieval
DMP	Data Management Plan
ETL	Extract, Transform, and Load
EU	European Union
FAIR	Findability, Accessibility, Interoperability, Reusability
FDP	FAIR Data Point
HMIS	Health Management Information System
IN	Implementation Network
KPI	Key Performance Indicator
OWL	Web Ontology Language
RDF	Resource Description Framework
VODAN	Virus Outbreak Data Network

2.1 Introduction

Data management has become one of the prime factors of concern in contemporary research in all fields of research. The volume and velocity of data is rapidly increasing, causing serious bottlenecks in data processing, storage and reusability. To tackle this issue, a multimodal process that advances the human-data relationship may offer a viable approach [1]. This is achieved by developing theoretical frameworks for automated data management and technological architectures that distribute data, as well as expanding human expertise.

However, these developments towards automated data processing pose numerous challenges, from the perspective of both society [2] and technology [3]. These challenges are magnified in the field of health, where privacy, security and patient-data ownership are critical concerns. Coincidentally, these data typically contain vital, yet untapped, information for the advancement of scientific research. Health data is by definition personal data, which may contain sensitive and personal information. The Universal Declaration of Human Rights

(1948) states that “No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence” [4]; therefore, personal data protection is enshrined within the foundation of international law.

The VODAN-Africa initiative, guided by the FAIR guidelines, provides a framework that addresses these concerns through a multimodal approach to data management and data stewardship [5]. By developing an architecture in which data is **F**indable, **A**ccessible (under well-defined conditions), **I**nteroperable and **R**eusable (FAIR), we may address technical concerns using modern metadata processing techniques, while data stewardship empowers scientific communities with expertise to interact with these data across their field in a meaningful way.

The way we deal with medical data within VODAN-Africa is inherently distributed in order to provide data sovereignty, however, there are concerns over the convergence between localised instances. To reconcile such localised instances with a common vocabulary, in this article we have developed a set of shared terminologies that allow for the unambiguous exchange of controlled vocabularies and development of consistent data stewardship expertise.

This article investigates and reviews the basic concepts and terminology in the context of the Virus Outbreak Data Network (VODAN) and specifically the VODAN-Africa implementation, established as an Implementation Network (IN) under the GO-FAIR initiative jointly with FAIR IN Africa. The VODAN-Africa initiative has been established as a pilot deployment to produce clinical patient data, which is by nature sensitive data. Important is the full retention of data ownership in residence, through data-visiting, and recognising the fragmented nature of the regulatory frameworks applicable in each locale.

This article sets out to review how data terminology can be defined in the context of health data management, for the investigation of VODAN-Africa. In addition we seek to facilitate further investigation of FAIR-based clinical patient data generation, processing and analytics within distributed and federated healthcare data applications.

2.2 Data concepts

To develop our terminology framework, first we thoroughly build upon the core terminologies used in the process of data management. The first concepts we need to develop for our framework are ‘data’, ‘information’ and ‘knowledge’ [6] as they are procured within a clinical setting. In the perspective of this framework, we start with unprocessed data, which are the first elements we encounter in the operational sphere in the data stack.

Metaphorically, data can be seen as the technological equivalent to the stimuli humans receive through their senses. These stimuli are raw bits of information and, before they are processed in the brain, are not attached to any meaning. Similarly, data entered in a computer, either through automated recording or human data entry, does not have any meaning until it is compartmentalised and processed. From the clinical perspective, meaning is central to subsequent application of data, which is defined through biosemantics.

Data A set of numeric values, characters and/or symbols.

The definition of data is very broad and includes both ordered and unordered data. In practice, the vast majority of data originates from observation, such as observational patient data, and is initially unstructured. To provide data with meaning, we need to process the data in accordance to standardised methods of formalisation. The three most common forms of data processing are: (1) select or sample the data relevant to the purpose by filtering, (2) compartmentalise data into separate attributes, and (3) provide an index to the data (i.e., a time-stamp, identifier, numeric ordering) [7].

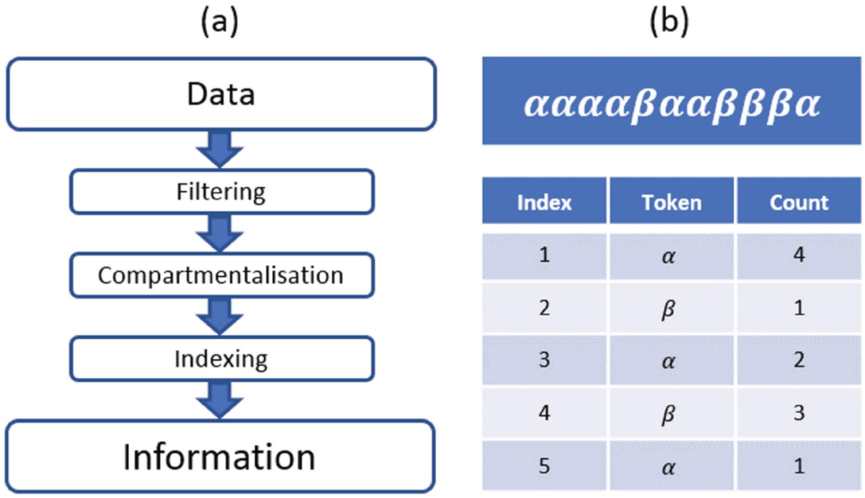
All the techniques that structure and give meaning to data are considered data processing techniques. The simplest example of this employed at VODAN-Africa is ad-hoc data processing with composite forms based on controlled vocabularies, in which the structure of the form indicates the assignment of entered data to specific attributes under specified conditions.

Information Data that has been structured and processed in such a way that meaning has been assigned to it, which can be interpreted and from which analyses can be drawn.

The process of transforming data into information involves giving structure to the data, which is primarily aimed at making the data suitable for human interpretability and machine interoperability. These processes can be either performed manually, i.e., by assigning certain data to a type or attribute field, or by automated methods based on ontology specifications.

An example of this can be found in the transcription of written medical documents. A digital image of a medical form consists of nothing but raw pixel values that can be rendered on a screen. In this context, the machine is not inherently able to determine whether or not a certain group of pixels has a specific meaning. We can, thus, state that the semantics of such an image cannot be directly derived by a machine from the raw data.

However, these data can be transcribed by human annotators, given they possess such domain knowledge. In the medical field this is traditionally performed by clinicians, but many such tasks can be performed data clerks and data stewards after training, which are extensively involved within VODAN-Africa. By gathering the data from the form, they can be entered into appropriate attribute fields in a digital format.



Index	Token	Count
1	α	4
2	β	1
3	α	2
4	β	3
5	α	1

Fig. 2.1.: (a) Flowchart indicating the generalised process to transform data towards information. (b) Example of data (top) and possible resulting information (bottom) (Plug, R. 2021)

In this way, the human annotator has assigned meaning to the visual data, based on their existing knowledge, and transformed these data into a structured format, which is information that can be used by both humans and machines without requiring additional context. These processes can also be automated, in this example optical character recognition (OCR) may be used to extract the characters, numbers and letters from the form – but these technologies typically fail to compartmentalise data further, are prone to error, requiring manual review and possess no accountability that data stewards have. While both methods produce information, the information is unequal in specificity and granularity [8].

Another factor we have to consider when processing data is that relationships may exist between data or derived information. There are many types of relationships that can exist between data and the type of relationship can depend on the type of data. For example, two numerical attributes may be correlated or one attribute may be associated with, or causal of, another attribute.

This is critical in the context of sensitive data processed in VODAN-Africa, as crucial to localised data methods are the context and meaning of these data. Analysing data in isolation may remove context, and thus meaning. Appropriate metadata and semantics, in the form of provenance, may be key to preserve these relationships when deidentification is applied to sensitive data.

By mapping the relationships between the information we have extracted from the data, we are transforming information into knowledge [7, 8], which is one of the primary methods used in VODAN-Africa. Knowledge typically takes the form of a graph representation, in which nodes identify instances that have attributes and the edges indicate relationships between such instances. This type of graph structure can be visualised for human interpretation, as well as traversed by computational algorithms for a process we consider knowledge discovery [9].

> **Knowledge** A tectonic description of information and the interconnected relationships between elements of information.

A widely used methodology to represent knowledge is the Resource Description Framework (RDF) [10]. This is a data structure framework that implements a machine interoperable language to represent semantic graphs. In this context, each node is a URI specifying a resource with associated attributes, and each edge is a directional relationship between two resources. The combination of the URI and the locale can be employed to produce globally unique identifiers when accessing and querying metadata across different services, which is important to enable unambiguous data access within VODAN-Africa.

As relational descriptions in RDF are primarily used for machine interoperability, and through linkages compatible with JSON data produced by non-relational health databases, they have no spatial structure. The visualisation of these graphs in complex relational schemas is non-trivial [11], but an RDF-based knowledge representation provides a very powerful machine interpretable data structure that can be readily used for relational knowledge discovery [12], which is one of the core aims from the knowledge base developed within VODAN-Africa.

Knowledge Discovery The derivation of new relational properties in a knowledge graph, based on the properties of the graph structure.

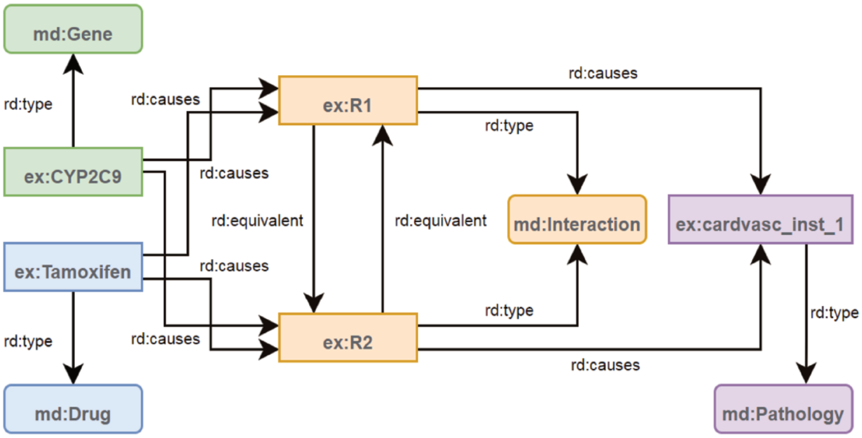
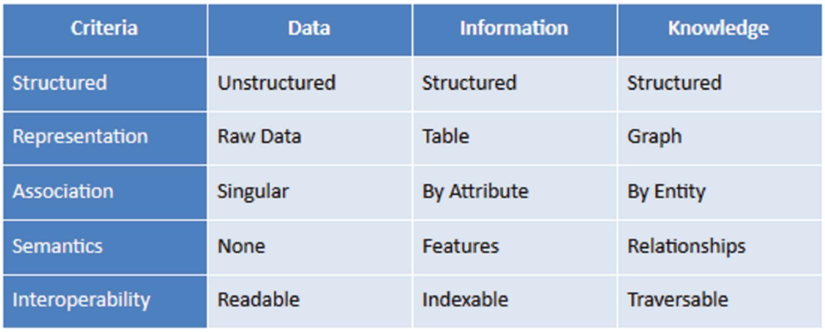


Fig. 2.2.: An example of an RDF graph for drug-gene interaction using knowledge discovery; equivalent interactions R1 and R2 have been associated with rd:equivalent (Plug, R. 2021)

Thus far, we have described the framework that incorporates data to produce information and knowledge graphs. The motivation behind this process is twofold: both to incorporate the domain-specific meaning of the data and to provide machine interoperability. The most important properties of these three core terminologies that will be used to develop the FAIR health data management framework are listed in Table 1.

As we have discussed in the previous section, the core principle underlying the transformation of data into information and knowledge is the attribution of meaning to the data. As meaning is fundamentally a philosophical concept, we need a formalised methodology to ascribe meaning to data.



Criteria	Data	Information	Knowledge
Structured	Unstructured	Structured	Structured
Representation	Raw Data	Table	Graph
Association	Singular	By Attribute	By Entity
Semantics	None	Features	Relationships
Interoperability	Readable	Indexable	Traversable

Fig. 2.3.: Properties of data, information and knowledge (Plug, R. 2021)

These formalisations are shaped by metadata, which in epistemology designates the self-referential denomination of data with respect to data [13]. The conceptual foundation of this formalisation is that meaning can be structured as data; for example, in the form of a description or a caption. These data can be used in reference to other data to attach meaning; in the above example a caption could be attached to an image to provide meaning to the image.

As a consequence, we can thus derive that metadata are the building blocks that allow us to transform data into information and knowledge [14]. For us to transform data into information, we have to specify metadata that conveys context over the particular data. Likewise, transforming information into knowledge requires the production of metadata that specifies the relationship between elements of information. In other words, metadata form the mechanism that provides a link to the insights with respect to the semantics of the data, primarily in facilitating information seeking, retrieval, understanding and use [13].

Metadata Data that describes other data in order to convey information that guides understanding, specificity, retrieval and interoperability.

Herein also lies the fundamental problem: with self-reference, there is always the risk of unresolved or inconsistent references. This is problematic in some complex data sets, where the metadata itself may require references to the data to convey its meaning. Another issue is that without some form of domain standardisation across an implementation network like VODAN, the meaning of the metadata may be ambiguous or unspecified [15].

To standardise metadata, we define different types of metadata based on the objective that is associated with the denotation [14]. To illustrate the paradigm, some metadata may be produced to aid human understanding, while other metadata describe properties for machine interoperability. We define three main archetypes of metadata, which form the building blocks of our data management framework in VODAN-AFRICA.

The first type of metadata we consider is metadata that is centred around human understanding, providing descriptions of or annotations about data. This type of contextual metadata provides the link between machine interoperable data and human interpretability.

Contextual Metadata Metadata that provides descriptions about data to aid human understanding.

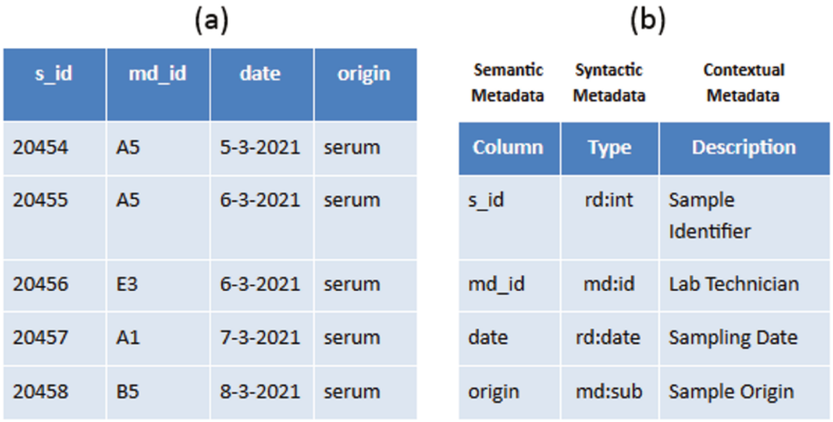
The next type of metadata we discuss is focused on the machine interpretability of the data – or what we consider the syntactic metadata, which provides information about the format of the data, the way the data should be operated on and the way the data is structured. Being able to specify the syntactic format of data is essential in cross-machine interoperability.

Syntactic Metadata Metadata that provides structural specifications about data to aid machine interoperability.

Finally, we consider semantic metadata that specifies the meaning of data, which is the broadest concept for which metadata can be produced [16]. These metadata define the broad context, and may be used to specify unique identifiers and link different concepts or data together. These metadata are central to the structure of interlinked data and form the building blocks of the concept of the semantic web, as proposed by Berners-Lee [17]. Semantic metadata is central to frameworks such as RDF to represent knowledge graphs [10] and the Web Ontology Language (OWL) [18], which is used to formalise knowledge representations [18] that are used by clinicians and implemented by data stewards in VODAN-AFRICA.

Semantic Metadata Metadata that associates objective meaning with the data in relation to other data.

An operational example of how these three types of metadata work in conjunction with one another in medical data records is provided in Table 2. The metadata in this table supports the entered data, such that the individual data points can be isolated using the semantic metadata, the data is interoperable due to the syntactic metadata providing instructions for machine interpretation, and the contextual metadata provides annotations on the relationship of the data to domain-specific knowledge.



(a)

s_id	md_id	date	origin
20454	A5	5-3-2021	serum
20455	A5	6-3-2021	serum
20456	E3	6-3-2021	serum
20457	A1	7-3-2021	serum
20458	B5	8-3-2021	serum

(b)

Semantic Metadata	Syntactic Metadata	Contextual Metadata
Column	Type	Description
s_id	rd:int	Sample Identifier
md_id	md:id	Lab Technician
date	rd:date	Sampling Date
origin	md:sub	Sample Origin

Fig. 2.4.: (a) Example data produced for the given metadata using a controlled vocabulary. (b) Metadata as data, describing the properties of the various metadata (Plug, R. 2021)

As shown in Table 2, what constitutes metadata cannot always be inferred simply by considering the attribute values. The table to the left (a) shows the classical example, where the metadata is structured as semantic metadata. These metadata provide a structural specification about the meaning of each different attribute in the data, in which each row is a uniquely indexed record in the table, which forms an essential part of the VODAN-AFRICA URI.

On the other hand, we can also construe metadata as the records themselves, as shown in the table to the right (b). We consider this synergy of ‘metadata as data’ [19], in which for each semantic identifier we also have the syntactic and contextual metadata associated with that semantic concept. The composite of these three elements forms the complete metadata specification of a particular concept in the information or knowledge specification of our domain, which formalises the data generation and traversal throughout VODAN-AFRICA.

Metadata Specification The complete specification of all metadata associated with a concept within a domain.

As there are potentially uncountable different methods by which metadata can be specified for linked concepts, a standardisation process is typical used within domain-specific knowledge bases [13, 15]. The baseline of VODAN-Africa community standardisation is expressed through the use of agreed-upon vocabularies, defined as controlled vocabularies, which limits the potential set of concepts to a finite and enumerable set.

Vocabulary A finite set of terms and symbols derived from expressions within a domain.

As vocabularies may continuously change and evolve as new concepts are generated by domain experts within VODAN-Africa, there is the inherent prospect that the vocabulary itself may become ambiguous. For example, in the case of synonyms, where two terms are linked to the same concept, or in the case of homonyms, where a single term may be linked to multiple concepts in a controlled vocabulary [20]. To maintain the specificity and integrity of the knowledge base, it is important that such ambiguities are avoided by using lemmatised concepts across VODAN-Africa in order to achieve convergence within the knowledge framework. For instance, if two research facilities use a different terminology for the same concept, it is important that these terminologies are grouped together as a single lemma, instead of being treated as separate entities for purposes of convergence within health communities.

In order to achieve this within VODAN-Africa, a centralised, controlled vocabulary can be used. These vocabularies are organised in such a way as to optimise the knowledge base, minimise ambiguities and streamline data retrieval in relational entity-based knowledge bases [21]. The controlled vocabulary consists of a curated list of terms used to transform information into knowledge, by associating these terms as metadata to convey the specification, links and descriptors of unique conceptual entities.

Controlled Vocabulary A curated set of terms and symbols from which concepts and relations between concepts can be expressed.

We can further specify this by formalising the method we use to structure a controlled vocabulary by the means of specified grammars to form an ontology [22]. These grammars define the way that terms within the controlled vocabulary can be used together. For instance, in a medical ontology we may choose that a phenotype expression can only be linked to an instance of a gene, but not to an instance of a pharmacological compound. By formally defining these constraints, we can ensure, by using an ontology, that only semantically valid, and uniquely identified, knowledge is created as a product of input data.

Ontology A domain-specific language from which knowledge can be represented as the product of a controlled vocabulary and semantic rules governed by formal grammar.

A concept that arises from the use of ontologies is that of templating metadata, which is an essential element of VODAN-Africa-wide data formalisation. As ontologies control for both the vocabulary and grammar of the knowledge base, any data entered within the knowledge base should belong to an entity within that knowledge base [22]. This limits the metadata that may be associated with data, and this can be expressed by constraining the metadata to a template format that controls for terms and semantic properties.

Metadata Template A set of semantically valid, domain-specific metadata specifications derived from constraints as specified by an ontology.

By using metadata templates in VODAN-Africa, which are produced from the domain ontology, we can standardise the way that products of data, information and knowledge are represented within an information system, in this instance a health information system. The standardisation of terms and semantics defined by metadata is a core element in producing data that is interoperable and reusable, and is key in the process of knowledge discovery.

2.3 FAIR health data management

As health facilities have started collecting more data about physiology, pharmacology and treatment efficacy, there has been an increasing need for the digitisation of health data to keep these increases in data volume manageable and usable. This is especially relevant to digitalisation in VODAN-Africa, across which a multitude of health facilities have thus far operated on manual data entry or handwritten patient records. Eysenbach describes these digitisation efforts as e-health, representing the relationship between medicine and computers and how this combination can benefit the healthcare and pharmacological industries [23].

However, because of the rapid development of data collection and healthcare information technologies, the academic definition of e-health extends to include the enhancement of health services and information supported by the onset of relevant technologies. This can be represented as the development and application of digital technologies in the field of medicine [24] in efforts to improve interoperability. Examples of health information in e-heath are patients' electronic health records (EHRs), genomic data, digital prescription, and even extending to remote diagnostics, each of which are data encompassed in VODAN-Africa.

Care facilities frequently use health key performance indicators (KPIs), over which VODAN-Africa bases the key analytical factors unique to each locale. These are employed to compare their performance to that of other care facilities, which makes it particular relevant in cross-facility analytics and knowledge exchange. Specially, to identify areas for improvement, and in addition, KPIs can be correlated with measures directly related to treatment efficacy within local context. For instance, average hospital stay and outpatient rate are some of the commonly used healthcare KPIs within VODAN-Africa, measured for various treatment types [25].

Healthcare key performance indicators (KPIs) A well-defined performance metric that is used to track, analyse, improve, and transform all essential healthcare operations in order to enhance patient satisfaction.

Different KPIs may be recognised at different levels of healthcare in VODAN-Africa, which addresses health from both a clinical as well as population level. From the perspective of a nation we are most interested in metrics such as life expectancy, while at the clinic level treatment outcomes and patient turnaround are critical. One of the primary issues that the VODAN-Africa implementation addresses is the need for both in residence and aggregate analytics, using a specifically designed data management framework [**23**, 26].

Data in residence Data produced and stored at a research institute or at the point-of-care, used to enable and enhance healthcare and scientific research, as well as to perform analytics.

The data that is present in residence within VODAN-Africa is stored in local database architectures, which are defined as data repositories, driven by local ownership [27]. The repository is the technical implementation of the system that collects, aggregates, manages and stores data in residence. What differentiates the repository from a standardised database is that the repository also maintains services for generating and maintaining domain specific ontologies, pooled from a central controlled vocabulary, and knowledge bases to support data management and access.

Data repository The point of storage and management of all data, information and knowledge relating to the primary purpose of a facility.

These operations, and the underlying operations performing these transactions, are part of a larger architecture, which we consider a health management information system (HMIS). Most of the current HMIS implementations within Africa are proprietary [27], which is a large drawback that VODAN-Africa seeks to address. Typical HMIS form the layer between the end-user (e.g., researchers and health professionals) and the data repository [28]. This allows for the management of access levels and interfacing directly with other applications that are used within departments of a healthcare facility.

Health management information system (HMIS) A system for entering, storing, maintaining, retrieving, and processing health data stored in repositories. Provides functionality to aid in the planning, management, and decision-making processes of healthcare institutions.

Two processes that are primarily monitored by a HMIS are data integrity and data quality, which are critical to the operation of a health facility. Within VODAN-Africa, data quality is maintained through provenance, rich metadata and domain specific accuracy measures, while data integrity is maintained by means of data redundancy and strictly regulated access and control patterns [29].

Data integration can be considered one of the main data management processes in operating an HMIS, which represents the process of combining data from various data sources into a single, unified and cohesive dataset with the purpose of supporting users with the consistent data access and delivery [30]. When consolidating healthcare data into a health information system, there are some challenges involved in the processing pipeline, which impose constraints on accessing data, the retention of data quality, and validation of data consistency [31]. The FAIR framework provides a workable solution to these issues through the accessibility and interoperability specifications, which in case of VODAN-Africa are transparent and locale-dependent [29].

Not all healthcare metadata are case-specific. There are some common data elements through VODAN-Africa such as patient age, gender, and marital status that are common in a lot of clinical datasets from the multitude of healthcare systems. Common domain specific data elements also exist in health metadata and are defined in biomedical ontologies, specified by the VODAN-Africa community. These describe common clinical metadata, which can be used to transform data to a common VODAN-Africa format, as well as for secondary analysis.

Common data elements (CDEs) Standardised terms or concepts that can be used or shared with other healthcare and research institutions as controlled vocabularies or ontologies for clinical research.

When doing clinical research, the data management plan (DMP) plays an important role. After the proposal stage and before the funding stage, the DMP helps researchers to organise the use of data and includes data management and data analysis during and after the research. In addition, it is a critical component in validating whether or not the data management process is compliant with local data regulations [32].

Data management plan (DMP) A formal written document that outlines the process for accessing or producing data; the standards for managing, describing, and storing data; and the system for handling and protecting the data during and after research.

The process specification involved in a DMP helps researchers to manage the research data specification and requirements, which in total specifies the data lifecycle [33]. Data lifecycle phases typically include data collection, data storage, data usage, data archiving and, finally, data destruction. For a viable DMP the entire process must be well-defined.

Data lifecycle An overview of all the stages of data existence from its production, storage, use, and reuse to destruction.

The process of data generation involves measuring or acquiring data according to a pre-specified collection protocol. While this process can differ across locales in VODAN-Africa, the steps afterwards are standardised [29]. After the data creation stage, the data must be stored and protected with different security levels within the organisation, based on specification and regulation. In the data usage phase, data can be read, analysed, manipulated, edited, and saved. Data archiving stores data as a backup without additional maintenance. Finally, data destruction removes the data from the repository, ensuring, from a security and privacy perspective, that the data can no longer be restored or subsequently used.

Contemporary data, information and knowledge management in healthcare and research faces emerging and ever-increasing difficulties in dealing with the challenges posed by big data [34]. Simple increases in computational performance, storage capacity and algorithm efficiency alone are not enough to handle the magnitude of data that is being generated [2]. For this reason, the were conceptualised by Wilkinson et al. [1], consisting of four foundational principles, namely: Findability, Accessibility, Interoperability, and Reusability.

These principles were developed in order to improve data management and stewardship and ensure transparency, reproducibility, and reusability for digital assets that contain not only data, but also related algorithms, tools and workflows [1]. These are the key principles that are used throughout the VODAN-Africa implementation of the VODAN-Africa health data management architecture.

The primary requirement of FAIR compliance with respect to data management, is the baseline specification for data discoverability through the concept of findability. For data to be findable, there must be a well-documented path to index, organise and query data through the use of unambiguously readable metadata and traversable knowledge graphs, defined by a standards-driven ontology specification.

Findable Health Data Health data should be discoverable by humans and machines through the use of metadata and data linkages defined by biomedical ontologies.

Once data has been properly indexed and integrated into a health information system for findability, there must be a well-specified method to perform a repository query. At the point of data access, typically implemented by an application programming interface (API), data queries are handled under well-defined conditions, such as methods of authorisation and credential verification audited by data stewards either in residence or at the MoH.

Accessible Health Data Data, information or knowledge in residence should be accessible, possibly in anonymised format, under well-defined and transparent authorisation conditions.

A critical component that revolves around the findability and accessibility of health data is the machine interoperability of the data throughout VODAN-Africa. For this, a baseline requirement is that the ontology, produced from the central controlled vocabulary, must be resolvable by all locales and the unique identifiers associated with the metadata must be unique.

The representation of knowledge, and the entity-attributed metadata through templating, must be interpretable by automated evaluation to make the underlying data machine-actionable. From the perspective of formal graph representation, this means that the knowledge graph that is implemented must be well connected. Semantic metadata that is not referenced or indexed by the health system is not operable, as the data pertaining to these metadata are not findable through automated methods in the repository.

Interoperable Health Data Health knowledge bases should be interlinked and operable for secure, automated data processing, storage and analysis across health facilities.

Through interoperability, by making the health data architecture well-specified, resolvable and machine-actionable, the conditions under which data becomes reusable are expressed in a formal framework. Interoperability throughout VODAN-Africa allows for techniques such as automated knowledge discovery [35] to maximise the information and knowledge that can be extracted from existing data, or combinations of old and new data.

For the reuse of data to comply with data protection regulations, it is essential that the reposited data within VODAN-Africa remains in good provenance by maintaining all associated metadata specified in the DMP. In addition, the laws of each VODAN-Africa locale under which accessibility is regulated must be well-documented, and both data and metadata has to be provided with a specification describing the conditions under which access may be provided.

Reusable Health Data Health data should be in good provenance, with documented metadata to allow for the replication or reuse of data across health facilities and locales.

The architecture of VODAN-Africa has been designed as a FAIR ecosystem, in which every aspect has been specified with the as key design elements. This is aimed at achieving the primary objective, which is to support the transnational reusability of medical (research) data and the exchange of knowledge, while maintaining data sovereignty [36].

Data sovereignty Data is reposited at the place of production, where full data ownership is retained and data is subject to local laws and regulations.

By keeping data in residence in VODAN-Africa, and maintaining the rights of the data owner, data controllers and processors work under the local laws and regulations of the jurisdiction. This ensures that the rights of the data subject, the person or community to whom the data pertain, are always maintained in accordance with the government processes influenced by local constituents. A key problem that hampers data reusability and the exchange of knowledge, is the lack of a framework in which data can be exchanged or used under controlled conditions outside the jurisdiction. This requires the architecture of VODAN-Africa to be inherently distributed. From the perspective of data localisation, each of the data repositories within the network form individual FAIR Data Points (FDPs) [26] that are compliant with GDPR [37] and further regulated under the data protection laws of the locale. Within the network, FDPs represent the individual repositories where data is both controlled and processed using FAIR compliant health management processes.

FAIR Data Point (FDP) A local data repository and accompanying services that are compliant with the FAIR Principles.

The design of this network is specified in the design of a FAIR digital health infrastructure by van Reisen et al. [29], where communication between FDPs is integrated in the Internet of FAIR Data and Services (IFDS) through the concept of data visiting. Conceptually, data visiting involves the provision of aggregate and inferential data, produced from the original data in residence at each of the FDPs, without exposing the actual data records. This allows for a robust, distributed community analytics framework, where meta-analyses can be performed on VODAN-Africa aggregate data while retaining full data sovereignty and is, thus, also compliant with regulatory frameworks in regard to privacy and data protection.

Data visiting Retrieval of aggregate analyses or statistics from a FAIR Data Point, where analysis processing is fully performed at the repository and no underlying data is exposed.

This ecosystem is defined as the Internet of FAIR Data and Services, where FAIR data is produced and interacted with through FAIR services, which interface through FDPs. To establish the process of data visiting within this ecosystem, unambiguous resource identification is required. These resources are conceptualised in a digital object model, where each resource has a unique identifier that is persistent as well as resolvable [38].

Unique, persistent and resolvable identifier (UPRI) A unique, persistent and resolvable identifier for digital objects.

A FAIR compliant system to support the data processing and management of the VODAN-Africa FDPs is implemented at the Center for Expanded Data Annotation and Retrieval (CEDAR) [39], which is responsible for the management of the ontologies, knowledge bases and all activities related to FAIR-based data processing. This provides individual facilities in VODAN-Africa with tools to perform both data controlling and data processing without requiring external parties, based on controlled vocabularies that are agreed upon through community and stakeholder driven decision making. The comprehensive implementation defined as the FDP, implemented as a repository managed by CEDAR with services that provide a data visiting interface, forms the central unit within the VODAN-Africa architecture.

2.4 Jurisdiction and data governance

The question of data ownership is both a legal and philosophical challenge, which plays a central role in VODAN-Africa. As data is non-tangible, from a legal standpoint data may be interpreted as intellectual property. However, some data are ‘matter of fact’, to which no rights can be attributed [40]. This is further complicated by the question of who the true legal owner of data is, and whether or not it is even possible to identify the legal owner of data, to which provenance plays a key role. Each of these matters may depend on the jurisdiction in which the data is produced and the geospatial location where the data is physically stored.

Data ownership The individual or party that has full control and legal rights over specified data, and who can, therefore, define the terms pertaining to access to and control of the data.

A baseline principle that must always be upheld for data governance in cross-national instances is data provenance [41]. Data is said to be in good provenance when meta-causality is upheld, i.e., the origin and the processes that generated the data are known and well-documented through a clear data-lineage. From the perspective of VODAN-Africa, provenance is a critical element for the data to have meaning in the place where it was produced, which increases its relevance, but also serves as a way to measure the data’s veracity. The quality of provenance in data is critical to an investigation of the environmental interactions of data in the context in which it was generated. Not only by locale, but also by data subject cluster. With proper data provenance, the question of data ownership can be addressed by means of identifying whom the subject of the data is, if applicable, and the party that initially collected or sampled the data.

Data provenance The documentation and updating of the origin and the processes that generated the data.

Apart from concerns about data ownership in VODAN-Africa, there are also legal and ethical concerns surrounding both collecting and storing data. Most of these legal concerns are focused on the privacy of subjects [42], which is further driven by the rapidly increasing scope and variety of the medical data that is being collected on individuals since the SARS-CoV-2 pandemic [43]. Data are by definition heterogeneous, as such different types of data may warrant different levels of legal protection. Medical data typically warrants the highest level of legal protection, due to the sensitive nature of such information [44, 45], which is one of the main concerns that VODAN-Africa stakeholders have [29]. In order to investigate the potential relevance of a FAIR-based health information system, the VODAN-Africa researchers performed FAIR-Equivalency analyses. These are analyses in which core policy documents of a particular locale or sovereign political entity are identified and investigated on the equivalency of their ambitions with FAIR-equivalent principles. The method is based on the fifteen facets of FAIR [29].

FAIR Equivalency A measure of the equivalency in ambitions expressed in a certain place regarding data supported E-health with the fifteen facets of the FAIR Guiding Principles.

The legal concerns surrounding the handling and storing of data are placed within the perspective of the jurisdiction in which the data resides. The legal policies and standards that are in place within a jurisdiction fall under the data governance and regulatory framework, which aim to standardise the way data is handled according to the applicable laws and regulations [46].

Data governance The enactment of regulations and policies surrounding the collection, handling and storage of data as well as the authorisation management of cross-border data flows.

When designing an information management system that can be localised, it is essential that it is compatible with the different modes of data governance – as in the applicable laws and regulations surrounding data in the place where it is produced. One approach that may be taken is an open source approach, where localisation is performed by manually customising every aspect of the implementation to comply with regulations. An information management system across different geographies requires that it be flexible to handle regulatory fragmentation across locales, as each implementation may use radically different methodologies to comply with the terms of the jurisdiction it operates under. Security of the data held in residence requires that a clear strategy is mobilised for encrypted data back-up on hard drives disconnected from the system whilst fully respecting the principle of localisation.

Data localisation The practice of repositing data (and their back-ups) at the location where the data has been produced.

An implementation of this is to use of ethnographic design principles across VODAN-Africa. Within the community that seeks convergence on an information system, all stakeholders representing each different locale are actively participating in the design and development process. This approach promotes transparency and allows for agreed-upon solutions to issues when differences in laws and regulations are identified. Through a participatory and collaborative ethnographic process, an implementation is created that provides a baseline for all stakeholders and well-documented options for divergence from the baseline when needed for any practical or regulatory reason.

Ethnographic design A participatory collaborative design principle that aims to satisfy the requirements of cross-national stakeholders.

At the centre of a participatory and collaborative ethnographic design is transparency about the process and implementation. As both data collection and data analysis are becoming increasingly complex and 'black-box', there is an increased need for conspicuousness when it comes to the intermediate processes by which data is stored and archived [47].

A step further is the concept of a completely transparent information system, in which non-sensitive data is anonymised and published in an interoperable and reusable manner. Such a concept is implemented in the European Open Science Cloud (EOSC) [48], while upholding the same principles with regards to ethnographic design and full-scale interoperability [49].

In relation to legal concepts regarding data, information and knowledge management, we use the General Data Protection Regulation (GDPR) as the foundational legislative frame of reference [50, 51]. The GDPR, as a framework, revolves around transnational legislation for increasing operational transparency, promoting integrity, necessitating confidentiality and specifying the constraints of data processing. This applies to personal data, which is data that pertains to a natural person and over which the natural person should have control.

Personal data Any data, information or directly resulting knowledge that relates to, and legally belongs to, the data subject (Article 4(1), GDPR).

At the centre of the GDPR framework is the legal arbitration between the data owner, data controller and data processor. While data ownership, as we have previously defined, pertains to the party that has control and legal obligation over a specified set of data, under the GDPR we fully recognise the rights of the individual from whom data has been collected. As VODAN-Africa provides full data provenance, this becomes feasible to implement over the entire implementation network. As a consequence, we assume the individual from whom data has been drawn to retain full ownership over their data, while another party may process or control data under strict guidelines. These guidelines are only exempt under documented derogations that are jurisdiction-specific, and typically cover matters of security, defence, public security and the judicial process (Article 23(1), GDPR) which overrule, by local means, the conditions defined by VODAN-Africa stakeholders.

For instance, medical data that has been collected to perform toxicological tests are sensitive in nature. While these data are stored and operated by the medical facility, from a data protection regulation framework perspective the data subject still has full legal rights over the data and the facility requires legal permission to use and store these data, unless a legal exemption clause was signed. Exemptions in relation to data ownership, such as data used for scientific research, are subject to strict regulations and typically require a DMP that involves a process of pseudonymisation or the anonymisation of the data to protect the data subject. The aggregation process, such as that used in VODAN-Africa, depersonalises data and as such they no longer pertain to a specific data subject and are thus not personal data.

Data subject A natural person about whom data has been collected and who can be identified, directly or indirectly, by reference to that data (Article 4(1), GDPR).

We consider here the difference between ‘data objects’, which we consider any non-human entity from which data can be sampled, as compared to ‘data subjects’, a term that exclusively covers data relating to a natural person. From the perspective of the data collector and regulator within VODAN-Africa, we can relate this to data from which we can, directly or indirectly, identify any natural person. In this instance, the data collector does not have full legal rights over the data, rather the rights remain with the data subject who needs to give exclusive and sole permission for data to be stored and used, which requires findability as a baseline property.

The conditional requirements under which a data subject may be able to provide permissions over their personal data falls under the GDPR, which stipulates that the data subject can only provide consent if given full information about the processing and use of their personal data. These conditions are typically given by the domain experts, that drive the semantic and purpose of data within VODAN-Africa.

This underlines the importance of data provenance in the implementation of an information system that holds data about data subjects. It is of critical importance to maintain well-documented contextual metadata that specifies the ownership of the data, the conditions under which the data may be used or processed, and the extent of the consent that has been provided by the data subject. It should also be noted that under the GDPR, consent can be withdrawn at any time and the data subject has the right to request a record of the personal data, as defined under right of access, as well as to have personal data erased.

Informed consent The voluntarily given, specific and unambiguous consent given by a data subject who is informed of all available data processing activities (Article 4(11), GDPR).

From the perspective of medical data processing, such as that performed in residence or in medical repositories, we are dealing with special categories of personal data. If a non-privileged party wishes to process these data in VODAN-Africa, they must receive explicit consent for every single purpose that the data will be used for and local regulations can impose limitations on the permissions that a data subject may give to other parties over special categories of personal data. As VODAN-Africa has a wide variety of legislative frameworks, these limitations may vary, but not be more permissive than the implementation.

There are exemptions for certified public services that require more permissive data processing capabilities to function, such as MoHs associated to VODAN-Africa. These categories allow for secure processing and storage under professional secrecy, by certified individuals, under strict conditions stipulated by the national regulating body, without receiving explicit consent (Article 9(3), GDPR). Examples in VODAN-Africa are if the processing and controlling of data is necessary for medical diagnosis, occupational medicine, provisional healthcare or the management of healthcare systems by individuals under non-disclosure.

Special categories of personal data Sensitive personal data that are subject to strict regulations, which may only be processed and used by legally certified parties (Article 9(1–3), GDPR).

In addition to the data subject, we identify two entities in VODAN-Africa that may handle personal data: the clinician as data controller and the data steward as data processors. The data controller is the contingent that is given the right to control personal data belonging to a data subject, which is typically provided through informed consent. The controller determines the conditions, purpose and means by which personal data is stored and used by the data processor. Under these conditions, from the perspective of medical data management, the data controller is typically the residence at which the data has been produced.

Data controller The entity that specifies the purpose for, and the means by, which personal data belonging to a data subject is processed (Article 24(1–3), GDPR).

The controller of the data is legally responsible for acquiring consent or legal permission, and providing a statement of purpose and DMP. The controller does not need to be a singular entity. Multiple organisations, such as VODAN-Africa, may form a group that jointly determines and states the purpose and conditions under which data may be stored and processed while complying with the GDPR guidelines.

While clinicians as controllers specify the purpose and means under which data is handled, the data steward as data processor is the party responsible for processing and storing the data on behalf of the data controller. It is the responsibility of the data processor to implement a data repositing process with sufficient security measures and the ability to certify the integrity and security of personal data that is stored within the locale. Potential security risks and measures taken to minimise these risks have to be documented in a data protection impact assessment (DPIA) report (Article 35(1), GDPR).

Data protection impact assessment (DPIA) Potential security risks and measures taken to minimise these risks have to be documented in a data protection impact assessment report (Article 35(7), GDPR).

The data controller and the data processor may in some instances be the same entity, for instance, a small clinic where medical professionals process data. However, data processing is typically covered by a specialised party, for example, a cloud service provider, that is contracted by the data controller. All responsibilities, legal obligations and non-disclosure stipulations must be documented in a contract between data controller and data processor.

Data processor The entity that is responsible for processing the complete lifecycle of the personal data belonging to a data subject on behalf of the data controller (Article 28(3), GDPR).

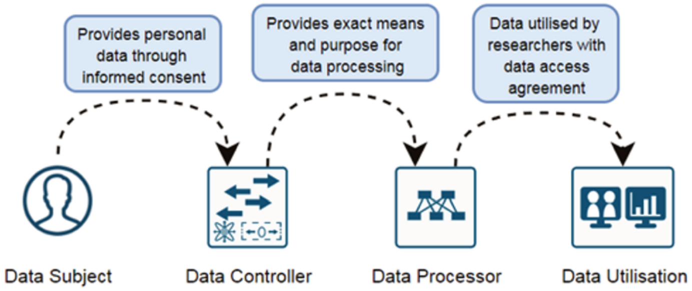

Provides personal data through informed consent
Provides exact means and purpose for data processing
Data utilised by researchers with data access agreement
Data Subject
Data Controller
Data Processor
Data Utilisation

Fig. 2.5.: (a) Diagram showing each of the steps between the data subject and legal use of personal data (Plug, R. 2021)

The GDPR applies to any identified or identifiable natural person. In order to process the information for research purposes in VODAN-Africa, a common technique that the data processor, in agreement with the data controller, may employ to provide privacy protections over accessed data is anonymisation. This involves replacing all directly and indirectly identifiable information in a data set with a unique identifier that does not disclose the identity of the data subject when records are retrieved, and thus cannot be linked to a data subject by combining separately stored data-sets. At the point of full anonymisation, such as aggregation used by VODAN-Africa, the GDPR no longer applies to the data, meaning that the data subject cannot be identified in any way and, thus, the data is not considered personal data.

Anonymisation Ensuring that personal data cannot be attributed to a data subject in any way, directly or indirectly, including by combining separately stored data sets (Preamble 26, GDPR).

Pseudonymisation is the process in which directly identifiable personal information is removed. The process of pseudonymisation is an important protection mechanism for sensitive data, such as medical data used with research exemption clauses, as the identity of the data subject is usually only of concern in extenuating circumstances or for verification of the integrity of the data.

Pseudonymisation Ensuring that personal data can only be attributed to a data subject indirectly, by utilising separately stored information for which access is strictly regulated (Article 4(5), GDPR).

By means of processing the different data available, the data can still lead to the data subject. As a result, the natural person is indirectly identifiable. It is not permissible that this data, which is not fully deidentified, leaves the localised instance in VODAN-Africa. Deidentification refers to the process through which data is not identifiable with a natural person to avoid the personal identity from being revealed by combining different data.

Deidentification Reducing the possibility of personal data linkage to a natural person by combining various data that in combination may reveal the identity of the data subject [52].

As the GDPR does not apply to completely anonymous data, a method that has been conceptualised in VODAN-Africa to improve the ease of data exchange for big data applications is to synthesise data based on the statistical properties of the original data belonging to the data subjects [53]. This process of data synthesis, in essence, extracts knowledge from the data through computational or mathematical processing, and then uses the knowledge to create new data that has not originated from a data subject.

Repositing synthetic data with proper provenance has certain benefits for VODAN-Africa, especially with regards to security and privacy, and increases the ease of data exchange. However, specific care has to be taken to ensure that combinations of the underlying distributions of these synthetic data does not contain the granularity that would allow indirect or approximate identification of the individuals from which these data were synthesised. This phenomenon is described as 'k-anonymity' [54]. Another point of concern is the quality of the data, as synthetic data is the result of sampling from a modelled distribution, rather than from a population that can be verified. The transparency and provenance in VODAN-Africa are important tools to uphold data quality when synthetic data is employed to model population health.

Synthetic data Data that has been generated from a measured distribution or computational process, and has not been obtained from direct measurement or observation.

Robust mechanisms for verification that may determine that synthetic data do indeed match the characteristics of the original data subjects through federated data, could ultimately result in synthetic data being verifiable through pseudonymous data, as their generative process could be linked to a population of data subjects. While these methods are developed in VODAN-Africa, GDPR has not yet elaborated on novel federated data concepts.

While the data steward as data processor within VODAN-Africa bears responsibility for the technical security aspects of a data repository, the hospital management as data controller has to perform due diligence through a Privacy Impact Assessment (PIA) documenting all identifiable information that will be obtained, the risks involved, and the conditions under which this data will be obtained. This documents the risk evaluation and impact assessment with respect to the risks to the rights of data subjects, which can be evaluated under GDPR throughout VODAN-Africa.

Finally, it is the responsibility of the data controller to notify the supervisory authority about data breaches, such as unauthorised access or access control failures. When managing health data, this would require immediate reporting to the regulatory health authority, such as the MoH of the relevant country under Article 33 of the GDPR, in accordance with Article 55 of the GDPR. While outside the EU this is not a legal requirement, VODAN-Africa supports transparency and liability over the security of personal data as an important safeguard. This underlines the well-documented and specified access and control patterns, in addition to record keeping of access within VODAN-Africa, are crucial when handling protected categories of personal data and form an essential basis of community wide trust in health data management.

Decentralised Control The establishment of the control is the responsibility of the establishment of the group of the undertakings of data production and processing in accordance with all the legal provisions that apply [Preambe 36, GDPR].

The decentralised control over the data processing is agreed with the different parties in Data Processing Agreements and Data Use Agreements. These assign specific responsibilities, following the divisions of tasks identified in GDPR as Data Processors, Data Controllers, Data

Protection Officers and Supervisory Authority [GDPR]. These responsibilities are identified in relation to each implementation environment where data production activity is taking place. The responsibilities are agreed with the different parties in Data Processing Agreements and Data Use Agreements. These agreements assign responsibilities concerning data production activity in accordance with the implementation environment in each place.

2.5 Discussion and conclusion

The purpose of this article is to develop a set of shared terminologies that allow for the unambiguous exchange of controlled vocabularies, and development of consistent data stewardship expertise throughout VODAN-Africa. At the core of the VODAN-Africa implementation lies the concept of knowledge management, which uses ontologies to manage data using graph representations that aid in findability and knowledge discovery in data where causality is highly relevant such as the health domain. The core elements of the architecture are transferable to other research areas and may be considered by other domains to establish data stewardship expertise and FAIR data networks. The core concepts, defined in this article, are each crucial to the deployment of a FAIR implementation network.

This article considers the elements involved in traditional health data management, identifies the challenges involved and discusses how these challenges are addressed within the FAIR architecture. Some of these challenges are technical in nature, while others deal with societal challenges such as compliance with regulations and the rights of individuals. These may vary in different locales and the help to bridge potentially fragmented realities concerning data management with different customs or rights awarded to protecting individuals and society.

Utilising both the GDPR, as well as the FAIR principles, and respecting the principle of personal privacy protection enshrined in the Universal Declaration of Human Rights, the VODAN IN shapes the way forward for sovereignty over health data, in the place where such data is produced and mindful of societal differences in relation to the management of the data.

By the definitions we have developed, we specify a framework of terms that build upon the VODAN-Africa architecture. This architecture is highly distributed and interoperable, ultimately managed and controlled in residence by data stewards that rely on unambiguous specifications. The data is available for aggregate analytics through data visiting computational techniques, always conditioned on permission for data visiting being permitted by the data controller of the health facility that produced the data.

This article was conceptualised as a review on how data terminology can be defined in the context of health data management, with a focus on aspects of FAIR and regulatory compliance. To this extent we have developed a comprehensive framework, that will support further development and deployment of FAIR data architectures in the domain of health, such as VODAN-Africa, and to modernise knowledge on health data management to educate a new generation of data stewards.

References

[1] M. Swan. „Philosophy of big data: Expanding the human-data relation with big data science services“. In: *Proceedings of the 2015 IEEE First International Conference on Big Data Computing Service and Applications* (2015), pp. 468–477 (cit. on pp. 22, 40).

[2] U. Sivarajah, M. Kamal, Z. Irani, and V. Weerakkody. „Critical analysis of big data challenges and analytical methods“. In: *Journal of Business Research* 70 (2017), pp. 263–286 (cit. on pp. 22, 40).

[3] A. L'Heureux, K. Grolinger, H.F. Elyamany, and M.A. Capretz. „Machine learning with big data: Challenges and approaches“. In: *IEEE Access* 5 (2017), pp. 7776–7797 (cit. on p. 22).

[4] United Nations. „Universal Declaration of Human Rights, Article 12“. In: *United Nations* (1948). Available at: https://www.un.org/en/about-us/universal-declaration-of-human-rights. Accessed 30 July 2021 (cit. on p. 23).

[5] M. Wilkinson, M. Dumontier, I.J. Aalbersberg, et al. „The FAIR Guiding Principles for scientific data management and stewardship“. In: *Scientific Data* 3.1 (2016), pp. 1–9 (cit. on p. 23).

[6] A. Liew. „Understanding data, information, knowledge and their interrelationships“. In: *Journal of Knowledge Management Practice* 7.2 (2007). ISSN: 1705-9232 (cit. on p. 24).

[7] S. Baskarada and A. Koronios. „Data, information, knowledge, wisdom (DIKW): A semiotic theoretical and empirical exploration of the hierarchy and its quality dimension“. In: *Australasian Journal of Information Systems* 18 (2013) (cit. on pp. 25, 27).

[8] P. Dalrymple. „Data, information, knowledge: The emerging field of health informatics“. In: *Bulletin of the American Society for Information Science and Technology* 37.5 (2011), pp. 41–44 (cit. on pp. 26, 27).

[9] A. Gold, A. Malhotra, and A.H. Segars. „Knowledge management: An organizational capabilities perspective“. In: *Journal of Management Information Systems* 18 (2001), pp. 185–214 (cit. on p. 27).

[10] N. Gibbins and N. Shadbolt. „Resource Description Framework (RDF)“. In: *Intelligence, Agents, Multimedia Group, University of Southampton* (2009) (cit. on pp. 27, 31).

[11] R. Chawuthai and H. Takeda. „RDF Graph visualization by interpreting linked data as knowledge“. In: *JIST* (2015) (cit. on p. 28).

[12] P. Heim, S. Hellmann, J. Lehmann, S. Lohmann, and T. Stegemann. „RelFinder: Revealing relationships in RDF knowledge bases“. In: *Semantic Multimedia. Lecture Notes in Computer Science* 5887 (2009) (cit. on p. 28).

[13] M.A. Sicilia. „Metadata, semantics, and ontology: Providing meaning to information resources“. In: *International Journal of Metadata, Semantics and Ontologies* 1.1 (2006), pp. 83–86 (cit. on pp. 29, 30, 33).

[14] R. Gartner. „Metadata: Shaping Knowledge from Antiquity to the Semantic Web“. In: *Springer International* (2016), p. 114 (cit. on pp. 29, 30).

[15] International Organization for Standardization and the International Electromechanical Commission (ISO/IEC). „Information technology: Metadata registries - Part 3: Registry metamodel and basic attributes“. In: *ISO/IEC 11179-3:2003(E)* (2003) (cit. on pp. 30, 33).

[16] G. Goos, J. Hartmanis, J. Van Leeuwen, et al. „The Semantic Web“. In: *Lecture Notes in Computer Science* (2011) (cit. on p. 31).

[17] M. Berners-Lee, J. Hendler, and O. Lassila. „The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities“. In: *Scientific American* 284.5 (2001), pp. 34–43 (cit. on p. 31).

[18] M. Dean, A. Schreiber, S. Bechofer, et al. „OWL Web Ontology Language – Reference“. In: *W3C Recommendation* (2004). Available at: http://www.w3.org/TR/owl-ref/. Accessed 30 July 2021 (cit. on p. 31).

[19] J. Greenberg. „Big metadata, smart metadata, and metadata capital: Toward greater synergy between data science and metadata“. In: *Journal of Data and Information Science* 2 (2017), pp. 19–36 (cit. on p. 32).

[20] J.J. Cimino, G. Hripcsak, S.B. Johnson, and P.D. Clayton. „Designing an Introspective, Multipurpose, Controlled Medical Vocabulary“. In: *Proceedings of the Annual Symposium on Computer Application in Medical Care* (1989). PMCID: PMC2245774, pp. 513–518 (cit. on p. 33).

[21] S. Jupe, B. Jassal, M. Williams, and G. Wu. „A controlled vocabulary for pathway entities and events“. In: *Database: The Journal of Biological Databases and Curation* (2014) (cit. on p. 34).

[22] M. Ashburner, C.A. Ball, J.A. Blake, et al. „Gene Ontology: Tool for the unification of biology“. In: *Nature Genetics* 25 (2000), pp. 25–29 (cit. on p. 34).

[23] G. Eysenbach. „What is e-health?“ In: *Journal of Medical Internet Research* 3.2 (2001), E20 (cit. on p. 35).

[24] WHO. „WHO guideline recommendations on digital interventions for health system strengthening“. In: *World Health Organization* (2019), p. 1 (cit. on p. 36).

[25] E.A. Amor and S.A. Ghannouchi. „Towards KPI-based health care process improvement“. In: *Procedia Computer Science* 121 (2017), pp. 767–774 (cit. on p. 36).

[26] M.V. Reisen, M. Stokmans, M. Basajja, et al. „Towards the tipping point for FAIR implementation“. In: *Data Intelligence* 2 (2020), pp. 264–275 (cit. on pp. 36, 43).

[27] M.V. Reisen, M. Stokmans, M. Mawere, et al. „FAIR practices in Africa“. In: *Data Intelligence* 2 (2020), pp. 246–256 (cit. on p. 37).

[28] P. Embi and P. Payne. „Clinical research informatics: Challenges, opportunities and definition for an emerging domain“. In: *Journal of the American Medical Informatics Association* 16.3 (2009), pp. 316–327 (cit. on p. 37).

[29] M. Reisen, F. Oladipo, M. Stokmans, et al. „Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research“. In: *Advanced Genetics* 2.2 (2021) (cit. on pp. 38, 40, 43, 46).

[30] F. Prasser, H. Spengler, R. Bild, J. Eicher, and K. Kuhn. „Privacy-enhancing ETL-processes for biomedical data“. In: *International Journal of Medical Informatics* 126 (2019), pp. 72–81 (cit. on p. 38).

[31] R. Gupta, M. Venkatachalapathy, and F.K. Jeberla. „Challenges in adopting continuous delivery and DevOps in a globally distributed product team: A case study of a healthcare organization“. In: *Proceedings of the 2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE)* (2019), pp. 30–34 (cit. on p. 38).

[32] S. Shastri, V. Banakar, M. Wasserman, A.C. Kumar, and V. Chidambaram. „Understanding and benchmarking the impact of GDPR on database systems“. In: *Proceedings of the VLDB Endowment* 13 (2020), pp. 1064–1077 (cit. on p. 39).

[33] M.E. Arass and N. Souissi. „Data lifecycle: From big data to SmartData“. In: *Proceedings of the 2018 IEEE 5th International Congress on Information Science and Technology (CiSt)* (2018), pp. 80–87 (cit. on p. 39).

[34] S. Shilo, H. Rossman, and E. Segal. „Axes of a revolution: Challenges and promises of big data in healthcare“. In: *Nature Medicine* 26 (2020), pp. 29–38 (cit. on p. 40).

[35] M. Weeber, J. Kors, and B. Mons. „Online tools to support literature-based discovery in the life sciences“. In: *Briefings in Bioinformatics* 6.3 (2005), pp. 277–286 (cit. on p. 42).

[36] M. Jarke, B. Otto, and S. Ram. „Data sovereignty and data space ecosystems“. In: *Business and Information Systems Engineering* 61 (2019), pp. 549–550 (cit. on p. 42).

[37] B. Mons. „The VODAN IN: Support of a FAIR-based infrastructure for COVID-19“. In: *European Journal of Human Genetics* 28 (2020), pp. 724–727 (cit. on p. 43).

[38] B. Mons. „FAIR science for social machines: Let’s share metadata knowlets in the Internet of FAIR Data and Services“. In: *Data Intelligence* 1 (2019), pp. 22–42 (cit. on p. 44).

[39] R.S. Gonçalves, M. O'Connor, M.M. Romero, et al. „The CEDAR Workbench: An ontology-assisted environment for authoring metadata that describe scientific experiments“. In: *The Semantic Web – ISWC 2017. Lecture Notes in Computer Science* 10588 (2017), pp. 103–110 (cit. on p. 44).

[40] M.W. Carroll. „Sharing research data and intellectual property law: A primer“. In: *PLoS Biology* 13 (2015) (cit. on p. 45).

[41] J. Wang, D. Crawl, S. Purawat, M. Nguyen, and I. Altintas. „Big data provenance: Challenges, state of the art and opportunities“. In: *Proceedings of the 2015 IEEE International Conference on Big Data (Big Data)* (2015), pp. 2509–2516 (cit. on p. 45).

[42] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo. „Protection of big data privacy“. In: *IEEE Access* 4 (2016), pp. 1821–1834 (cit. on p. 46).

[43] A. Zwitter and O. Gstrein. „Big data, privacy and COVID-19: Learning from humanitarian expertise in data protection“. In: *Journal of International Humanitarian Action* 5 (2020) (cit. on p. 46).

[44] E. Dove and M. Phillips. „Privacy law, data sharing policies, and medical data: A comparative perspective“. In: *Medical Data Privacy Handbook* (2020), pp. 639–678 (cit. on p. 46).

[45] J.M. Rumbold and B.K. Pierscionek. „The Effect of the General Data Protection Regulation on Medical Research“. In: *Journal of Medical Internet Research* 19 (2017) (cit. on p. 46).

[46] J. Winter and E. Davidson. „Big data governance of personal health information and challenges to contextual integrity“. In: *The Information Society* 35 (2019), pp. 36–51 (cit. on p. 46).

[47] M. Mostert, A. Bredenoord, M. Biesaart, and J. Delden. „Big data in medical research and EU data protection law: Challenges to the consent or anonymise approach“. In: *European Journal of Human Genetics* 24 (2016), pp. 1096–1096 (cit. on p. 48).

[48] EOSC Executive Board. „Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC)“. In: *European Open Science Cloud* 1.0 (2020) (cit. on p. 48).

[49] B. Mons, C. Neylon, J. Velterop, et al. „Cloudy, increasingly FAIR: Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud“. In: *Information Services and Use* 37 (2017), pp. 49–56 (cit. on p. 48).

[50] European Parliament and Council. „Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)“. In: *Official Journal of the European Union* L119 (2016), pp. 1–88 (cit. on p. 48).

[51] P. Voigt and A.V. Bussche. „The EU General Data Protection Regulation (GDPR): A practical guide“. In: *Springer* (2017) (cit. on p. 48).

[52] M. Hintze. „Viewing the GDPR through a de-identification lens: A tool for compliance, clarification and consistency“. In: *International Data Privacy Law* 8.1 (2018) (cit. on p. 54).

[53] A. Goncalves, P. Ray, B.C. Soper, et al. „Generation and evaluation of synthetic patient data“. In: *BMC Medical Research Methodology* 20 (2020) (cit. on p. 55).

[54] P. Samarati and L. Sweeney. „Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression“. In: *Technical Report SRI-CSL-98-04, Computer Science Laboratory, SRI International* (1998) (cit. on p. 55).