Special Issue: Launching an International FAIR Data Network for COVID Data

Pre-Publication 2021

Article 2: Terminology for a FAIR Framework for the Virus Outbreak Data Network

Authors: Plug R., Liang Y., Aktau, A., Basajja, M., Oladipo, F., Van Reisen, M.

Cite as:

Plug R., Liang Y., Aktau, A., Basajja, M., Oladipo, F., Van Reisen, M., (2021) Terminology for a FAIR Framework for the Virus Outbreak Data Network. Pre-publication. Prepared for: Special Issue: Launching an International FAIR Data Network for COVID Data. Data Intelligence.

Article 2: Terminology for a FAIR Framework for the Virus Outbreak Data Network

Authors: Plug R., Liang Y., Aktau, A., Basajja, M., Oladipo, F., Van Reisen, M.

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. The Virus Outbreak Data Network (VODAN) architecture builds on these 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. In this article we provide an overview of 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 architecture.

Keywords: data management, distributed data, federated data, data governance, FAIR, 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

1. Introduction

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

However, these developments towards big data pose numerous challenges, from the perspective of both society [2] and technology [3]. These challenges are magnified in the field of health data management, where privacy, security and 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.

FAIR provides a framework that addresses these concerns through a multimodal approach to data management and data stewardship [5]. Providing an architecture in which data is findable, accessible (under well-defined conditions), interoperable and reusable (FAIR) addresses technical concerns, while data stewardship empowers scientific communities with expertise to interact with these data management technologies.

FAIR provides a framework that is inherently distributed in order to provide data sovereignty, however, there are concerns over the convergence between localised instances of FAIR implementation. To reconcile such localised instances with a common vocabulary, we have developed a set of shared terminologies that allow for the unambiguous exchange and development of data stewardship expertise.

This article investigates the use of the FAIR Data Principles in the context of the Virus Outbreak Data Network (VODAN)-Africa, established as an Implementation Network (IN) under the GO-FAIR initiative jointly with FAIR IN Africa. The VODAN-Africa activity has been established as a first test to produce clinical patient data, which is by nature sensitive data, with the full retention of data ownership in residence, through data-visiting, and recognising the fragmented nature of the regulatory frameworks applicable in each place.

The investigation of such data in place is relatively easy, but this exercise is difficult when undertaken across countries and continents in an international community, such as VODAN-Africa. This article sets out the conceptual framework for the investigation of VODAN-Africa, exploring how a data network can be set up to facilitate the investigation of clinical patient data in full compliance with local and international regulatory frameworks, which benefits the use of digital data at point of care.

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]. Within this framework we start with data, which are the first elements we encounter in the operational sphere of technology.

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.

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 and is unstructured. To provide data with meaning, we need to process the data. 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].

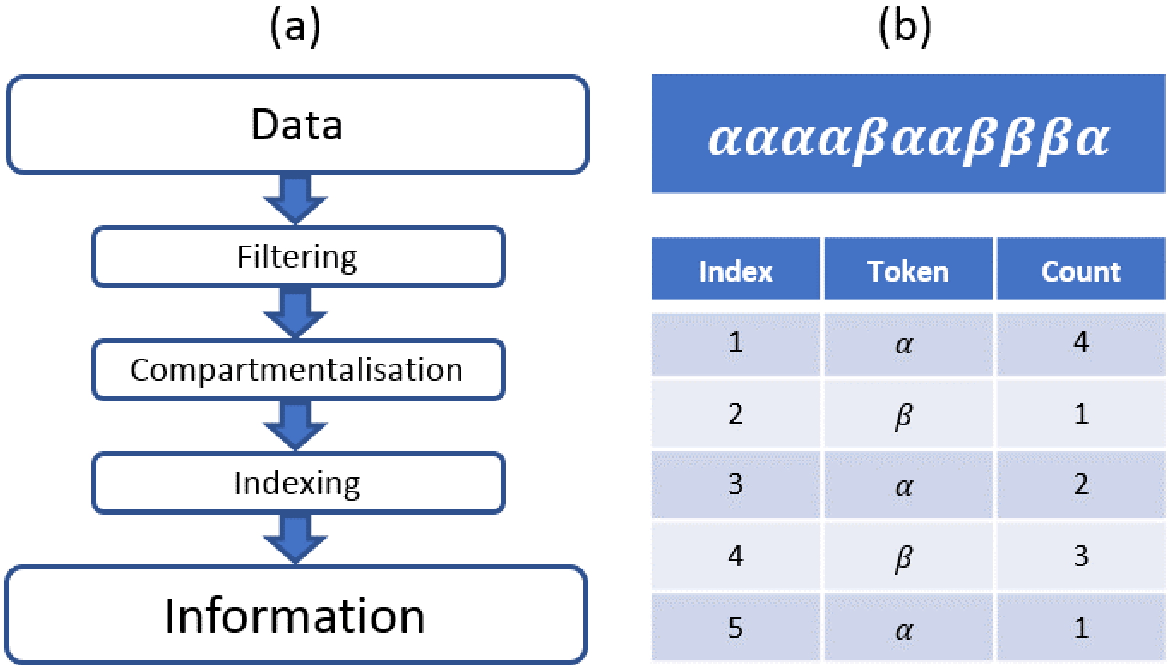
All the techniques that structure and give meaning to data are considered data processing techniques. The simplest example of this is entering data into a composite form, in which the structure of the form indicates the type of data that should be entered and assigns entered data to specific attributes.

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.

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. By gathering the data from the form, they can be entered into appropriate attribute fields in a digital format. In this respect, the human annotator has assigned meaning to the visual data 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. While both methods produce information, the information is unequal in specificity and granularity [8].



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

Figure 1. (a) Flowchart indicating the generalised process to transform data towards information **(b) Example of data (top) and possible resulting information (bottom)**

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.

By mapping the relationships between the information we have extracted from the data, we are transforming information into knowledge [7, 8]. 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. As relational descriptions in RDF are primarily used for machine interoperability, 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].

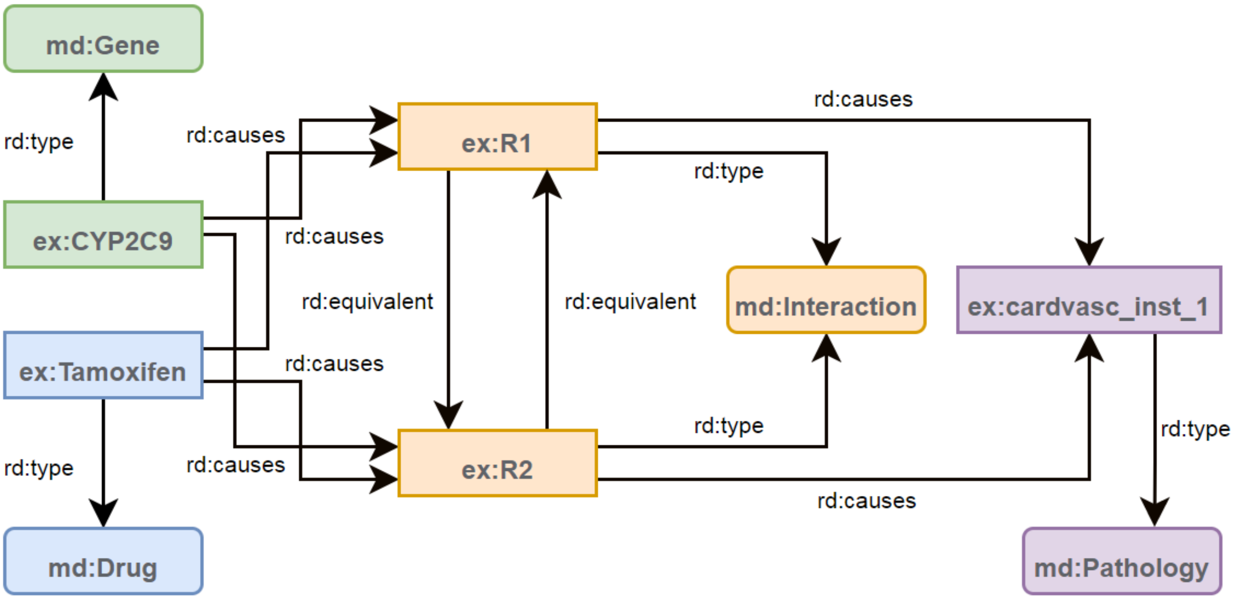
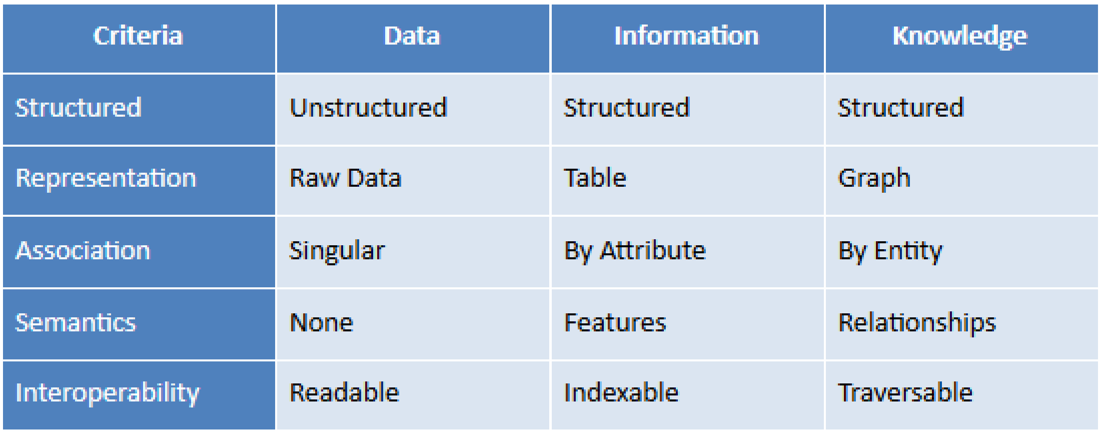


Figure 2. An example of an RDF graph for drug-gene interaction using knowledge discovery; equivalent interactions R1 and R2 have been associated with rd:equivalent (Plug, R. 2021)

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

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.

Table 1. Properties of data, information and knowledge



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

3. Metadata, semantics and ontologies

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.

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. 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 standardisation, the meaning of the metadata may be ambiguous or unspecified [14].

To standardise metadata, we define different types of metadata based on the objective that is associated with the denotation. 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.

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.

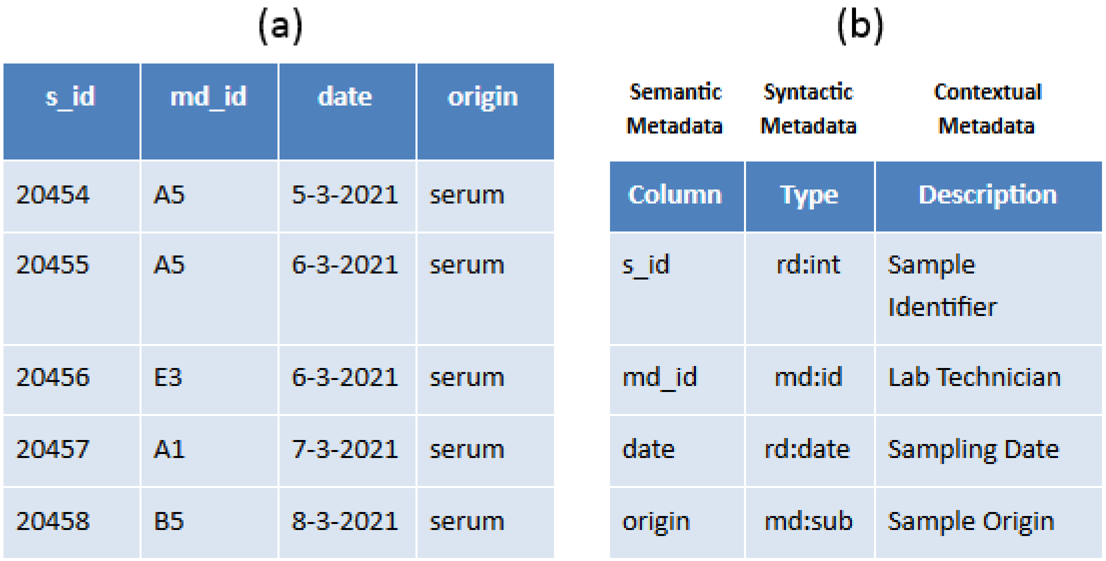
Syntactic metadata Metadata that provides structural specifications about data to aid machine interoperability.

Finally, we consider semantic metadata that specifies the meaning of data, which is the broadest concept for which metadata can be produced. 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 [15]. Semantic metadata is central to frameworks such as RDF to represent knowledge graphs [10] and the Web Ontology Language (OWL), which is used to formalise knowledge representations [16].

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.

Table 2. (a) Example data produced for the given metadata using a controlled vocabulary (b) Metadata as data, describing the properties of the various metadata



(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

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.

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’ [17], 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.

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, 14]. The baseline of standardisation is expressed through the use of agreed-upon 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, 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. To maintain the specificity and integrity of the knowledge base, it is important that such ambiguities are avoided and that concepts are expressed by their lemmatised version in order to achieve convergence within the conceptual 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.

In order to achieve this, a 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 [18]. 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 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. 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 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. 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. 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, which are produced from a domain ontology, we can standardise the way that products of data, information and knowledge are represented within an information system. The standardisation of terms and semantics defined by the metadata is a core element in producing data that is interoperable and reusable, and is key in the process of knowledge discovery.

4. Health data management

As health facilities have started collecting more data about patient's physiology, pharmacology and treatment efficacy, there has been an increasing need for the digitisation of health data, such as medical 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 [19].

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 straightforwardly represented as the development and application of digital technologies in the field of medicine [20]. Examples of health information in e-heath are patients' electronic health records (EHRs), genomic data, digital prescription, and even extending to remote surgery.

Care facilities frequently use health key performance indicators (KPIs) to compare their performance to that of other care facilities and to identify areas for improvement; in addition, KPIs can be correlated with measures directly related to treatment efficacy. For instance, average hospital stay is one of the commonly used healthcare KPIs measured for various treatment types [21], which represents the average time that patients stay in the hospital, and can be calculated as the total stay duration divided by total number of discharges. Changes in health as a result of metrics or specific health-care investments or interventions on participants in a clinical trial are known as factors related to common KPIs that affect health outcomes.

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. 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 VODAN-Africa addresses is the need for both in residence and aggregate analytics, using a specifically designed data management framework [22, 23].

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 is stored in local database architectures, which are defined as data repositories, driven by local ownership [24]. 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 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.

As a consequence, the repository is primarily responsible for maintaining information and knowledge through data, using the structure defined by the data architect. The core operations that can be performed on a repository are similar to that of a database, which are, fundamentally, data transactions and queries.

Repository query Any request for access to data, information or knowledge stored in a repository, based on a well-defined access pattern.

Repository transaction Any single unit of change that is applied to the data or knowledge base that is being held in the repository, based on a well-defined mechanism that retains data integrity.

These transactions, and the underlying repository performing these transactions, are part of a larger architecture, which we consider a health management information system (HMIS). This system forms the layer between the end-user (e.g., researchers and health professionals) and the data repository [25]. This allows for the management of access levels and interfacing directly with other applications within healthcare.

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 [26].

Data integration can be considered one of the main data management processes in operating an HMIS; it 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 [27]. In the healthcare industry, data integration plays an important role in helping doctors identify illnesses and medical diseases based on the integrated data from various healthcare systems. To perform an auto data integration process that connects and routes data from source systems to destination systems, the data integration technique ‘extract, transform, and load’ (ETL) can be applied.

Extract, transform, and load (ETL) A process that extracts data from disparate source systems, transforms it by applying various computations such as concatenations, and then loads the harmonised data into a data warehouse system.

When consolidating healthcare data into the data warehouse system, there are some challenges involved in the ETL process, which impose constraints on accessing data, the retention of data quality, and validation of data consistency. The FAIR framework provides a workable solution to these issues through the accessibility and interoperability principles. While working with the ETL process, it is important to maintain uptime during system upgrades. Continuous integration and delivery is a methodology used to implement this [28].

Continuous integration/ continuous delivery (CI/CD) A set of automation processes to provide continuous uptime while pushing modifications from the code base to the service machines.

Not all healthcare (meta)data are specific; there are some common data elements such as patient age, gender, and marital status that are common in a lot of clinical datasets from various healthcare systems. Common domain specific data elements also exist in health metadata and are defined in biomedical ontologies. These describe commonly used clinical data and can be used in clinical settings for patient care as well as for secondary data 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 [29].

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 [30]. 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 steps in the data creation stage are: acquiring existing data outside of the organisation, manually inputting new data within the organisation, and capturing data generated by devices used in accordance with a data collection protocol. 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.

5. Jurisdiction and data governance

The question of data ownership is an often-debated topic, and can be interpreted in a legal or philosophical sense. 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 [31]. 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. Each of these questions may depend heavily 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.

In addressing the question of data ownership, a baseline principle that must be upheld is data provenance [32]. 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 (data-lineage). Provenance is critical for reproducibility. Provenance, as an attribute of the data, is critical 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. For scientific purposes, 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. From the quality of data provenance, the question of data ownership can be addressed by means of identifying whom the subject of the data is (if available) and the party that initially collected or sampled the data.

Data provenance The origin and the processes that generated the data are known, well-documented and kept current.

Apart from concerns about data ownership, there are also legal and ethical concerns surrounding both the collecting and storing of data. Most of these legal concerns are focused on the privacy of subjects [33], 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 [34]. 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 [35, 36].

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 [37].

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.

Data localisation The practice of repositing data at the location where the data has been produced.

An implementation of this is to use an ethnographic design principle. 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 an optimised baseline for all stakeholders and streamlined, 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 [38].

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) [39], while upholding the same principles with regards to ethnographic design and full-scale interoperability [40].

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 [41, 42]. The GDPR, as a framework, revolves around transnational legislation for increasing operational transparency, promoting integrity, necessitating confidentiality and specifying the constraints of data processing.

General Data Protection Regulation (GDPR) The regulatory framework developed by the European Commission to provide uniform data security and privacy policies that enhance the rights of the data subject.

The GDPR 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. Where such processes take place outside the European Union (EU) and international law applies, the regulation also applies, even if controllers are not established inside the EU. 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 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).

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.

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, 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.

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. In addition, there must be no coercion and consent must be provided unambiguously by the data subject to enable another party to be able to process and store personal data about the data subject.

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, 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.

There are exemptions for certified public services that require more permissive data processing capabilities to function. 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 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 identified two parties that may handle personal data: the data controller and the data processor. 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 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 controllers specify the purpose and means under which data is handled, the data processor is the party responsible for processing and storing the data on behalf of the data controller. The processor does not specify what happens with the data, but only implements the requirements that have been set forward by the controller. It is the responsibility of the data processor to then implement an architecture with sufficient security measures and the ability to certify the integrity and security of personal data that is stored within their implementation. 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 be the same entity, for instance, a hospital with its own technical staff and data repository. However, data processing is typically covered by an external party, for example, a cloud service provider, that is contracted by the data controller. All the responsibilities, legal obligations and non-disclosure stipulations must be documented in a legal contract between the data controller and the 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).

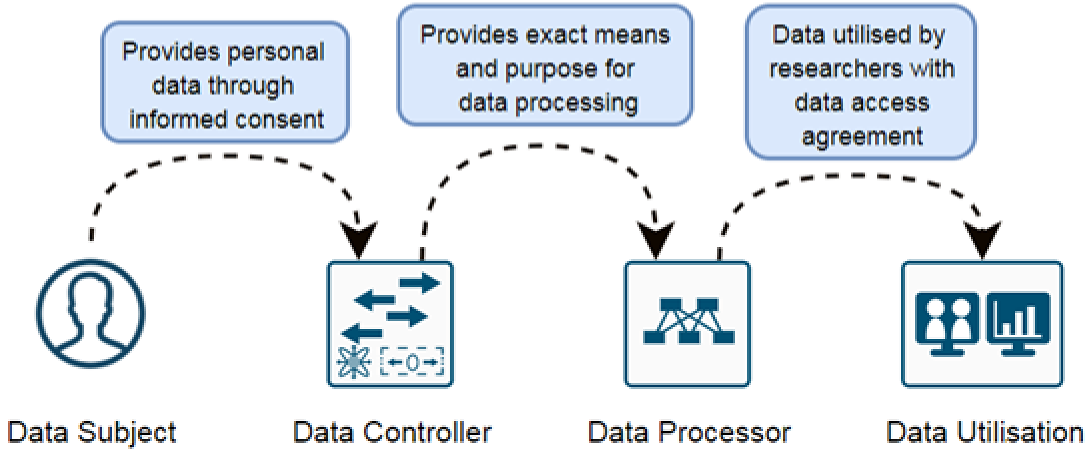

Provides personal data through informed consent
Provides exact means and purpose for data processing
Data utilised by researchers with data access agreement
Data Subject
Data Controller
Data Processor
Data Utilisation

Figure 3: 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 statistical or research purposes, a common technique that the data processor, in agreement with the data controller, may employ to increase 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 cannot be linked to a data subject by combining separately stored data-sets. At the point of full anonymisation, 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, but 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.

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).

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. A step further is to completely anonymise data, by means of eliminating all data from the records that may be used to identify a natural person as the data subject, at which point the data is no longer regulated by the GDPR.

As the GDPR does not apply to completely anonymous data, a method that has been conceptualised 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 [43]. 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.

Repositioning synthetic data has certain benefits, especially in regard 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’ [44]. 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.

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, could ultimately result in verifiable synthetic data being equivalent to pseudonymous data, as their generative process could be linked to a population of data subjects. These are novel developments which the GDPR has not yet elaborated on.

While the data processor bears responsibility for the technical security aspects of a data repository, the 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.

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 ministry of health of the concerned country under Article 33 of the GDPR, in accordance with Article 55 of the GDPR. This underlines the fact that well-documented and specified access and control patterns, in addition to record keeping of access, are crucial when handling protected categories of personal data.

6. FAIR data architectures

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 [45]. 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 FAIR Principles 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. These are the key principles that are used in the VODAN-Africa health management system.

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	Data should be discoverable by humans and machines through the use of metadata and data linkage.

Once data has been properly indexed and integrated into an 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.

Accessible Data should be accessible under well-defined conditions, and the conditions for access should be transparent.

A critical component that revolves around the findability and accessibility of the data is the machine interoperability of the data. For this, a baseline requirement is that the ontology (produced from the controlled vocabulary) must be resolvable 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 system is not operable, as the data pertaining to these metadata are not findable through automated methods.

Interoperable Data should be interlinked and operable for automated data processing, storage and analysis.

Through interoperability – by making the architecture well-specified, resolvable and machine-actionable – the conditions under which data becomes reusable are expressed in a formal framework. This resolves questions about the existence of research data and the conditions that are required to access these data. In addition, interoperability allows for techniques such as automated knowledge discovery [46] 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 remains in good provenance by maintaining all associated metadata specified in the DMP. In addition, the laws under which accessibility is regulated must be well-documented and both data and metadata provided with a licence describing the conditions under which access was provided.

Reusable Data should be in good provenance with documented metadata to allow for the replication or reuse of data.

The architecture of VODAN-Africa has been designed as a FAIR ecosystem, in which every aspect has been specified with the FAIR Principles 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 [47].

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, 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 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) [22] that are compliant with GDPR [23] 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. [26], 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 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 [48].

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) [49], which is responsible for the management of the ontologies, knowledge base and all activities related to data processing. This provides individual facilities with powerful 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 entire package of the FDP, implemented as a repository managed by CEDAR with services that provide a data visiting interface, forms the VODAN-Africa architecture.

7. Discussion and conclusion

The purpose of this article is to present a set of agreed terms in the context of VODAN. At the core of VODAN-Africa lies the concept of knowledge management, which uses ontologies to manage data using graph representations that aid in findability and knowledge discovery in data relevant to health. The core elements of the architecture are transferable to other content areas and may be considered by other communities established to set up FAIR data networks. The core concepts, defined in this article, are critical to the realisation of a FAIR IN.

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 FAIR Principles 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. All data is kept, managed and controlled in residence by the VODAN-Africa architecture, as well as the FAIR services. Newly developed technical concepts such as the visiting of data held in local repositories, under local control and governed by local regulatory frameworks, show the potential of FDPs to facilitate collaboration on big data analytics in fragmented policy geographies. This allows for an increase in cooperation in sensitive domains, such as health, while retaining non-disclosure and full data ownership.

References

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

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

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

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

[5] Wilkinson, M., Dumontier, M., Aalbersberg, I.J., et al.: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(1), 1–9 (2016).

[6] Liew, A.: Understanding data, information, knowledge and their inter-relationships. *Journal of Knowledge Management Practice* 7 (2007).

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

[8] Dalrymple, P.: Data, information, knowledge: The emerging field of health informatics. *Bulletin of the American Society for Information Science and Technology* 37(5), 41–44 (2011).

[9] Gold, A., Malhotra, A., Segars, A.H.: Knowledge management: An organizational capabilities perspective. *Journal of Management Information Systems* 18, 185–214 (2001).

[10] Gibbins, N., Shadbolt, N.: *Resource Description Framework (RDF).* Intelligence, Agents, Multimedia Group, University of Southampton, Southampton (2009).

[11] Chawuthai, R., Takeda, H.: RDF Graph visualization by interpreting linked data as knowledge. *JIST* (2015).

[12] Heim, P., Hellmann, S., Lehmann, J., Lohmann, S., Stegemann, T.: RelFinder: Revealing relationships in RDF knowledge bases. In: T.S. Chua, Y. Kompatsiaris, B. Mérialdo, W. Haas, G. Thallinger, W. Bailer (eds), *Semantic Multimedia*. SAMT Lecture Notes in Computer Science, vol. 5887, Springer, Berlin (2009).

[13] Sicilia, M.A.: Metadata, semantics, and ontology: Providing meaning to information resources. *International Journal of Metadata, Semantics and Ontologies* 1(1), 83–86 (2006).

[14] International Organization for Standardization and the International Electromechanical Commission (ISO/IEC): *Information technology: Metadata registries – Part 3: Registry metamodel and basic attributes, ISO/IEC 11179-3:2003(E)*. International Organization for Standardization, Geneva (2003).

[15] Berners-Lee, M., Hendler, J., Lassila, O.: The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* 284(5), 34–43 (2001).

[16] Dean, M., Schreiber, A., Bechofer, S., Harmelen, F.V., Hendler, J., Horrocks, I., MacGuinness, D., Patel-Schneider, P., Stein, L.: OWL Web Ontology Language – Reference. (2004). Available at: http://www.w3.org/TR/owl-ref/. Accessed 30 July 2021.

[17] Greenberg, J.: Big metadata, smart metadata, and metadata capital: Toward greater synergy between data science and metadata. *Journal of Data and Information Science* 2, 19–36 (2017).

[18] Jupe, S., Jassal, B., Williams, M., Wu, G.: A controlled vocabulary for pathway entities and events. Database: *The Journal of Biological Databases and Curation* (2014).

[19] Eysenbach, G.: What is e-health? *Journal of Medical Internet Research* 3(2), E20 (2001).

[20] WHO: *WHO guideline recommendations on digital interventions for health system strengthening*. World Health Organization, Geneva, p.1 (2019).

[21] Amor, E.A., Ghannouchi, S.A.: Towards KPI-based health care process improvement. *Procedia Computer Science* 121, 767–774 (2017).

[22] Reisen, M.V., Stokmans, M., Basajja, M., Ong'ayo, A., Kirkpatrick, C.R., Mons, B. Towards the tipping point for FAIR implementation. *Data Intelligence* 2, 264–275 (2020).

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

[24] Reisen, M.V., Stokmans, M., Mawere, M., Basajja, M., Ong'ayo, A., Nakazibwe, P., Kirkpatrick, C.R., Chindoza, K.: FAIR practices in Africa. *Data Intelligence* 2, 246–256 (2020).

[25] Embi, P., Payne, P.: Research paper: Clinical research informatics: Challenges, opportunities and definition for an emerging domain. *Journal of the American Medical Informatics Association* 16(3), 316–327 (2009).

[26] Van Reisen et al.: Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research. *Advanced Genetics* (2021).

[27] Prasser, F., Spengler, H., Bild, R., Eicher, J., Kuhn, K.: Privacy-enhancing ETL-processes for biomedical data. *International Journal of Medical Informatics* 126, 72–81 (2019).

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

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

[30] Arass, M.E., Souissi, N.: Data lifecycle: From big data to SmartData. In: *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, pp. 80–87 (2018).

[31] Carroll, M.W.: Sharing research data and intellectual property law: A primer. *PLoS Biology* 13 (2015).

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

[33] Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., Guo, S.: Protection of big data privacy. *IEEE Access* 4, 1821–1834 (2016).

[34] Zwitter, A., Gstrein, O.: Big data, privacy and COVID-19 – learning from humanitarian expertise in data protection. *Journal of International Humanitarian Action* 5 (2020).

[35] Dove, E., Phillips, M.: Privacy law, data sharing policies, and medical data: A comparative perspective. In: Aris Gkoulalas-Divanis, Grigorios Loukides (eds), *Medical Data Privacy Handbook*, Springer International Publishing, pp. 639–678 (2020).

[36] Vretta, M.: *The new EU General Data Protection Regulation (GDPR) in medical data and clinical research*. International Hellenic University, Thessaloniki (2019).

[37] Winter, J., Davidson, E.: Big data governance of personal health information and challenges to contextual integrity. *The Information Society* 35, 36–51 (2019).

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

[39] EOSC Executive Board: *Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC).* Version 1.0 15 February 2021, European Open Science Cloud (2020). Available at: https://eosc.eu/sites/default/files/EOSC-SRIA-V1.0_15Feb2021.pdf. Accessed 30 July 2021.

[40] Mons, B., Neylon, C., Velterop, J., Dumontier, M., Santos, L.O., Wilkinson, M.: Cloudy, increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud. *Information Services & Use* 37, 49–56 (2017).

[41] 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). *Official Journal of the European Union*, L119, pp. 1–88 (2016).

[42] Voigt, P., Bussche, A.V.: *The EU General Data Protection Regulation (GDPR): A practical guide*. Springer, Cham (2017).

[43] Goncalves, A., Ray, P., Soper, B.C., Stevens, J.L., Coyle, L., Sales, A.: Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology* 20 (2020).

[44] Samarati, P., Sweeney, L.: *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. Technical Report SRI-CSL-98-04, Computer Science Laboratory, SRI International, Menlo Park (1998).

[45] Shilo, S., Rossman, H., Segal, E.: Axes of a revolution: Challenges and promises of big data in healthcare. *Nature Medicine* 26, 29–38 (2020).

[46] Weeber, M., Kors, J., Mons, B.: Online tools to support literature-based discovery in the life sciences. *Briefings in Bioinformatics* 6(3), 277–86 (2005).

[47] Jarke, M., Otto, B., Ram, S.: Data sovereignty and data space ecosystems. *Business & Information Systems Engineering* 61, 549–550 (2019).

[48] Mons, B.: FAIR science for social machines: Let's share metadata knowlets in the Internet of FAIR Data and Services. *Data Intelligence* 1, 22–42 (2019).

[49] Gonçalves, R.S., O'Connor, M., Romero, M.M., Egyedi, A.L., Willrett, D., Graybeal, J., Musen, M.: The CEDAR Workbench: An ontology-assisted environment for authoring

metadata that describe scientific experiments. In: C. D'Amato et al. (eds) *The Semantic Web – ISWC 2017*. Lecture Notes in Computer Science, Springer, Cham, vol. 10588, pp. 103–110 (2017).