RESEARCH PAPER

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, 2331 GL Leiden, the Netherlands

[2]Kampala International University, 260101 Kampala, Uganda

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

Keywords: Data management; Distributed data; Federated data; Data governance; FAIR Guidelines; FAIR Data and Services; FAIR Data Point; FAIR framework

Citation: Plug, R., Liang, Y., Aktau, A., Basajja, M., Oladipo, F., Van Reisen, M.: Terminology for a FAIR framework for the Virus Outbreak Data Network-Africa. Data Intelligence 4(4), 698–723 (2022). doi: 10.1162/dint_a_00167
Submitted: March 10, 2021; Revised: June 10, 2022; Accepted: July 15, 2022

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 Guidelines have been developed to address these challenges. The Virus Outbreak Data Network (VODAN) architecture builds on these principles, using the European Union's 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.

ACRONYMS

CEDAR	Center for Expanded Data Annotation and Retrieval
DMP	data management plan
ETL	extract, transform, and load
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable
FDP	FAIR Data Point

† Corresponding author: Ruduan Plug, Leiden University (Email: r.b.f.plug@umail.leidenuniv.nl; ORCID: 0000-0001-5146-6116).

GDPR	General Data Protection Regulation
HMIS	health management information system
IN	Implementation Network
KPI	key performance indicator
OWL	Web Ontology Language
RDF	Resource Description Framework
URI	universal resource identifier
VODAN	Virus Outbreak Data Network

1. INTRODUCTION

Data management has become one of the prime factors of concern in all fields of contemporary 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 by expanding human expertise.

However, these developments towards automated data processing pose numerous challenges, from the perspective of society [2] and technology [3]. These challenges are magnified in the field of health, where privacy, security and the ownership of patient data 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, in Article 12, that "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence" [4]; therefore, personal data protection is enshrined in the foundations of international law.

The Virus Outbreak Data Network (VODAN) 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 Findable, Accessible (under well-defined conditions), Interoperable and Reusable (FAIR), we may address technical concerns about the use of 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 that way we deal with medical data within VODAN is inherently distributed, in order to ensure 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 VODAN, and specifically VODAN-Africa, which was established as an Implementation Network (IN) under the GO FAIR initiative, jointly with FAIR IN-Africa [6]. The VODAN-Africa initiative has been established as a pilot deployment to produce clinical patient data, which is by nature sensitive data (Article 1 of this Special Issue) [7]. 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 [6].

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 the further investigation of FAIR-based clinical patient data generation, processing and analytics within distributed and federated healthcare data applications.

2. DATA CONCEPTS

To develop our terminology framework, we built upon the core terminologies used in the process of data management. The first concepts we developed for our framework were 'data', 'information' and 'knowledge' [8], as they are procured within a clinical setting. In 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 into 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 the subsequent application of data, which is defined through biosemantics.

Data — Data is a set of numeric values, characters and/or symbols.

This 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 with standardised methods of formalisation. The three most common forms of data processing are: (i) select or sample the data relevant to the purpose by filtering, (ii) compartmentalise data into separate attributes, and (iii) provide an index to the data (i.e., a time-stamp, identifier, numeric ordering) [9].

All the techniques that structure and give meaning to data are considered data processing techniques. The simplest example of this employed at VODAN 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 — Information is 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, provided 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), who are extensively involved in VODAN. By gathering the data from the form, the data can be entered into appropriate attribute fields in a digital format.

In this way, the human annotator assigns meaning to the visual data, based on their existing knowledge, and transforms 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; for 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, unlike data stewards. While both methods produce information, the information is unequal in terms of specificity and granularity [10].

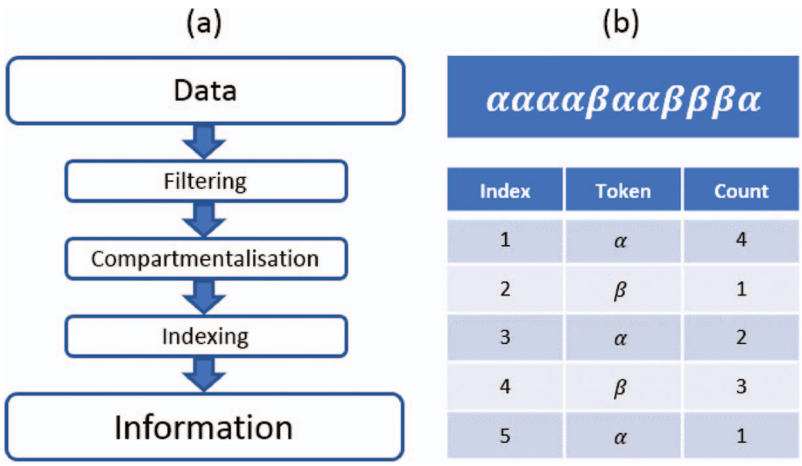

Figure 1. (a) Flowchart indicating the generalised process to transform data into information [11]. (b) Example of data (top) and possible resulting information (bottom) [11].

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 important in the context of the sensitive data processed in VODAN, as the context and meaning of these data are crucial to localised data methods. 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 [9, 10], which is one of the primary methods used in VODAN. 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 [12].

Knowledge Knowledge is 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) [13]. This is a data structure framework that implements a machine interoperable language to represent semantic graphs. In this context, each node is a universal resource identifier (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 a globally unique identifier when accessing and querying metadata across different services, which is important to enable unambiguous data access within VODAN.

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 [14], but an RDF-based knowledge representation provides a very powerful machine interpretable data structure that can be readily used for relational knowledge discovery [15], which is one of the core aims of the knowledge base developed within VODAN.

Knowledge discovery Knowledge discovery is 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.

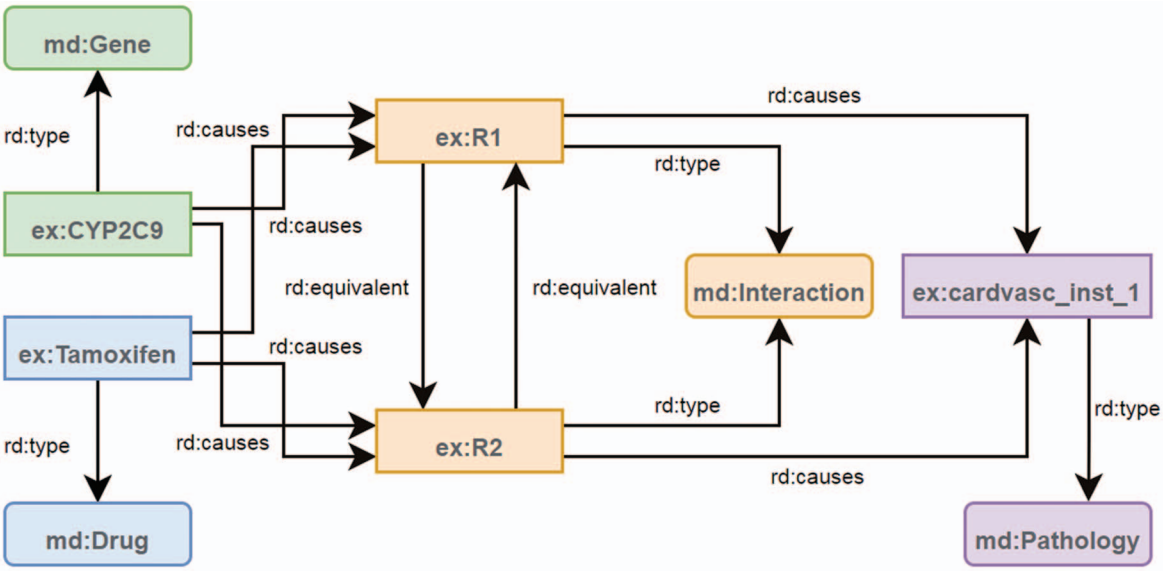

Figure 2. An example of an RDF graph for drug-gene interaction using knowledge discovery; equivalent interactions R1 and R2 have been associated with rd:equivalent [11].

Table 1. Properties of data, information and knowledge [11].

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

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 [16]. 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 derive that metadata are the building blocks that allow us to transform data into information and knowledge [17]. 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 [16].

Metadata Metadata is 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, in which 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 [18].

To standardise metadata, we define different types of metadata based on the objective that is associated with the denotation [17]. 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.

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

Finally, we consider semantic metadata, which specifies the meaning of data and is the broadest concept for which metadata can be produced [19]. 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 et al. [20]. Semantic metadata is central to frameworks such as RDF to represent knowledge graphs [13] and the Web Ontology Language (OWL) [21], which is used to formalise knowledge representations [21] that are used by clinicians and implemented by data stewards in VODAN.

Semantic metadata Semantic metadata is 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 of data produced for the given metadata using a controlled vocabulary [11]. (b) Metadata as data, describing the properties of the various metadata [11].

(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, which forms an essential part of the VODAN 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' [22], 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.

Metadata specification Metadata specification is 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 [16, 18]. The baseline of VODAN 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 Vocabulary is 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, 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 [23]. To maintain the specificity and integrity of the knowledge base, it is important that such ambiguities are avoided by using lemmatised concepts across VODAN 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 the purposes of convergence within health communities.

In order to achieve this within VODAN, 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 [24]. 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 Controlled vocabulary is a curated set of terms and symbols from which concepts and relationships 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 [25]. 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 Ontology is 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-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 [25]. This limits the metadata that may be associated with data, and can be expressed by constraining the metadata to a template format that controls for terms and semantic properties.

Metadata template — A metadata template is a set of semantically valid, domain-specific metadata specifications derived from constraints specified by an ontology.

By using metadata templates in VODAN, 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.

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, across which a multitude of health facilities have thus far operated using manual data entry or handwritten patient records. Eysenbach describes these digitisation efforts as eHealth, representing the relationship between medicine and computers and how this combination can benefit the healthcare and pharmacological industries [26].

However, because of the rapid development of data collection and healthcare information technologies, the academic definition of eHealth 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 [27] in an effort to improve interoperability. Examples of health information in eHeath are patients' electronic health records (EHRs), genomic data, digital prescription, and even extending to remote diagnostics, each of which are data encompassed in VODAN.

Care facilities frequently use health key performance indicators (KPIs), based on which VODAN defines 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. KPIs can be specially used to identify areas for improvement. In addition, KPIs can be correlated with measures directly related to treatment efficacy within the local context. For instance, average hospital stay and outpatient rate are some of the commonly used healthcare KPIs within VODAN, measured for various treatment types [28].

Healthcare key performance indicators (KPIs) — Healthcare key performance indicators are 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, which addresses health at both the clinical as well as the 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 VODAN-Africa addresses is the need for both in residence and aggregate analytics, using a specifically designed data management framework [29, 30].

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

The data that is present in residence within VODAN is stored in local database architectures, which are defined as data repositories, driven by local ownership [31]. 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	A data repository is 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 HMISs in Africa are proprietary [31], which is a large drawback that VODAN seeks to address. Typical, the HMIS forms the layer between the end-user (e.g., researchers and health professionals) and the data repository [32]. This allows for the management of access levels and for interfacing directly with other applications that are used within departments of a healthcare facility.

Health management information system (HMIS)	A health management information system is a system for entering, storing, maintaining, retrieving, and processing health data stored in repositories. It 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 [6].

Data integration can be considered one of the main data management processes in operating an HMIS, and 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 [33]. When consolidating healthcare data into a HMIS, 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 [34]. The FAIR framework provides a workable solution to these issues through the accessibility and interoperability specifications, which in case of VODAN are transparent and locale-dependent [6].

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

Common data elements (CDEs) Common data elements are 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 [35].

Data management plan (DMP) A data management plan is 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 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 [36]. 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 The data lifecycle is 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, the steps afterwards are standardised [6]. After the data creation stage, the data must be stored and protected with different security levels within the organisation, based on the specifications and regulations. 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 [37]. 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 Guidelines were conceptualised by Wilkinson et al. [5], 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 [5]. These are the key principles that are used throughout the VODAN-Africa implementation of the VODAN health data management architecture.

The primary requirement of FAIR compliance with respect to data management, is the baseline specification for data to be discoverable 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 is findable when it is 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 relevant ministry of health.

Accessible health data Health is accessible when health data, information or knowledge in residence is accessible, possibly in an 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. 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 are interoperable when they are 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 become reusable are expressed in a formal framework. Interoperability throughout VODAN allows for techniques such as automated knowledge discovery [38] 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 remains in good provenance, which is done by maintaining all associated metadata specified in the DMP. In addition, the laws of each VODAN locale under which accessibility is regulated must be well-documented, and both data and metadata have to be provided with a specification describing the conditions under which access may be provided.

Reusable health data	Health data is reusable when it is in good provenance, with documented metadata to allow for the replication or reuse of data across health facilities and locales.

The architecture of VODAN has been designed as a FAIR ecosystem, in which every aspect has been specified, with the FAIR Guidelines 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 [39].

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

By keeping data in residence in VODAN, and maintaining the rights of the data owner, data controllers and processors work under the local laws and regulations in the jurisdiction. This ensures that the rights of the data subject are always maintained in accordance with the government processes, which are 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) [29] that are compliant with the General Data Protection Regulation (GDPR) [30] 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 FAIR Data Point is a local data repository (with accompanying services) that is compliant with the FAIR Guidelines.

The design of this network is specified in the design of a FAIR digital health infrastructure by van Reisen et al. [6], in which 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, in which meta-analyses can be performed on VODAN aggregate data, while retaining full data sovereignty, and is, thus, also compliant with regulatory frameworks in regard to privacy and data protection.

Data visiting	Data visiting refers to the 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, in which each resource has a unique identifier that is persistent as well as resolvable [40].

Unique, persistent and resolvable identifier (UPRI)	This refers to a unique, persistent and resolvable identifier for digital objects.

A FAIR compliant system to support the data processing and management of VODAN-Africa FDPs is implemented at the Center for Expanded Data Annotation and Retrieval (CEDAR) [41], 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 architecture.

4. JURISDICTION AND DATA GOVERNANCE

The question of data ownership is both a legal and philosophical challenge and plays a central role in VODAN. 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 [42]. 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, in 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 owner	The data owner is the individual or party who 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 [43]. 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, 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.

For scientific purposes, the quality of data provenance is critical to an investigation of the environmental interactions of data in the context in which it was generated, not only in terms of the locale, but also the data subject cluster. From the quality of data provenance, the question of data ownership can be addressed by means of identifying who the subject of the data is, if applicable, and the party that initially collected or sampled the data.

Data provenance	Data is in good provenance when the origin of the data and the processes that generated the data are known, well-documented and kept current.

Apart from concerns about data ownership in VODAN, 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 [44], 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 [45]. 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 [46, 47], which is one of the main concerns of VODAN stakeholders [6].

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 [48].

Data governance	Data governance is the enactment of regulations and policies surrounding the collection, handling and storage of data, as well as the authorisation and 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, in which 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 Data localisation is the practice of repositing data at the location where the data was produced.

An implementation of this is to use ethnographic design principles across VODAN. Within the community, which seeks the convergence of information systems, 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 An ethnographic design is a participatory collaborative design 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 transparency when it comes to the intermediate processes by which data are stored and archived [49].

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) [50], while upholding the same principles with regards to ethnographic design and full-scale interoperability [51].

In relation to legal concepts regarding data, information and knowledge management, VODAN uses the GDPR as the foundational legislative frame of reference [52, 53]. The GDPR 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 Personal data is 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 over and legal obligations in relation to a specified set of data, under the GDPR we fully recognise the rights of the individual from whom data has been collected. As VODAN provides full data provenance, this becomes feasible to implement over the entire implementation network. As a consequence, we assume that the individual from whom data has been drawn retains 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's 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 must have 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 anonymization of the data to protect the data subject. The aggregation process, such as that used in VODAN, depersonalises data and, as such, they no longer pertain to a specific data subject and are, thus, not considered personal data.

Data subject — A data subject is 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, 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 their 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 permission to store and use their personal data fall 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, who drive the semantic and purpose of data within VODAN.

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 — Informed consent is consent that is voluntary, specific and unambiguously 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, 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 covers a wide variety of legislative frameworks, these limitations may vary, but should 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 ministries of health associated with VODAN. 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 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 — Special categories of personal data refers to 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 that may handle personal data: the clinician as data controller and the data steward as 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 was produced.

Data controller — The data controller is 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 by 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 at 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 cases, be the same entity, for instance, in 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 data processor is 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).

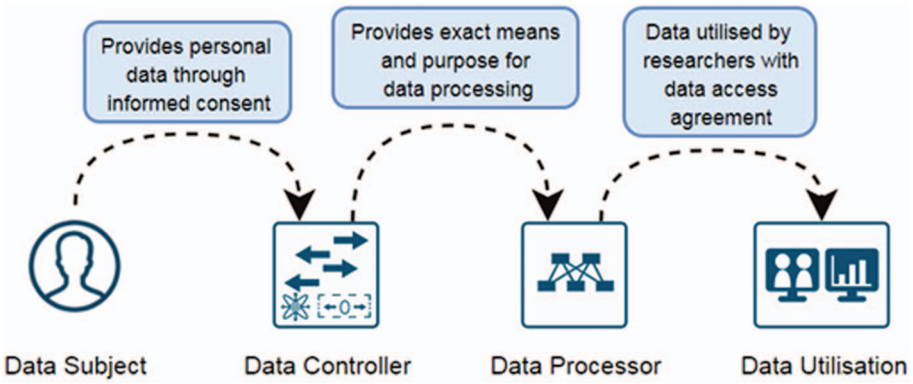

Figure 3. Diagram showing each of the steps between the data subject and legal use of personal data [11].

The GDPR applies to any identified or identifiable natural person. In order to process the information for research purposes in VODAN, a common technique that the data processor, in agreement with the data controller, may employ to provide privacy protection over accessed data is anonymization. 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 anonymization, such as aggregation used by VODAN, 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.

Anonymization Anonymization is the process of 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. In VODAN, these data are not allowed to leave the localised instance.

Pseudonymisation Pseudonymisation is the process of ensuring that personal data can only be attributed to a data subject indirectly, by utilising separately stored information, to 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.

As the GDPR does not apply to completely anonymous data, a method that has been conceptualised in VODAN 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 [54]. 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, especially in relation 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' [55]. 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. In VODAN, transparency and provenance are important tools for upholding data quality when synthetic data is employed to model population health.

Synthetic data — Synthetic data is 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, the GDPR has not yet elaborated on novel federated data concepts.

While the data steward, as data processor within VODAN, bears responsibility for the technical security aspects of a data repository, the clinicians, as data controllers, have 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 the 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 the GDPR throughout VODAN.

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 relevant country under Article 33 of the GDPR, in accordance with Article 55 of the GDPR. While outside the European Union (EU) this is not a legal requirement, VODAN supports transparency and liability for the security of personal data as an important safeguard. This underlines the well-documented and specified access and control patterns, as well as record keeping of access, within VODAN, which are crucial when handling protected categories of personal data and form an essential basis for community trust in health data management.

5. DISCUSSION AND CONCLUSION

The purpose of this article was to develop a set of shared terminologies that allow for the unambiguous exchange of controlled vocabularies and the development of consistent data stewardship expertise throughout VODAN. At the core of the implementation 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 when 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 in 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 Guidelines 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 Guidelines, and respecting the principle of personal privacy protection enshrined in the Universal Declaration of Human Rights, VODAN-IN shapes the way forward for sovereignty over health data, in the place where the data is produced and mindful of societal differences in relation to the management of the data. Through the definitions we have developed, we specify a framework of terms that build upon the VODAN architecture. This architecture is highly distributed and interoperable, ultimately managed and controlled in residence by data stewards that rely on unambiguous specifications.

This article was conceptualised as a review of 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 the further development and deployment of FAIR data architectures in the domain of health, such as VODAN-Africa, and the modernisation knowledge on health data management to educate a new generation of data stewards.

ACKNOWLEDGEMENTS

I would like to acknowledge the help of Prof. Dr Mark Musen and John Graybeal of the Stanford Biomedical Informatics Research institute for research support; Misha Stocker for management and coordination of the Special Issue (Volume 4); and Susan Sellars for copyediting and proofreading. We would also like to acknowledge VODAN-Africa, the Philips Foundation, the Dutch Development Bank FMO, CORDAID, and the GO FAIR Foundation for supporting this research.

AUTHORS' CONTRIBUTIONS

Ruduan Plug (ruudplug@gmail.com, 0000-0001-5146-6116): Writing—original draft preparation, investigation, conceptualization. Yan Liang (liangyan12138@gmail.com, 0000-0002-3860-725X): Writing—original draft preparation, investigation. Aliya Aktau (aleka.aktau@gmail.com, 0000-0003-4942-2725): Writing—review and editing, investigation. Mariam Basajja (mariam.basajja@gmail.com, 0000-0001-7710-8843): Writing—review and editing, investigation. Francisca Oladipo (francisca.oladipo@kiu.ac.ug, 0000-0003-0584-9145): Writing—review and editing, validation, project administration. Mirjam van Reisen (mirjamvanreisen@gmail.com, 0000-0003-0627-8014): Supervision, validation, conceptualization, project administration.

CONFLICT OF INTEREST

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

ETHICS STATEMENT

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

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). doi:10.1109/BigDataService.2015.29

[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). doi: 10.1016/j.jbusres.2016.08.001

[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). doi: 10.1109/ACCESS.2017.2696365.

[4] United Nations: Universal Declaration of Human Rights. United Nations General Assembly (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., Appleton, G., Axton, M., Baak, A., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1), 1–9 (2016). https://doi.org/10.1038/sdata.2016.18

[6] Van Reisen, M., Oladipo, F., Stokmans, M., Mpezamihgo, M., Folorunso, S., Schultes, E., et al.: Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research. Advanced Genetics 2(2) (2021). doi: 10.1002/ggn2.10050

[7] Van Reisen, M., Oladipo, F., Mpezamihigo, M., Plug, R., Basajja, M., Aktau, A., Purnama Jati, P.H., Nalugala, R., Folorunso, S., Amare, Y.S., Abdullahi, I., Afolabi, O.O., Mwesigwa, E., Taye, G.T., Kawu, A., Ghardallou, M., Liang, Y., Osigwe, O., Medhanyie, A.A., Maware, M.: Incomplete COVID-19 data: The curation of medical health data by the Virus Outbreak Data Network-Africa. Data Intelligence 4(4), 673–697 (2022)

[8] Liew, A.: Understanding data, information, knowledge and their inter-relationships. Journal of Knowledge Management Practice 7(2) (2007). ISSN: 1705-9232
[9] 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). doi: 10.3127/ajis.v18i1.748
[10] 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). doi: 10.1002/bult.2011.1720370512
[11] Plug, R.: Presentation. Unpublished (2021)
[12] Gold, A., Malhotra, A., Segars, A.H.: Knowledge management: An organizational capabilities perspective. Journal of Management Information Systems 18, 185–214 (2001). doi: 10.1080/07421222.2001.11045669
[13] Gibbins, N., Shadbolt, N.: Resource Description Framework (RDF). Intelligence, Agents, Multimedia Group, University of Southampton, Southampton (2009). doi: 10.1081/E-ELIS4-120043688
[14] Chawuthai, R., Takeda, H.: RDF Graph visualization by interpreting linked data as knowledge. In: G. Qi, K. Kozaki, J. Pan, S. Yu (eds) Semantic Technology, Joint International Semantic Technology Conference (JIST), Lecture Notes in Computer Science, Vol. 9544, Springer, Cham (2015). doi: 10.1007/978-3-319-31676-5_2
[15] 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). doi: 10.1007/978-3-642-10543-2_21
[16] Sicilia, M.A.: Metadata, semantics, and ontology: Providing meaning to information resources. International Journal of Metadata, Semantics and Ontologies 1(1), 83–86 (2006). doi: 10.1504/IJMSO.2006.008773
[17] Gartner, R.: Metadata: Shaping knowledge from antiquity to the Semantic Web. Springer International, Cham, p. 114 (2016). doi:10.1080/17583489.2017.1301713
[18] 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). ISO, Geneva (2003).
[19] Goos, G., Hartmanis, J., Leeuwen, J.V., Hutchison, D., Pan, J.Z., Chen, H., Kim, H.G., Li, J., Wu, Z., Horrocks, I., Mizoguchi, R., Wu, Z.: The Semantic Web. Lecture Notes in Computer Science (2011). doi: 10.1007/978-3-642-29923-0
[20] 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)
[21] 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 [Online]. (2004). Available at: http://www.w3.org/TR/owl-ref/. Accessed 30 July 2021
[22] 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). doi: 10.1515/jdis-2017-0012
[23] Cimino, J.J., Hripcsak, G., Johnson, S.B., Clayton, P.D.: Designing an introspective, multipurpose, controlled medical vocabulary. Proceedings of the Annual Symposium on Computer Application in Medical Care, pp. 513–518 (1989)
[24] 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). doi: 10.1093/database/bau060
[25] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H.L., et al.: Gene ontology: Tool for the unification of biology. Nature Genetics 25, 25–29 (2000). doi: 10.1038/75556

[26] Eysenbach, G.: What is e-health? Journal of Medical Internet Research 3(2), E20 (2001). doi: 10.2196/JMIR.3.2.E20
[27] WHO: WHO guideline recommendations on digital interventions for health system strengthening. World Health Organization, Geneva, p. 1 (2019)
[28] Amor, E.A., Ghannouchi, S.A.: Towards KPI-based health care process improvement. Procedia Computer Science 121, 767–774 (2017). doi: 10.1016/J.PROCS.2017.11.099
[29] Van Reisen, M., Stokmans, M., Basajja, M., Ong'ayo, A., Kirkpatrick, C., Mons, B.: Towards the tipping point for FAIR implementation. Data Intelligence 2, 264–275 (2020). doi: 10.1162/dint_a_00049
[30] Mons, B.: The VODAN IN: Support of a FAIR-based infrastructure for COVID-19. European Journal of Human Genetics 28, 724–727 (2020). doi: 10.1038/s41431-020-0635-7
[31] Van Reisen, M., Stokmans, M., Mawere, M., Basajja, M., Ong'ayo, A.O., Nakazibwe, P., Kirkpatrick, C.R., Chindoza, K.: FAIR practices in Africa. Data Intelligence 2, 246–256 (2020). doi: 10.1162/dint_a_00047
[32] 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). doi: 10.1197/jamia.M3005
[33] 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). doi: 10.1016/J.IJMEDINF.2019.03.006
[34] 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). doi: 10.1109/ICGSE.2019.00020
[35] 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). doi: 10.14778/3384345.3384354
[36] 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). doi: 10.1109/CIST.2018.8596547
[37] Shilo, S., Rossman, H., Segal, E.: Axes of a revolution: Challenges and promises of big data in healthcare. Nature Medicine 26, 29–38 (2020). doi: 10.1038/s41591-019-0727-5
[38] 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). doi: 10.1093/BIB/6.3.277
[39] Jarke, M., Otto, B., Ram, S.: Data sovereignty and data space ecosystems. Business & Information Systems Engineering 61, 549–550 (2019). doi: 10.1007/S12599-019-00614-2
[40] 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). doi: 10.1162/dint_a_00002
[41] 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). doi: 10.1007/978-3-319-68204-4_10
[42] Carroll, M.W.: Sharing research data and intellectual property law: A primer. PLoS Biology 13 (2015). doi: 10.1371/journal.pbio.1002235
[43] 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). doi: 10.1109/BigData.2015.7364047
[44] Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., Guo, S.: Protection of big data privacy. IEEE Access 4, 1821–1834 (2016). doi: 10.1109/ACCESS.2016.2558446

[45] Zwitter, A., Gstrein, O.: Big data, privacy and COVID-19—learning from humanitarian expertise in data protection. Journal of International Humanitarian Action 5 (2020). doi: 10.1186/s41018-020-00072-6

[46] 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). doi: 10.1007/978-3-319-23633-9_24

[47] Rumbold, J.M., Pierscionek, B.K.: The effect of the General Data Protection Regulation on medical research. Journal of Medical Internet Research 19 (2017). doi: 10.2196/jmir.7108

[48] Winter, J., Davidson, E.: Big data governance of personal health information and challenges to contextual integrity. The Information Society 35, 36–51 (2019). doi: 10.1080/01972243.2018.1542648

[49] 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). doi: 10.1038/ejhg.2015.239

[50] 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 (EOSC) (2020). Available at: https://eosc.eu/sites/default/files/EOSC-SRIA-V1.0_15Feb2021.pdf. Accessed 30 July 2021

[51] 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). doi: 10.3233/ISU-170824

[52] European Parliament and Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, 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). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&rid=2. Accessed 7 September 2021

[53] Voigt, P., Bussche, A.V.: The EU General Data Protection Regulation (GDPR): A practical guide. Springer, Cham (2017). doi: 10.1007/978-3-319-57959-7

[54] 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). doi: 10.1186/s12874-020-00977-1

[55] 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). doi: 10.1.1.37.5829