Universiteit Leiden

MSc. ICT in Business and the Public Sector

The Design of Ontology Based Access Control for Virus Outbreak Data Network

Name: Mingyu Huang

Student-no: s2820730

Date: 12/10/2022

1st supervisor: Prof. Dr. Mirjam van Reisen

2nd supervisor: Dr. Katy Wolstencroft

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Acknowledgement

First of all, I would like to express my deepest gratitude to my first supervisor, Prof. Dr. Mirjam van Reisen for her kind guidance of both this thesis and lectures such as Regulatory Governance for Data Science and Data Science in Practice. I also want to thank my second supervisor, Dr. Katy Wolstencroft for her constructive revision suggestions and valuable insights through her presentation, which inspired me to research the topic of semantic web.

Secondly, I would like to thank all friends from the VODAN family for their insightful inputs in interviews which play an essential role in this thesis, especially Ruduan Plug, Putu Hadi Purnama Jati, and Samson Yohannes. I also want to express my appreciation for Seth Okeyo, Erik Flikkenschild, and interview participants from the Dutch government for their valuable opinions and contribution to this research.

Lastly, I would like to express my deepest appreciation to my beloved family for their unconditional support and guidance along this journey. And many thanks to my friends and fellow students in ICT in Business for the companion and encouragement.

Abstract

In the past 20 years, the exponentially increasing amount of data especially in the healthcare industry have posed challenges associated with data governance, and data ownership. In this context, the FAIR data principle was proposed by (Wilkinson, et al., 2016) including Findability, Accessibility, Interoperability and Reusability principle with intent to enhance the data stewardship. In addition, FAIR principle increases interoperability of massive scientific data by deploying FAIR Data Point (FDP) as a way to create, collect, and store distributed data. The data generated under the guidance of FAIR principle is described by rich metadata and identified by Unique Resource Identifiers, the interoperability of data is thus enhanced, and it is accessible and readable once published. FAIR data has been broadly adopted globally in several domains because using FAIR principle as a solution to manipulate open data does not break data privacy but increase the trust among stakeholders of FAIR-related project. For instance, Virus Outbreak Data Network-Africa (VODAN-Africa) project has built up trust with 88 collaborative countries in Africa because of its strict alignment with FAIR principle. Concrete mechanisms or solutions are required to implement in support of the current practice, whereas FAIR guideline and GDPR have already promoted high interoperability of data and privacy protection respectively from regulatory perspective. The goal of this research is to investigate data privacy problems existing in current data flow architecture, and to design an ontology-base access control (OBAC) mechanism for managing healthcare data with triple-patterns. The research and implementation of OBAC bases on the following considerations. First, modern data curation strategies develop towards the decentralized fashion, the use of ontologies can model this distributed pattern in line with the interoperability requirement of FAIR principle. By linking separated data production nodes together, we can form the domain knowledge, which enables the secondary research and reasons new knowledge. Second, VODAN-Africa project and other open data initiatives in the healthcare sector creates massive ontologies described by rich metadata. Simple protection mechanisms such as allowing users to define the accessibility (public/private) of these linked ontologies limit the availability of open data even though they enhance the security level of the system. Therefore, it is necessary to implement access control based on the underlying metadata to constrain the access to these ontologies. Last but not least, OBAC aligns with the accessibility requirement of FAIR principle in which data communication protocol should enable authentication and authorization mechanism.

Keywords: FAIR principle, Ontology-Based Access Control, GDPR, Data governance

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Problem Statement

To address the data ownership challenge, VODAN-A has launched a Minimal Viable Product (MVP) with internal dashboard where medical data (patient registries and local facilities' information) is collected and visualized, as well as external dashboard where statistics are aggregated by data queried from local facilities. This is a milestone of VODAN-A project, which symbolizes its production and reuse of high-quality medical data start in African geographies. As to the data ownership preservation, VODAN-A internal dashboard has been equipped with the authentication and authorization procedure to manage the visibility of statistics within facilities based on users' role. This prevents internal resources from being visited by irrelevant or unauthorized users to some extent. In addition, aggregated statistics are only visible to external visitors. In this circumstance, the aggregation mechanism hides details of data in its source, which not allows internal resources to leave its residence and to be queried for external use. Role-based access control mechanism within the internal dashboard together with aggregation approach in the external dashboard ensures high level of data safety in the current architecture. However, potential data accessibility risks and the lack of preventive measures which assess the consequences of these risks have been identified by Purnama Jati, P. H., the data privacy specialist of VODAN-A (Purnama Jati, P. H., et al., 2022).

First of all, VODAN-A has not implemented any access control mechanisms over data template instances and scientific data collected by these templates in residential *Allegrograph* repositories, which are also known as FAIR Data Point (FDP). The incomplete access control might potentially violate data ownership and stewardship, which leads to the trust risk among VODAN-A stakeholders, as what (van Reisen, M., et al., 2021) argued. Moreover, this deficiency violates FAIR principle which recommends that (meta)data are released with an unambiguous, accessible data usage license and that an authentication and authorization procedure is introduced. Finally, it disobeys GDPR in which data should be processed securely by using technical methods such as encryption and organizational measures such as composing data privacy policy and constraining access to sensitive data.

Current internal VODAN-A dashboard is using role-based access control (RBAC) for managing the visibility of aggregated statistics, whereas it is hypothesized that RBAC is not an optimal mechanism for Allegrograph repositories where the linked ontologies are stored in RDF triples and data query is enabled by SPARQL. Unlike facilities with internal structured organization hierarchy, there is no role distinction in the external linked data sharing network, so it is not appropriate to utilize the role as the only factor to control the access to data exposed externally through SPARQL queries. The further analysis of RBAC is discussed in the findings chapter in detail. Therefore, it is necessary to design an alternative access control over scientific research data stored in FDPs.

1.2 Research Gap

Security and privacy preservation are important topics which deserve our attention in the context of data sharing. Technical solutions to these concerns including the symmetric encryption scheme (Elger, B. S., et al., 2010), the multi-factor authentication (Park, 2011), and pseudonymisation mechanism (Quantin, C., et al., 2011) have been sufficiently researched to manage the electronic health records (EHR). However, a limited number of studies research access control over linked FAIR data related to semantic web technology and ontology philosophy in the academia up to the time. Existing access control models have performed well in most internal information systems and sufficient studies have been carried out in conventional models such as Bell-LaPadula security model, Clark-Wilson integrity model.

These models made sure the data resources are persistent and accurate by preventing these resources from malicious modifications by unauthorized users (Fernández-Alemán, J. L., et al., 2013). In the healthcare domain, (Zhang, R., & Liu, L., 2010) applied anonymous user signatures to maintain the integrity of patients' HER in the cloud computing environment. Moreover, mature mechanisms such as RBAC, ABAC, and SBAC have been adopted by healthcare information systems. For instance, (Røstad, 2008) proposed a more accurate RBAC by separating roles into system roles (i.e., patients, physicians) and user roles which are defined by users themselves; (Peleg, M., et al., 2008) introduced the 'context' to deal with emergent cases where default RBAC policies should be overrided. Nevertheless, there is still the lack of discussions regarding access control over linked ontological metadata in a localized context, especially in the healthcare domain.

The Center for Expanded Data Annotation and Retrieval (CEDAR) project is a workbench for creating scientific data annotated by rich controlled vocabularies in biomedical domain (Musen M., et al., 2015). CEDAR platform plays significant role in VODAN-A. First, VODAN-A has deployed CEDAR as its core data entry point where research data is collected using CEDAR templates which connect to external controlled vocabularies. Second, templates and research data are stored locally and securely at its source by leveraging permission management mechanism of the CEDAR system. Biomedical ontologies are used to model scientific data for the sake of making distinctions among FAIR datasets and creating connection with relevant ones in the data opening/sharing network, whilst Resource Description Framework (RDF) is used to represent and expose FAIR linked data in the machine-readable format. Therefore, it remains unknown whether existing data access control models can satisfy requirements or handle privacy concerns raised by VODAN-A project due to its distributed characteristic and opening nature, which inspires this research to explore an adaptable one among all solutions at present.

In the future, a large volume of scientific research data and healthcare data will be collected using CEDAR templates in compliance with FAIR principle. By the concept of data visiting, VODAN-A envisages these FAIR data is exposed to authorized users (i.e., researchers, clinicians) and processed by allowing their algorithms travelling to local facilities. According to one survey, the majority of scientists (56%) consistently agree that access control mechanisms are required when users make their research data open (Fecher, B., et al., 2015). Based on these considerations, the emphasis of this study is put on the design of one eligible mechanism over SPARQL queries for accessing management of FAIR metadata based on its ontological structure.

1.3 Research Questions

The proposed research questions are centered around the design of an eligible access control mechanism in order to achieve higher level of data ownership and stewardship in VODAN-A based on current situation of access control in VODAN-A repositories.

The design of these interview question follows a broad-to-narrow principal, it starts from researching access control implemented within all functional components of VODAN-A in a wider scope, then dives into a specific component where (potential) data privacy risks have been identified, finally ends with conceptualizing approaches to tackle these challenges or mitigate negative impacts. The research questions and sub-questions are as follows:

RQ1. What are data protection techniques implemented in all components respectively within VODAN architecture?

- What types of (meta)data does VODAN collect and process in each component?
- What types of scientific (meta)data are being protected by which access control in VODAN?
- What is the difference between VODAN internal dashboard and public one in terms of data visibility?

RQ2. What would be the feasible access control mechanisms for Allegrograph-based semantic data queries?

- What data is stored in Allegrograph repository and is queried through SPARQL?
- What are features of AC models in the academia at present?
- What are use cases of these AC mechanisms?

RQ3. What is the difference between stakeholders' opinions discovered in terms of the requirement of semantic data queries?

- What eligible features should OBAC have to control FAIR data queried through SPARQL?

RQ4. How does the design of Ontology-Based Access Control (OBAC) over semantic data queries look like?

- What are components and relationship of them in OBAC architecture?
- What requirements and rules should be documented in OBAC policy?
- How to translate current data use agreement into OBAC policy?
- How to represent OBAC policy in a machine-readable way?
- How to measure the performance of OBAC over SPARQL queries?

1.4 Objectives

The primary goal of this study is to bridge the research gap and privacy concerns in VODAN-A system by proposing a feasible access control based on current dataflow established in the localization architecture. The objectives of each research question are listed below (see Table 1):

Table 1: Research questions and corresponding objectives

	Research Question	Objective
1	What are data protection techniques implemented in all components respectively within VODAN architecture?	To point out problems in VODAN based on assumptions: (1) current data protection techniques cannot fit semantic data queries; (2) there are no authentic ACs in all components; (3) there are ACs but with defects
2	What would be the feasible access control mechanisms for Allegrograph-based semantic data queries?	Interpret the theory of ACs and analyze their use cases to prove that among all other ACs, OBAC is a feasible solution for SPARQL queries for FAIR data
3	What is the difference between stakeholders' opinions discovered in terms of the requirement of semantic data queries?	By comparison, to conclude features OBAC should have in Allegrograph query component of VODAN based on the inputs from in-depth interviews with stakeholders
4	How does the design of OBAC over	To design OBAC architecture including

	semantic data queries look like?	dataflow, components and their relationship, AC policy and its representation

1.5 Research Approach

The methodology in this research includes a qualitative research method and a research design. The data collection starts from the literature review whose aim is to explore data access control mechanisms or models in the academia at present. One of them will be selected as the prototype which becomes the basis of the proposed approach for VODAN-A. The qualitative labelling will be applied to these literatures to extract critical features from existing models to satisfy requirements of VODAN-A. Then, interviews with stakeholders will be carried out to evaluate data protection approaches and privacy concerns in the current VODAN-A architecture. The interviews will be recorded and be converted into transcripts for qualitative labelling as well, envisaged results would be a summary about the safety level of VODAN-A architecture based on different participants' points of view. All processed data will be serving the following research design phase. An implementation plan consisting of the architecture, and the workflow will be proposed for the management of access and permission control. The principle of this design should not only be applicable to the data flow in current VODAN-A architecture, but also be ease of maintenance. The access control model will enhance the data ownership by enabling custom and flexible data accessing policies, it thus strengthens the trust among VODAN-A community.

1.6 Related work

Initially we start to analyze articles published by researchers of VODAN-A to figure out problems or (potential) risks in general and identify the design of access control in the future, the following Table 2 reveals different investigation aspects pertaining to the protection mechanisms in VODAN-A:

Table 2: The comparison between VODAN-A related articles

Literature	**Concern**	**Protection techniques**	**Permission configuration factors**	**Aspects**	**Data governance level**
Van Reisen, et al., 2022	Data ownership	Encryption	Role	Technical, legal, operational	Template, repository, facility, country
Plug, R., et al., 2022	Data ownership, privacy	Data depersonalization, credential validation, authorization, data management plan with specified control patterns	Role	Technical	Residence, country, metadata
Purnama Jati, P. H., et al.,2022	Data ownership, privacy	Data processing agreement, well-documented secure specification	Role distinctions	Legal, operational	Template, repository, facility, country
Stocker, M., et al., 2022	Data interoperability (Cost of duplication)	-	Membership, geography	-	Country
Folorunso, S., et al., 2022	Data interoperability	-	-	Algorithmic	API
Oladipo, F. O., et al.,	Privacy, misrepresentatio	Data management plan, customized training for data	-	Education	Country

2022	n of data, security	stewards			
Akindele, A. T., et al., 2022	Low quality training system	-	-	Education	Country
Basajja, M., et al., 2022	Poor data interoperability in Africa	-	-	Technical	Catalog, metadata
Chindoza, K., et al., 2022	Data integrity, FAIR adoption	Licensing, authorization/authentication	Time	Operational	Country
Luiz Olavo Bonino da Silva Santos, et al., 2022	Data security	Firewall, configuration of FDP index, authentication	Role, index of FDP, data manipulation (read, add, edit)	Technical	Metadata
Van Reisen, M., et al., 2021	Data ownership	Data localization	Location	Technical, legal	Template, repository, facility, country
Mawere, M., et al., 2020	Data ownership	Scrutiny mechanism	Location	Administrative	Country

These articles comprehensively discuss FAIR-related or VODAN-related concerns and solutions from diverse perspectives, covering technologies, data governance, legal frameworks, educations, and other aspects. Basajja et al., argue that the poor data interoperability which is not in line with FAIR guideline and the difficulty of reusing data in Uganda are due to the lack of particular technical solutions, leading to limited actions which are allowed to perform on data (i.e., user can only read data). Similar concern also exists in Zimbabwe, which is proved by Chindoza, K., et al., the authors suggest that the purpose, conditions, and ways of processing data should be licensed in a standardized agreement, and this kind of agreement should be unambiguously represented using human- and machine-readable format.

To address this interoperability challenge, Folorunso, et al., focus on addressing the data interoperability challenge by a generic designing machine learning model from a technical perspective. Another effective approach is education. For example, Oladipo, F. O., et al., and Akindele, A. T., et al., make preparations to the online education project – DISH involving FAIR principle, digital technologies, analytical algorithms, and regulatory frameworks to train data stewards in African countries, which contributes to VODAN-A's demand for qualified data stewardship specialists.

Another concern mentioned in most articles is the deficiency of data ownership in the current phase, which hinders the building of a trustworthy environment for VODAN-A partners, as identified by Plug, R., van Reisen, M., and Purnama Jati, P. H. Corresponding solutions contain administrative process (unambiguous specification, privacy impact assessment), legal approaches (DPA, GDPR), and the permission configuration mechanism. Ultimately, a data control and governance scheme should be conceptualized and implemented, which covers multiple layers ranging from the repository level to the nation level.

1.7 Ethical Consideration

In this study, data collection/processing methodologies such as the design of interview questions and collected data such as the transcripts may involve sensitive information of relevant stakeholders. Therefore, data management plan is designed appropriately to avoid any disclosure of private information. First, a consent form including the background, the objective, and data handling methodologies of this research is well-documented (see Section 3.2.3) and sent to selected participants before interviews. Second, personal information of interviewees will not be exposed, for example, appellations such as “Interview A” will be used to replace their real names during the discussion and analysis in this thesis. Third, all recordings and transcripts will be deleted within one month after this thesis is submitted to the educational system of university.

1.8 Social relevance

The data leakage issue of GGD in the Netherlands raises public awareness that any information systems especially those involve sensitive personal identities should prevent illegal disclosure or malicious theft of data. A possible reason is the lack of qualified cybersecurity experts who can identify vulnerabilities leading to the disclosure of data in the system and understand the consequences of these risks. For example, in the iBestuur Digitalization Conference held on 14 September 2022, several Dutch government officers mentioned the biggest concern they are facing is the insufficiency of human resources within the security department (ibestuurcongres, 2022). Moreover, they also blamed the data leakage problem on the immature mechanism of the risk impact assessment. Another social issue identified is the data outflow without creating value locally in the Netherlands. Private enterprises resell the data from Dutch citizens to tech giants (i.e., Google, Meta) leading to less benefits of data-driven service or research to residents. Despite this challenge, government officers still have a positive attitude towards data sharing especially in the healthcare sector since it can aggregate resources and knowledge from EU member states, which thus facilitate the collaboration and the quick response to the emerging pandemic.

From legal perspective, the security specification including processing/storage techniques, preventive assessment of risks, and permission control mechanisms should be well documented and guided by high-level security policies. However, as pointed out by one of digitalization officers, data processing activities are still not transparent to most Dutch citizens even though GDPR has been applied among EU member states for a long time. They are not aware of how their personal data is processed and stored. Therefore, technical mechanisms as the supplement should be conceptualized to assist the regulatory framework. From technical perspectives, concrete implementation of security models should guarantee that data manipulation activities conform to security policies. These mechanisms should prevent unauthorized modification of data from data processors or any data-driven applications.

To conclude, first, insights from participants of iBestuur Conference confirm that the loss of data value happens in not solely developing countries but developed nations with better infrastructure and regulatory framework. This challenge can be addressed by the concept of data localization proposed by members of VODAN-A. Second, it proves that the necessity of collaboration among EU countries is consistent with VODAN-A’s prospect of facilitating response to the pandemic by knowledge sharing across geographies. Third, the lack of eligible security specialists indicates that mentorship programs are necessary to train data stewards and help all stakeholders form the mindset towards the consistent goal of high-level security in VODAN.

1.9 Thesis Outline

This thesis is structured as follows. In the chapter 2, theories used in this study will be demonstrated in terms of introductions and their relationship, followed by methodologies given in the chapter 3. Subsequently, introduction of Virus Outbreak Data Analysis Network-Africa

(VODAN-Africa) in terms of its functionalities and potential concerns will be presented in the chapter 4. Findings of preliminary interviews and the analysis of VODAN-related articles pertaining to access control models and other data protection strategies identified in VODAN-A architecture will be given in chapter 5. Chapter 6 is the literature analysis pertaining to existing access control mechanisms in the academia and their use in practice. Chapter 7 is the findings of in-depth interviews revealing the requirements of stakeholders when designing ontology-based access control within VODAN-A. Chapter 8 will illustrate the design of ontology-based access control model. Finally, the discussion is given in the chapter 8 and chapter 9 is the conclusion.

2. Theoretical framework

In this chapter, important theories will be introduced, including access control, semantic web, data regulatory framework, and FAIR data principle. In addition, the motivations of selecting these theories and the relationship among these theories would be illustrated. Relationships between theories is demonstrated in the following Figure 1.

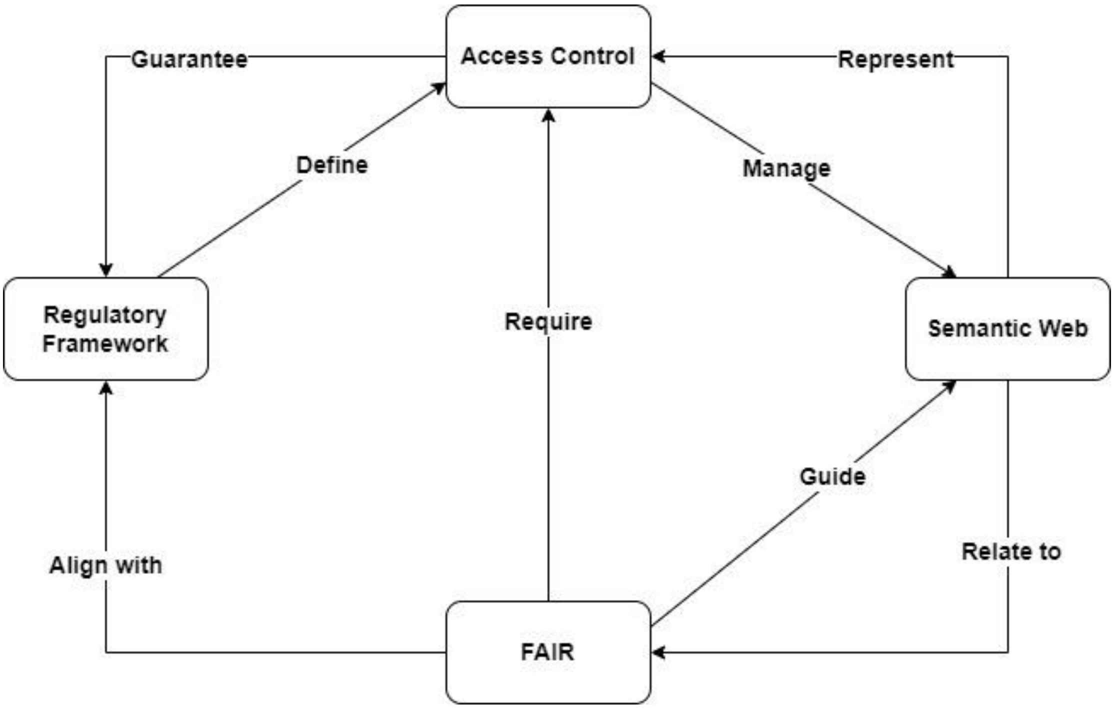


Figure 1: Relationships between theories

2.1 Semantic web

In healthcare domain, the meaning of clinical data is key to the further application of data, which is described by bio-semantics (Plug, R., et al., 2022). However, massive data obtained from observation-based diagnosis is originally unstructured, therefore these unstructured data should be processed in accordance with formalized approaches widely adopted within the domain so that useful information could be extracted for human to interpret and machine to interoperate. Data collection pipeline in VODAN-A is employed with standard templates based on controlled vocabularies to structuralize and give meaning to data, which enables both interpretability and interoperability for machines. This transforming process can be executed by automatic technologies based on ontology.

By creating the relationships among useful information extracted from raw data, information is converted into knowledge. By linking knowledge together, we construct knowledge graph (also known as semantic web) where nodes are representation of knowledge and edges demonstrate relationships between knowledge entities. Semantic web can be visualized for human, and analyzed by machines for knowledge reasoning.

In general, semantic web is commonly presented by Resource Description Framework (RDF). Under this framework, each node in semantic web is the data resource marked by universal resource identifier (URI) while each edge is the relationship between two data resources (Brewster, 2020). Most structured, semi-structured or unstructured data can be mapped to triple in which subject stands for the entity, predicate refers to the attributes of entity, object refers to the concrete value of entity's attribute. For large scale data management, graph database is recommended to store knowledge graphs.

As depicted in Figure 1, the semantic web technology puts FAIR's proposition into practice. On one hand, semantic web ensures data is FAIR in terms of the interoperability and findability: (1) the reasoning capability of semantic web guaranteed the data is interoperable; (2) the linkage property of semantic web makes sure data is findable, especially the use of URI creates a globally unique identifier when querying data across facilities in VODAN-A, which plays an important role to reduce ambiguity of data access (Plug, R., et al., 2022).

On the other hand, FAIR principle offers guideline to produce semantic data with high quality. FAIR requires data to be interoperable, the creation of qualified references (I3 principle) from one dataset to other datasets can be achieved by adding more associated triples on a semantic web. Moreover, the ontology reduces ambiguity of different data sets by introducing rich metadata as controlled vocabularies (I2 principle) to describe the application context of data or indicate the interrelations between these data entities. In this way, better data interoperability between machines is achieved and it decreases the burden of reusing scientific data to benefit both local clinics in African geographies and researchers. In addition, RDF is used in semantic web to represent data in a machine-interoperable manner, which aligns with I1 principal in which a standard language for knowledge presentation is required.

2.2 Access Control

VODAN-A will develop a layer between end-users such as researchers and professionals in healthcare domain. On the basis of the requirement of integrating heterogeneous data resources and delivering consistent data accessing, the control over data accessing levels is required with purpose of transparent data management.

The data accessing gateway is typically implemented by the Application Programming Interface (API), data queries through API should be controlled under richly defined and transparent prerequisites such as reliable authentication/authorization mechanisms at either the facility level or the country level (Plug, R., et al., 2022).

In practice, Trusted Computer System Evaluation Criteria (TCSEC) of U.S. government categorized access control mechanisms into two main streams: Discretionary Access Control (DAC) and Mandatory Access Control (MAC) (Defense, 1986). MAC mechanism refers to a type of access management by which a data resource repository limit the capability of data users to access or perform specific operations based on multiple clearance levels of users which is defined by the repository system (Ware, 1979). Mechanisms such as role-based access control (RBAC) and attribute-based access control (ABAC) are consistent this type of control. In contrast, DAC is a data control method whose accessibility is manually determined and authorized by owners of data resources without any interventions of machines or system stewards (David F. Ferraiolo, D. Richard Kuhn, 1992). Common implementation of DAC is the use of Access Control List (ACL) in early stages and Bell-LaPadula model with the access permission matrix. These mechanisms or models will be further discussed in the Chapter 6.

From macro perspective, each user subjects and each FDP repository as the container of individual data in VODAN-A network are modelled using ontology and are linked in the large network to shape knowledge. Users such as hospital staff and medical researchers can extract knowledge or any piece of data from this network, this is the huge value that semantic web brings. However, as illustrated in Figure 2, if access control mechanism is not implemented, any users (no matter under authorization or not) can reach all nodes defined in this semantic web without restrictions, namely that data sovereignty cannot be guaranteed, and some users' privacy may be violated.

From a micro perspective, protection approaches should also cover any underlying metadata describing each node in this graph since these metadata may indirectly associate with sensitive personal identifiers. Moreover, a finer-grained access control mechanism should also be capable of granting an access decision based on metadata.

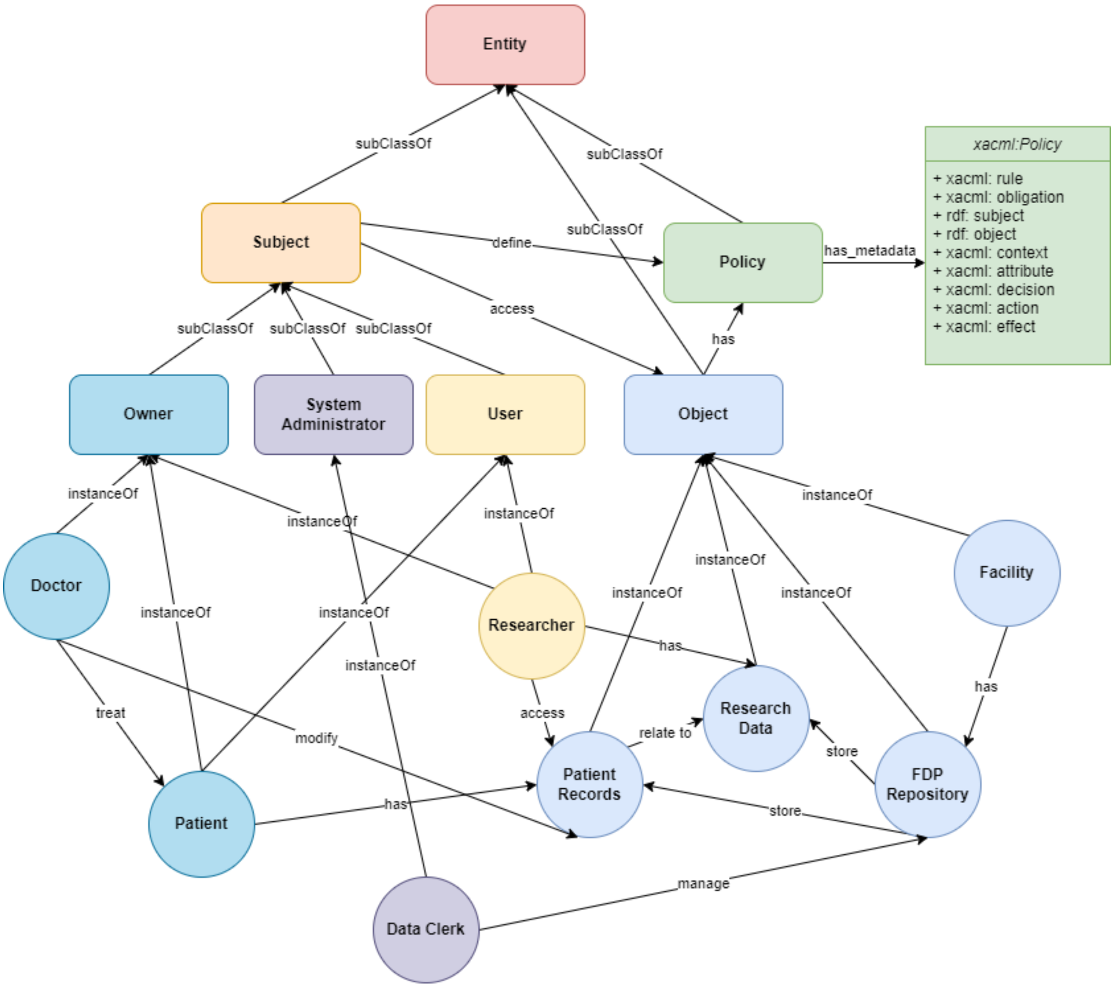


*Figure 2: Partial ontology structure of VODAN-A. Note: this graph does **not** reflect all entities and all relations defined in VODAN-A's ontologies library, only those related to this research are shown. Source: created by Mingyu Huanng, the author of thesis, based on the BioPortal biomedical ontologies repository established by VODAN-A team (https://bioportal.vodana.health/ontologies) and the secondary data from interviews.*

In this research, access control model is proposed based on the ontological semantic web implemented in VODAN-A (Basajja, M., 2022). In semantic web, data together with its attributes and its relationship with other data forms its ontology. Ontology utilizes massive attributes or connections as possible to accurately describe and identify an object in the world. For example – in VODAN-A, biomedical ontologies such as the terminology of diseases or viruses and general ontologies such as terms used in local clinics are defined in the BioPortal. These ontologies are described by controlled vocabularies widely adopted in the healthcare sector. By linking these ontologies, VODAN-A has established its own knowledge graph which makes it possible to discover further knowledge. Access control designed based on this ontological structure of semantic web can enable high flexibility of documenting access control policy.

2.3 Regulatory Framework

The preservation of data ownership using legal approaches is key issue to be considered in VODAN-A. This is because, on one hand, data itself is regarded as a non-tangible object; on the other hand, it is difficult to define legal owners of data. These challenges possibly rely on the jurisdiction of data entry points where data is originally produced and stored (Plug, R., et al., 2022). A reliable data entry point can be referred as the physical infrastructure with a secure data storage and exchange environment. In addition, from legal perspective, the purpose and approaches by which data is handled by data processors should be included in data use agreements in compliance with local legal system.

Access control mechanisms functions as an essential mechanism to complement regulatory framework, and vice versa. Jurisdiction, as a high-level legal approach, can maintain the data sovereignty to some extent. However, another critical problem which constrains the reusability of scientific data is the lack of robust access control mechanisms where data is exchanged under agreed-upon controlled conditions outside the jurisdiction. In this research, access control policies are designed meeting requirements of local data protection regulations.

Repositories aligning with FAIR principle ensure data sovereign in residence. The localization architecture of VODAN-A allows local facilities to customize each detail of deployment to align with local legal framework. In this regard, the deployment of healthcare information management systems across African countries can tailor their strategies to satisfy with local regulation. By keeping data in residential FDP repositories, it only allows to process metadata without the exposure of actual data, and it also hides the details of actual data by aggregation techniques. Data sovereign is preserved; therefore, it complies with GDPR, in this case, data users can process data under the guidance of regulatory frameworks.

2.3.1 GDPR

General Data Protection Regulation is European Union’s new generation of law regarding the data privacy and security (Regulation (EU) 2016/679, 2016), which imposes restrictions on any organizations whose activities involve the processing of personal data, both within EU and outside EU (Wolford, n.d.). It aims at specifying constraints of data processing activities and resolving conflicts between entities such as data processors, data subjects, and data controllers.

Data subject is the owner of data by nature who has the full control of the data (Regulation (EU) 2016/679, 2016). The protection of data ownership, especially one associated with personal privacy is the center of GDPR. In the context of VODAN-A, data subject can be patients who register their personal information in the system. In the future, VODAN-A will expose its data to more research institutes, who can also be considered as data subject who generate experimental or scholarly data. The decision on the exposure of data should be made or delegated to data controllers by data subjects

Data controller is the squad who is assigned the permission to control data owned by data subjects (Regulation (EU) 2016/679, 2016). The controller is responsible for defining the conditions, responsibilities, the usage, and approaches by which data is utilized and is stored by the data processor. The data controller also takes legal responsibilities to offer data use agreements and obtain consent. According to this definition, data controllers in VODAN-A can be clinicians and legal representatives in residence where data is generated. In different geographies, they are required to set rules of data accessing control in relation to the local legal framework and define data sensitivity level depending on the actual context.

Data processor is the entity who is responsible to store and process data on behalf of the data controller (Regulation (EU) 2016/679, 2016). It is emphasized that they should develop a secure data storage infrastructure equipped with the capability to interoperate data efficiently. In

VODAN-A, data clerks in local clinics are assigned as the role as the data processor, mainly taking a duty to maintain the data repository.

2.4 FAIR data principle

In modern society, the occurrence of a large volume of data especially in healthcare sector increases difficulties in managing them. The technological advancement in terms of the computing power, the storage capacity, and the performance of analytical algorithms cannot keep the pace with the requirement of handling with massive data generated within decades. The FAIR data principle was proposed by Wilkinson (Wilkinson, M. D., et al., 2016) including Findability, Accessibility, Interoperability and Reusability principle with intent to enhance the data stewardship in the big data era.

FAIR healthcare data on the semantic web can be traversed and interoperated by analytical algorithms. Effective machine interoperability relies on the realization of findability and accessibility. To achieve this, all VODAN-A facilities must resolve ontologies which are derived from controlled vocabularies and these ontologies must be assigned with unique identifiers (Plug, R., et al., 2022).

FAIR is the supplement to regulatory framework. FAIR’s R1.1 illustrates that data must be released with an unambiguous data usage license, which requires (meta)data to be remained at residences where policies specifying the conditions and prerequisites under which access. This complies with the requirement of GDPR in which data users must clearly inform data owners about the intention of using data. It is also important that the reuse of medical data follows data protection regulations. The data use agreement of each facility should be well-documented.

From legal perspective, it can maintain the data sovereignty to some extent. However, another critical problem which constrains the reusability of scientific data is the lack of technical mechanisms where data is exchanged under agreed-upon controlled conditions outside the jurisdiction. This issue was identified by Stocker et al., they highlighted FAIR data guidelines are supported and implemented by a series of tools, otherwise they will lose effect. For instance, FAIR principle (A1.2) requires authorization and authentication processes offered by access control to protect data from being visited by unauthorized users.

The FAIR data principle and underlying facets are the following (Wilkinson, M. D., 2016):

Data is Findable:

F1: (meta)data is assigned with a globally consistent and persistent identifier

F2: data is described with rich metadata

F3: metadata explicitly include the identifier of the data is describes

F4: (meta)data is registered or indexed in a discoverable resource

Data is Accessible:

A1: (meta)data is searchable by its identifier using a standard data communication protocol (i.e. Http)

A1.1: the protocol is open, free, and universally implementable

A1.2: the protocol allows for an authentication and authorization process, where necessary

A2: metadata is permanently accessible, even when the raw data is no longer available

Data is Interoperable:

I1: (meta)data use a formal accessible, shared, and broadly applicable language for knowledge representation

I2: (meta)data use vocabularies that follow FAIR principles

I3: (meta)data include qualified references to other (meta)data

Data is Reusable:

R1: (meta)data are richly described with a plurality of accurate and relevant attributes

R1.1: (meta)data are released with a clear and accessible data usage license

R1.2: (meta)data are associated with detailed provenance

R1.3: (meta)data meet domain-relevant community standards

The ‘FAIRification’ workflow following FAIR Guidelines is put forward by (Jacobsen et al., 2020). As depicted in Figure 3, this approach converts less structured (meta)data into semantic FAIR data through three coherent steps including pre-FAIRification, FAIRification, and post-FAIRification. The milestones that VODAN-A team has already achieved are centered around step 1 (pre-FAIRification) and step 2 (FAIRification).

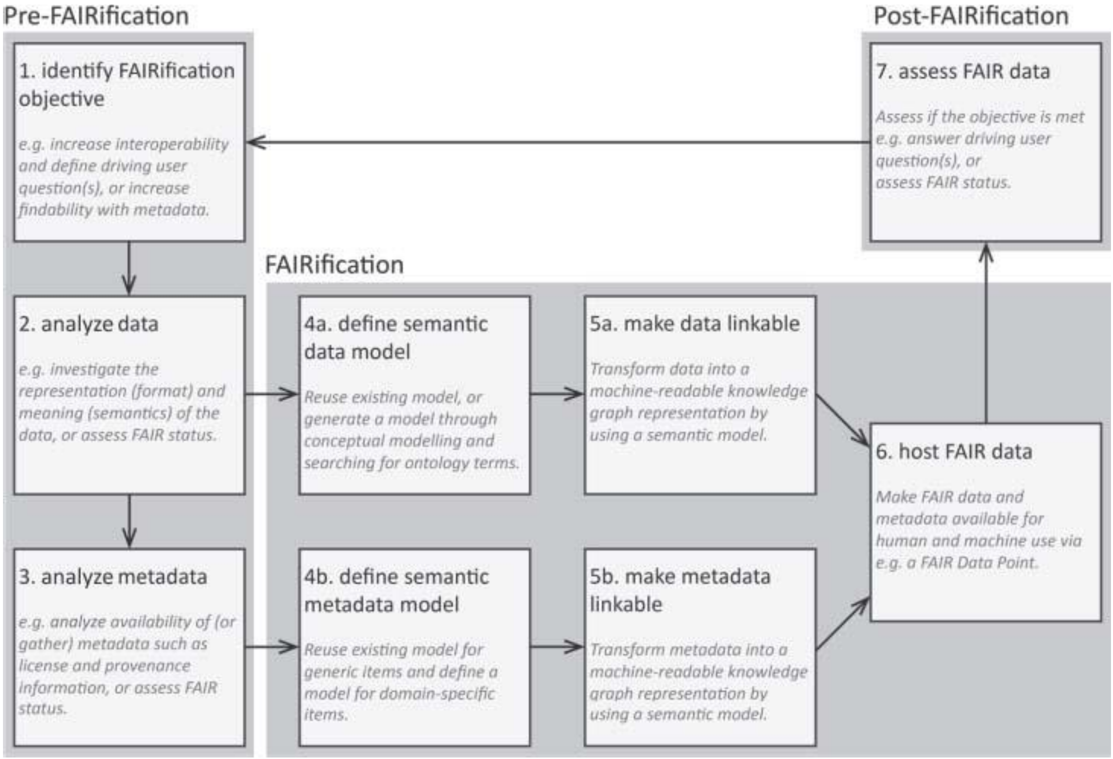

Pre-FAIRification
Post-FAIRification
1. identify FAIRification objective
e.g. increase interoperability and define driving user question(s), or increase findability with metadata.
7. assess FAIR data
Assess if the objective is met e.g. answer driving user question(s), or assess FAIR status.
FAIRification
2. analyze data
e.g. investigate the representation (format) and meaning (semantics) of the data, or assess FAIR status.
4a. define semantic data model
Reuse existing model, or generate a model through conceptual modelling and searching for ontology terms.
5a. make data linkable
Transform data into a machine-readable knowledge graph representation by using a semantic model.
6. host FAIR data
Make FAIR data and metadata available for human and machine use via e.g. a FAIR Data Point.
3. analyze metadata
e.g. analyze availability of (or gather) metadata such as license and provenance information, or assess FAIR status.
4b. define semantic metadata model
Reuse existing model for generic items and define a model for domain-specific items.
5b. make metadata linkable
Transform metadata into a machine-readable knowledge graph representation by using a semantic model.

Figure 3: FAIRification workflow. Source: (Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L. O. B., Mons, B., Schultes, E., Roos, M., & Thompson, M. , 2020)

In this research, focus is put on the third step (post-FAIRification) in which FAIR data is discoverable through the Internet and can be accessed by authorized data processors with strict access control mechanisms, both administrative and technical.

2.5 Conclusion

In this chapter, four theoretical frameworks and their interrelationships are discussed. These theories are not independent, they offer support or guidance each other. Suitable access control over semantic web is researched in this study for various motivations. First of all, it aims to align with the Readability principle from FAIR, in which raw data is described by rich metadata. There are equivalent concepts in VODAN project. Currently, VODAN is exploiting the CEDAR template in which each field can be populated by a collection of specific terms defined in BioPortal. In this context, fields are considered to be the metadata describing the meaning of dataset, which is equal to the subject of triples, while controlled vocabularies are equivalent to object of triples. By this process, data collected via CEDAR's templates is naturally aligned with the readable principle. Secondly, VODAN is dealing with huge volume of knowledge in the healthcare domain whose ontology in terms of classes, sub-classes, and relationship of classes is complex. Semantic web technology is driven by massive linked data so it can support this requirement of handling complicated data in life science. Thirdly, the nature of VODAN is localized data curation platform, it desires to engage more healthcare facilities in Africa to share data as possible while data itself still store at its sources.

3. Methodology

The research will carry out a comprehensive literature review as a theoretical foundation to gain insight into existing access control mechanisms. Followed by the research design in which a tailored access control will be proposed, interviews relating to ethnography will be designed to engage relevant stakeholders of VODAN-A such as technical team, country coordinators, data clerks in residence, and policy/security officers.

3.1 Literature Reviews

The purpose of retrieving literatures is to investigate as much research about access control in terms of theories and practical applications in the academia as possible. By comparison between these access controls and requirements of VODAN-A- in terms of the use of semantic ontologies as well as distributed storage/hosting, the selection scope of these access controls will gradually narrow down to a smaller range and one of them is going to be selected ultimately as a prototype used in the design step.

Also, useful traits of other access control technologies will be integrated into the prototype. Some features or rationales of access control models are displayed in the Table 3. Among all existing models or mechanisms, SBAC has similar features as OBAC. In this research, we use ontologies to model context or situation information and aim to leverage the reasoning capability of semantic web to extract knowledge. Afterwards, we propose that machines can generate new policies based on this knowledge. Meanwhile, OBAC also refers to a mechanism to protect the internal ontologies developed by VODAN team since these ontologies involved group knowledge. In this case, we use role distinctions, interval time, location, and other associated metadata as criteria to define the access control policies to protect the internal ontologies. Therefore, We hypothesize that Ontology-Based Access Control (OBAC) is applicable to FAIR data queries based on ontological representation and linkage structure of current data in VODAN-A. OBAC is independent from data resources, it can be thus reused and embedded in other applications where data accessing management is required (Brewster, 2020). OBAC would be more efficient in the scenario where data is initially expressed with the ontology. This saves the time of data transformation.

Table 3: The comparison of access control associated literatures

Literature	Ontology/ metadata relevance	Protection mechanism	Permission configuration Factors	Action modes	Elements	Domain
(Bell, D. E., & La Padula, L. J., 1976)	No	Clearance (user + data)	Security level	Append, read, write	Subject, action, object	Military
(Biba, 1977)	No	Clearance (user + data)	Integrity level	Read, write, execute	Subject, action, object	Military
(Clark, D. D., Wilson, D. R., 1987)	No	Well-formed transaction	Integrity level	Transaction procedure (a mix of read/write/execute)	Subject, transaction, object	Business
(Brewer, D. F., Nash, M. J., 1989)	No	Separation of duty	Integrity level	Transaction procedure with duties separated	Subject, transaction, object	Finance, banking
(David F. Ferraiolo, D. Richard Kuhn, 1992)	No	RBAC	Role distinction	Transaction	Subject, transaction, object, role	Civilian government

(Khan, M. F. F., Sakamura, K., 2012)	Yes	RBAC + ontology	Role, environment	Transaction	Subject, transaction, role, object, context	Healthcare
(Motta, G. H., & Furuie, S. S., 2003)	Yes	RBAC + Context	Role, context	Read, write, execute	Subject, action, role, object, context	EHR Management
(Peleg, M., Beimel, D., Dori, D., & Denekamp, Y., 2008)	Yes	SBAC	Role, emergent situation	Read, write, execute	Subject, action, object, situation, role	EHR Management
(Shen, H. B., & Hong, F., 2006)	Yes	ABAC	Attributes/ Metadata	Append, read, write	Subject, action, object, attributes	Web service
(Priebe, T., et al., 2006)	Yes	ABAC with ontology	Attributes	Read, write, execute	Subject, action, object, attributes	-
(Padia, A., et al., 2015)	Yes	ABAC with ontology	Attributes	Read, write, execute	Subject, action, object, attributes	Publishing
(Masoumzadeh, A., & Joshi, J. B., 2011)	Yes	OBAC	Ontology properties	Read, write, execute	Subject, predicate, object	Social network analysis
(Kiran, G. M., & Nalini, N., 2020)	Yes	OBAC	Ontology properties	Read, write, execute	Subject, predicate, object	Cloud computing
(Brewster, 2020)	Yes	OBAC	Ontology & metadata	Read	Subject, predicate, object	Civil governance

Literature reviews about available access control policy standards as well as relevant technical frameworks are also performed. Due to the lack of resources about OBAC, the retrieve strategy is adapted to expand the search scope for more indirect articles which can provide useful guidelines. Moreover, website resources such as technical tutorials are also included as supplementary materials to find required information. As a result, XACML standard and Authzforce framework are selected as the guidance of designing OBAC for VODAN-A. The details of them are elaborated in the research design section.

3.2 Interviews

The design of interview follows guidelines of an ethnographic research. The ethnographic research is an approach to investigate a problem by engaging participants who have points of view and knowledge in specific domains (Van Reisen, et al., 2022). It aims at solving the problems which require comprehensive definition and structure (Björklund, 2013).

3.2.1 Objectives

There are several research questions to be addressed through interviews: (1) the use of access control or other equivalent protection techniques in each functional component of current VODAN-A MVP; (2) opinions of stakeholders towards the necessity of access control; (3) contents to be written in data accessing policies aligning with requirements.

The first task is to give comprehensive assessment of VODAN-A MVP architecture in terms of its data category, data flow, existing data access control implemented in each functional module. Then, based on these information, potential data accessing risks would be diagnosed.

These data are assumed to collect from interviews with the technical team of VODAN-A.

The second task focuses on research current data protection strategies established in facilities of member countries at the facility level, followed by the overall willingness to introduce access control in current architecture at the country level. Therefore, interview participants would be country coordinators or data clerks in facilities.

The third task is centered around the composition of data accessing policies. Information including the obligations, accessing conditions, and regulation compliance would be collected. Thus, data privacy/security specialists of VODAN-A team will be selected as interview participants.

Moreover, stakeholders will be engaged during the whole research design in order to ensure proposed OBAC prototype would be adopted, follow-up interviews will thus proceed to acquire feedbacks of stakeholders across member countries (local clinics, technical team, privacy specialists) in terms of envisaged features.

Processed data from interviews as well as findings from literature reviews, will be used to fill missing pieces in the puzzle of proposed access control prototype.

3.2.2 Interview design

In the previous section, several categories of participant are identified depending upon objectives of research questions. The design of interview also follows several principals.

Factors such as participants' roles and knowledge will be taken into consideration so that bias could be mitigated. Feedbacks of the same problem are likely to vary among different stakeholders, for example, technical team may implement access control in current internal dashboard, but data clerks still can access data across facilities. To handle this, comparison groups are set to validate the results gained from another one.

In addition, the definition of terminology should be explained to participants without ambiguity to avoid misunderstanding. For instance, policy means a set of ideas of what to do in particular situations that has been agreed to officially by a group of people in common sense according to the definition in the dictionary; while data accessing policy in this study refers to machine-readable files with XACML syntax, which consist of a collection of rules, obligations and advice fulfilled by data processors when accessing data.

Interviews will also cover participants from different countries to make sure general consensus is reached among VODAN-A community. Table 4 elaborates the target groups and motivations.

Table 4: Interview participant group and selection motivation

Participant Group	Selection Motivation
Data stewards	(1) To investigate the data visibility from data processors' perspective.
VODAN-A security/privacy specialist	(1) To explore the envisaged features of access control; (2) To identified data to be protected; (3) To figure out current administrative approaches of data protection; (4) To explore what to write in accessing policies.
VODAN-A technical team	(1) To research current access control established in internal dashboard from technical perspective; (2) To find out propositions about access control over SPARQL queries through Allegrograph.
Relevant stakeholders	(1) To discover point of views towards the necessity of

	access control in different countries.
Digitalization conference	(1) To discover social concerns in terms of the data governance and security at the administrative level

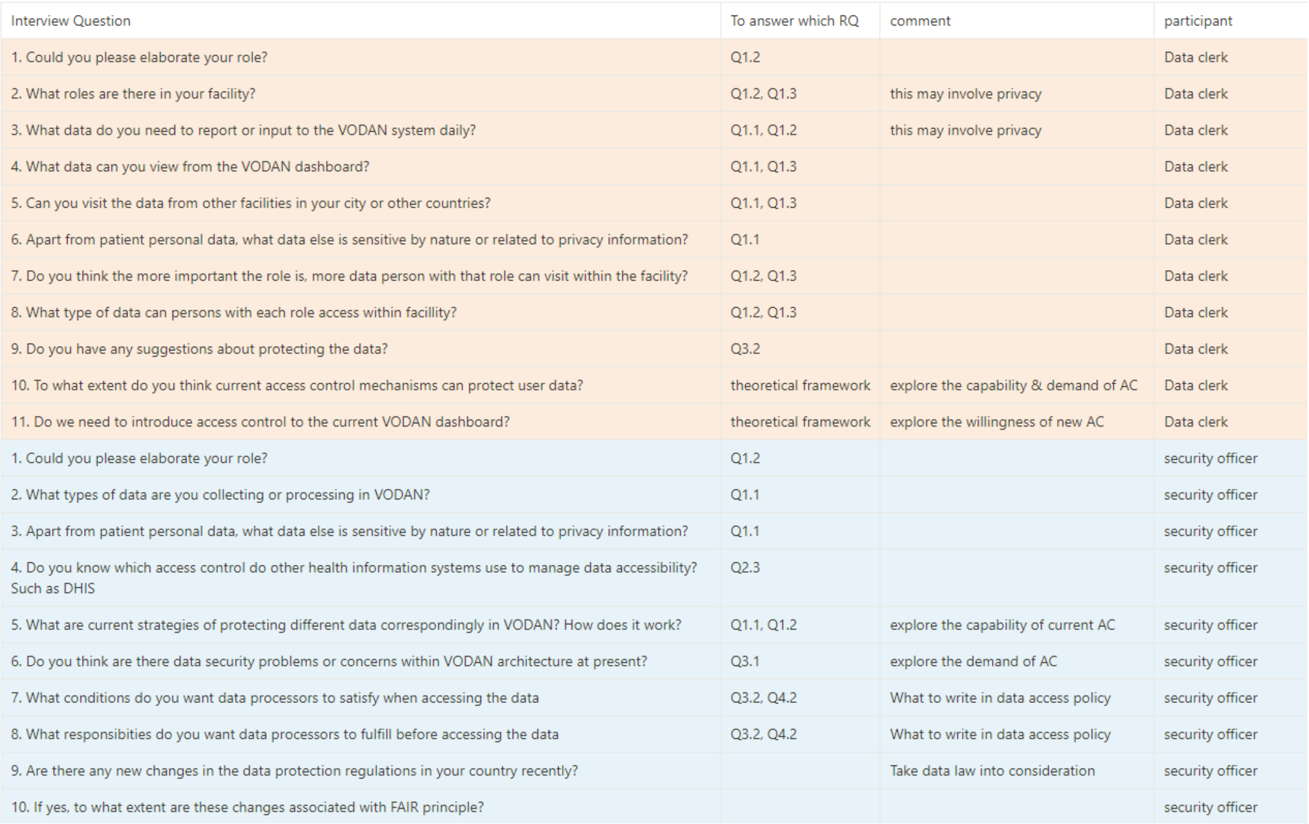
3.2.3 Interview questions

Interview questions are designed to answer proposed research questions and are tailored to specific participant groups. Complete interview questions are listed in the Appendix C. Several examples (see Figure 4) as well as a consent statement are as follows:

Thank you very much for your time as the participant of this interview. My name is Mingyu Huang, and I am working on one project of improving data ownership in local facilities within VODAN by the appropriate use of access control models. This interview will take around 40 minutes.

The general aim of this interview is to achieve a high quality of access control by investigating current data protection mechanisms on data pipelines of VODAN and designing a new one especially tailored to the FAIR data queries. The outcome of this project can be shared with you if you are willing to know.

Several questions will be asked, there are no correct or wrong answers, please feel free to elaborate your opinions regarding any information that you think is relevant to specific questions. This interview is fully anonymous, and your personal information would be protected in a proper way without sharing with others. However, some questions may involve sensitive information, in case you are not willing to answer, you do not need to answer. Furthermore, this interview will be recorded and be converted into transcripts for qualitative labelling and further analysis, all recordings and transcripts will be deleted within one month after final report is submitted to the educational system of Leiden University.



Interview Question	To answer which RQ	comment	participant
1. Could you please elaborate your role?	Q1.2		Data clerk
2. What roles are there in your facility?	Q1.2, Q1.3	this may involve privacy	Data clerk
3. What data do you need to report or input to the VODAN system daily?	Q1.1, Q1.2	this may involve privacy	Data clerk
4. What data can you view from the VODAN dashboard?	Q1.1, Q1.3		Data clerk
5. Can you visit the data from other facilities in your city or other countries?	Q1.1, Q1.3		Data clerk
6. Apart from patient personal data, what data else is sensitive by nature or related to privacy information?	Q1.1		Data clerk
7. Do you think the more important the role is, more data person with that role can visit within the facility?	Q1.2, Q1.3		Data clerk
8. What type of data can persons with each role access within facillity?	Q1.2, Q1.3		Data clerk
9. Do you have any suggestions about protecting the data?	Q3.2		Data clerk
10. To what extent do you think current access control mechanisms can protect user data?	theoretical framework	explore the capability & demand of AC	Data clerk
11. Do we need to introduce access control to the current VODAN dashboard?	theoretical framework	explore the willingness of new AC	Data clerk
1. Could you please elaborate your role?	Q1.2		security officer
2. What types of data are you collecting or processing in VODAN?	Q1.1		security officer
3. Apart from patient personal data, what data else is sensitive by nature or related to privacy information?	Q1.1		security officer
4. Do you know which access control do other health information systems use to manage data accessibility? Such as DHIS	Q2.3		security officer
5. What are current strategies of protecting different data correspondingly in VODAN? How does it work?	Q1.1, Q1.2	explore the capability of current AC	security officer
6. Do you think are there data security problems or concerns within VODAN architecture at present?	Q3.1	explore the demand of AC	security officer
7. What conditions do you want data processors to satisfy when accessing the data	Q3.2, Q4.2	What to write in data access policy	security officer
8. What responsibities do you want data processors to fulfill before accessing the data	Q3.2, Q4.2	What to write in data access policy	security officer
9. Are there any new changes in the data protection regulations in your country recently?		Take data law into consideration	security officer
10. If yes, to what extent are these changes associated with FAIR principle?			security officer

Figure 4: Interview questions for data clerk and security/privacy officers respectively

3.3 Research design

For ensuring proposed OBAC model to be adopted, the design would follow the advisory in terms of technical guidance and data protection law compliance from experts (technical team, privacy officer) in VODAN and stakeholders (data stewards, country coordinators) from VODAN-A member countries. Also, components in our proposed architecture will refer to similar frameworks or projects such as the architecture of Automatic Contracting Tool from the smashHit project (Chhetri, T. R., et al., 2022). Moreover, frameworks or solutions which driven the implementation of OBAC are also discussed, such as XACML, ODRL, CDMM, SPL frameworks.

4. Virus Outbreak Data Network-Africa (VODAN-Africa)

4.1 Phase 1: to address the outflow of valuable data by localization concept

Basajja et al. identified that a key problem commonly existing in most African countries is the outflow of data without creating better healthcare service for local communities especially those vulnerable ones (Van Reisen, 2022). This phenomenon is witnessed by the fact that data produced in local health facilities is delivered to the ministry of health for administrative affairs without being utilized at the points of healthcare, and is even removed from Africa to benefit somewhere else outside this continent (Basajja, M., Nambobi, M., 2022), which further increases the global discrepancy of healthcare quality. This urgent issue is due to the lack of robust data curation strategies, which leads to the design of an innovative digital architecture under the guidance of FAIR principles, namely Virus Outbreak Data Network - Africa (VODAN-A). In the first phase of VODAN-A, FAIR data points are deployed in selected health facilities equipped with a data visualization dashboard to reposit and reuse data locally. This process is also called data localization. The concept of data localization allows data to be stored at its sources without being exposed to external parties, which requires the engagement of the locales.

Up to the time of writing this thesis, data interoperability experiment of FAIR Data Points had been carried out in two countries by (Basajja et al., 2021): one as the user endpoint was deployed at Leiden University Medical Centre in the Netherlands, while another one as the data entry point was situated at Kampala International University in Uganda. The outcome proved that conceptual design of the FAIR-based digital architecture allows the data query at the facility, country and continent level while retains data at its local entry points, which preserves the data ownership. The Figure 5 shows the conceptual design of VODAN project.

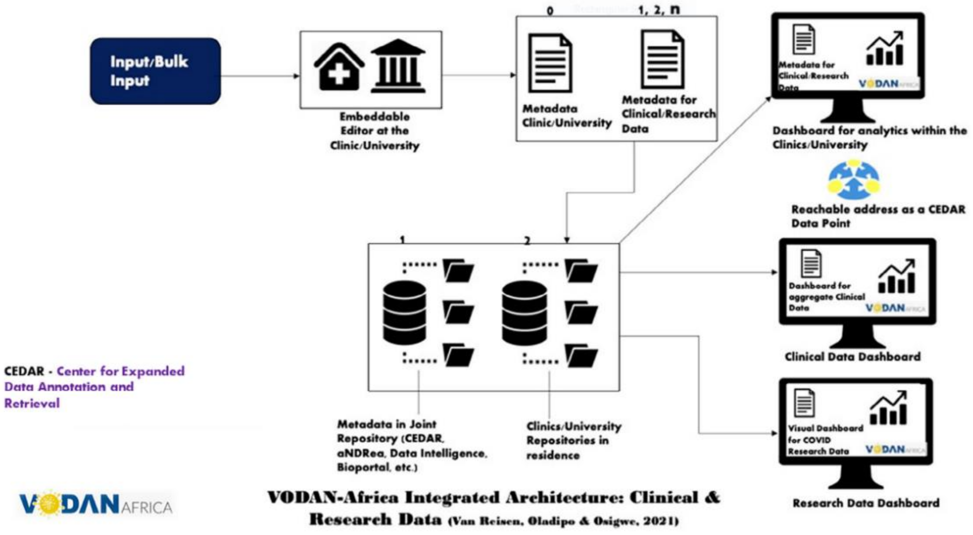
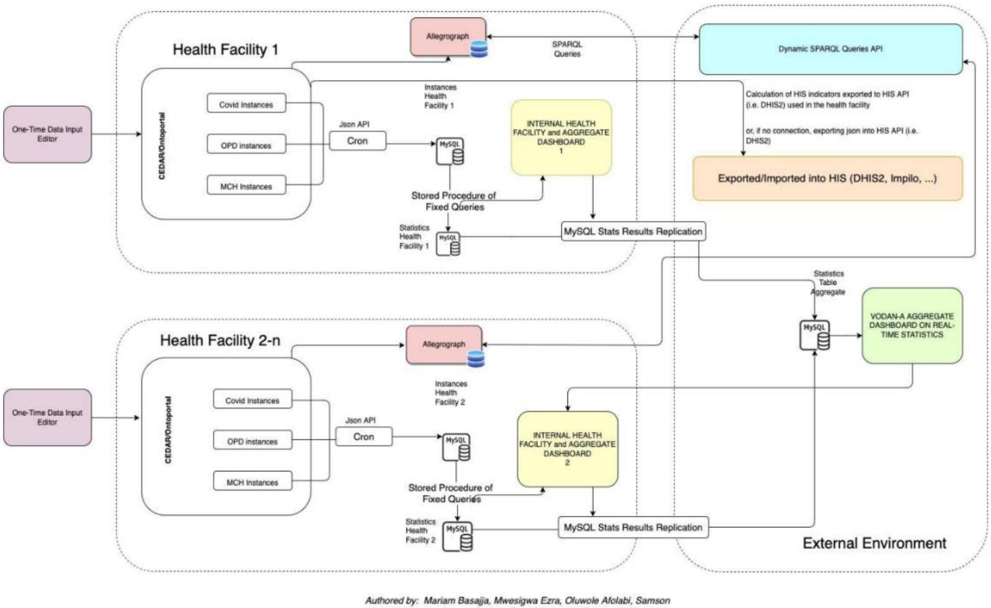

Input/Bulk Input
Embeddable Editor at the Clinic/University
0
1, 2, n
Metadata Clinic/University
Metadata for Clinical/Research Data
Metadata for Clinical/Research Data
VODAN AFRICA
Dashboard for analytics within the Clinics/University
Reachable address as a CEDAR Data Point
1
2
Dashboard for aggregate Clinical Data
Clinical Data Dashboard
Visual Dashboard for COVID Research Data
Research Data Dashboard
CEDAR - Center for Expanded Data Annotation and Retrieval
Metadata in Joint Repository (CEDAR, aNDRea, Data Intelligence, Bioportal, etc.)
Clinics/University Repositories in residence
VODAN AFRICA
VODAN-Africa Integrated Architecture: Clinical & Research Data (Van Reisen, Oladipo & Osigwe, 2021)

Figure 5: VODAN-A localization architecture. Source: (van Reisen, M., et al., 2021)

4.2 Phase 2: to address the lack of data sovereignty by the VODAN-A MVP

In addition to the concern pertaining to the loss of value extracted from local data, ambiguous data interoperability is caused by unclear expression of data ownership in existing digital health information systems, which is placed at the top of agenda in the second phase of VODAN-A project. The objective of the next step is to build up or enhance the trust among existing or potential participants of VODAN-A. According to van Reisen et al., the production of structured data tightly relates to the pipeline in which data is created on the basis of a localized setting which is the outcome of social classification (Van Reisen, 2022). Data sovereignty (also described as data ownership) in the localized context is defined as the right of data subjects to impose the control over digital data which is created and reposited in a geographical location. Currently, data sovereignty in African geographies is constrained by existing monopolistic platforms such as DHIS2, which separate data from the sovereignty of locations where data is created (Mawere, M., & van Stam, G., 2020). This phenomenon extremely undermines the capability of the locale to process and host data, leading to the data resources depreciating in value in residence and creating value somewhere else (van Reisen, M., et al., 2021).

From technical perspective, those deficiencies and critical issues summarized in phase 1 promote the further implementation of a new localization architecture in the phase 2, namely VODAN-A-MVP. The following technical implementation of this design was led by Oladipo et al. and carried out by the international collaboration of engineers. Folorunso et al. (Folorunso, S., 2022) put forward a workflow where analytical tools can be applied to data of each FAIR Data Point. (Akindele, A. T., et al., 2022) and (Oladipo et al., 2022) reported work that the technical team has made to build up the capacity of localized data storage and processing. The Figure 6 illustrates the data exchange architecture of VODAN-MVP.


Health Facility 1
Allegrograph
SPARQL Queries
Dynamic SPARQL Queries API
One-Time Data Input Editor
CEDAR/Ontoportal
Covid Instances
OPD instances
MCH Instances
Instances Health Facility 1
Json API
Cron
MySQL
INTERNAL HEALTH FACILITY and AGGREGATE DASHBOARD 1
Calculation of HIS indicators exported to HIS API (i.e. DHIS2) used in the health facility
or, if no connection, exporting json into HIS API (i.e. DHIS2)
Exported/Imported into HIS (DHIS2, Impilo, ...)
Stored Procedure of Fixed Queries
Statistics Health Facility 1
MySQL Stats Results Replication
Statistics Table Aggregate
VODAN-A AGGREGATE DASHBOARD ON REAL-TIME STATISTICS
Health Facility 2-n
Allegrograph
One-Time Data Input Editor
CEDAR/Ontoportal
Covid Instances
OPD instances
MCH Instances
Instances Health Facility 2
Json API
Cron
MySQL
INTERNAL HEALTH FACILITY and AGGREGATE DASHBOARD 2
Stored Procedure of Fixed Queries
Statistics Health Facility 2
MySQL Stats Results Replication
External Environment

Authored by: Mariam Basajja, Mwesigwa Ezra, Oluwole Afolabi, Samson Yohannes Amare, Ruduan Plug, Mirjam van Reisen, 2022

Figure 6: The architecture of VODAN-A MVP. Source: authored by Mariam Basajja et al.

4.3 Value propositions

The core competitive advantage of VODAN-Africa is its capacity of reuse and interoperating research data following the FAIR guideline. It increases the data quality by leveraging bio-ontologies with rich metadata to model research data in healthcare domain. Metadata such as the experiment context, purpose, and expected results enables high interpretation for both humans and machines. Moreover, it enhances the data reusability. By linking different ontologies together, the semantic web is formed, which helps both researchers and machines understand the 'meaning' of original data, thus makes it more convenient to reproduce the experiment and to reason new knowledge.

The broad cooperation with interdisciplinary experts and concerned stakeholders is another advantage of VODAN-Africa, which enables fast problem-shooting and innovation in a diverse and complex context such as the heterogeneity of various regulatory frameworks. (Amare, Y. S, et al., 2022)[1] stated that the context diversity should be taken into consideration in order to implement an alternative digital platform. To be specific, the engagement of local practitioners and engineers indeed realizes the data localization strategy in which they are responsible for implementing and maintaining the localized data repository.

4.4 Next phase: to enhance the data sovereignty in the data visiting environment

In the upcoming phase, the team will focus on the capacity construction of data visiting, which allows analytical algorithms and applications to access the data locally without making data leave its sources. Therefore, a more robust data governance scheme is required, from both operational and technical perspectives.

However, the vulnerability of data sovereignty maintenance has been identified in the current MVP architecture. The deficiency of data ownership can be interpreted as the fact that each data processor is able to access data without any restriction in the current architecture. For example, Purnama Jati et al. (Purnama Jati, P. H., 2022) identified issues regarding the data accessing control, they pointed out that users still can view aggregated statistics in VODAN-A's internal dashboard without sufficient restriction. Therefore, we hypothesize that the access control mechanism implemented in current VODAN-A system is not robust, there is still a gap between the long-term goal where data sovereignty is fully controlled by data entry points and current access control strategy in terms of high flexibility and customization.

This concern can be solved by both administrative and technical approaches. Regarding administrative operations, solutions such as documenting data processing agreement and introducing privacy impact assessment (PIA) are recommended by the privacy specialist. From technical perspective, (Purnama Jati, P. H., et el., 2022) also suggest that access control mechanism should be designed to support the execution of high-level administrative solutions, leading to the objective of this research – the discovery of an access control mechanism, which can satisfy meet requirements in facilities and respond to changing demands in the locale.

5. Data Protection Approaches in VODAN-A

In this chapter, data access control and other data protection mechanisms implemented in VODAN-A-MVP will be discussed. Moreover, several categories of data that VODAN-A is handling with are analyzed in terms of their sensitivity.

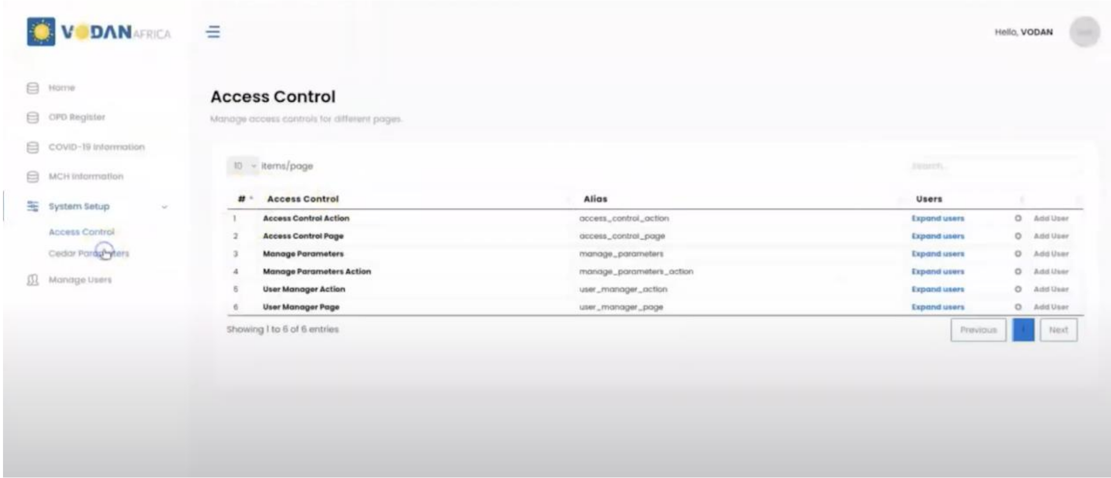
5.1 Access control in VODAN-A

Currently, VODAN-A has implemented one access control mechanism to limit the visibility of web pages of its dashboard based on users' identity. As depicted in Figure 7, current accessing control strategy focuses on the visibility control of various pages in this dashboard. Each web page is regarded as separate resource to be visited, while single user belongs to a data processor whose permission is authorized once they have been added to the user group of specific pages. Based on this observation, we hypothesize that VODAN-A is using RBAC mechanism because roles are group oriented as discussed by (David F. Ferraiolo, D. Richard Kuhn, 1992). Using this mechanism, system administrators can realize data accessing management in a straightforward way by manually adding or removing users, which plays an important role in controlling a small amount of resource at the beginning phase.

However, on one hand, the increase in either the number of active users or the scale of this system in terms of functional modules and web pages will pose a challenge to data accessing management and maintenance, especially in VODAN-A community where a large volume of data is collected across countries.

On the other hand, current strategy only limits permission to view dashboards while it has not completely imposed control over data itself. This "all or nothing" approach cannot achieve data governance at a granular level and satisfy requirement of data localization. That is, once users with certain privileges pass the authentication process, they can view all resources falling into the category that their roles are allowed to visit. Only if a system includes granularity of control, it can be certified with the top security level - Division A in accordance to the TCSEC requirement (Vincent C. Hu, et al., 2017). Various categories of data may require different levels of stewardship and legal preservation, as data in terms of sensitivity is heterogeneous by nature. Healthcare data involve sensitive information reflecting the historical, current, or future physical or mental health states of data subjects, they thus deserve the highest level of protection according to most legal frameworks such as GDPR and local data regulations across African geographies. For example, according to Recital 35 of GDPR, sensitive healthcare data refer to preliminary information derived from the medical testing or examination of a human body while these data can also be secondary clinical diagnosis records produced by professional physicians based on aforementioned preliminary data. From technical perspective, it thus requires a more precise and flexible access control mechanism centered around the data, which can dynamically define data accessing policies based on multiple factors (i.e. roles of data processors, sensitivity level of data, legal contexts across nations).

According to Mawere and Van Stam, data sovereignty in Africa continent is currently framed on the basis of national and sub-national levels in compliance with local legal frameworks (Mawere, M., & van Stam, G., 2020). Therefore, data accessing control strategies should align with this context to protect data sovereignty defined at different levels. Because of the localization nature of VODAN-A, it also requires each locale to define data accessing policies and maintain policy repositories in order to tailor their own data accessing control strategies to local regulations. However, localization does not mean permissive data processing without any guidelines, these customized solutions should be built on a commonly agreed-upon framework in VODAN-A community.



Access Control

Manage access controls for different pages.

10 items/page

Search...

#	Access Control	Alias	Users	
1	Access Control Action	access_control_action	Expand users	○ Add User
2	Access Control Page	access_control_page	Expand users	○ Add User
3	Manage Parameters	manage_parameters	Expand users	○ Add User
4	Manage Parameters Action	manage_parameters_action	Expand users	○ Add User
5	User Manager Action	user_manager_action	Expand users	○ Add User
6	User Manager Page	user_manager_page	Expand users	○ Add User

Showing 1 to 6 of 6 entries

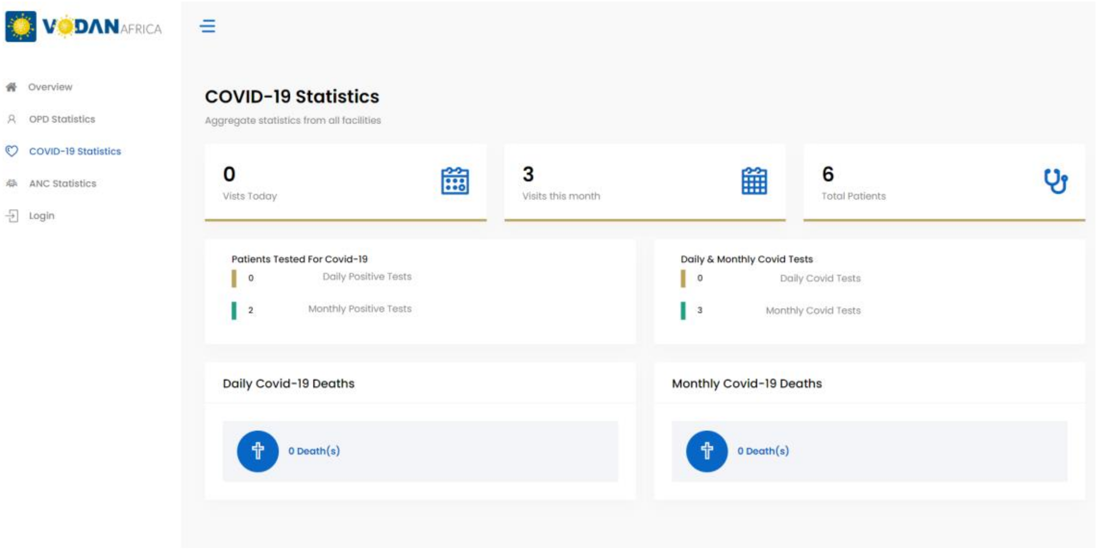
Figure 7: Screenshot of access control in current VODAN-Africa system. Source: 68th VODAN Africa & Asia Webinar

5.2 Data protection approaches in VODAN-A

From data localization's perspective, it allows data is stored at the locations where it is produced. Apart from the access control mechanism implemented in VODAN-A currently, other mechanisms such as data anonymization in the form of aggregation, data pseudonymisation, and data synthesis are identified through investigation. These mechanisms make sure private data is under good protective condition as well as any exemptions associated with data ownership are also strictly controlled.

5.2.1 Aggregation

GDPR is not applicable to fully anonymous data (Recital 26, GDPR), therefore depersonalization is a practical approach for VODAN-A to comply with this regulation. It involves removing either direct or indirect identifiers. In healthcare services, patients may be directly identified based on their name, telephone numbers, email address and photographs when they register for these services. Patients can also be identified by indirect information including medical treatment, diagnosis records. For instance, in order to make room for the reuse of scientific data, VODAN-A is currently using aggregation approach to depersonalize data without the exposure of data subject (Plug, R., et al., 2022). This anonymization process firstly removes any direct and indirect connections between personal data and data subject, then assigns a new URI to data resources without revealing any information regarding data subjects. As a consequence, aggregated data does not associate with any specific identified or identifiable person so they are not considered to be sensitive data any more, which can be thus accessed and processed for the research purpose. As depicted in Figure 8, only aggregated statistics reflecting the situation of COVID in certain areas are displayed in the public dashboard in which individual personal data is hidden and is reposited securly in residence.



COVID-19 Statistics

Aggregate statistics from all facilities

0 Vists Today

3 Visits this month

6 Total Patients

Patients Tested For Covid-19

0	Daily Positive Tests
2	Monthly Positive Tests

Daily & Monthly Covid Tests

0	Daily Covid Tests
3	Monthly Covid Tests

Daily Covid-19 Deaths

0 Death(s)

Monthly Covid-19 Deaths

0 Death(s)

Figure 8: Screenshot of aggregated statistics on VODAN-Africa public dashboard. Source: https://dashboard.VODAN-A.health/covid-statistics

5.2.2 Pseudonymisation

Pseudonymisation is a process of converting identifiable personal data into a unique artificial identifier by approaches such as encryption. Personal information which is decoupled from the original data is stored at secure locations separately. The difference between pseudonymisation and anonymization is the way that sensitive identifiers are handled: anonymization totally removes any sensitive parts, so data subjects are inaccessible and traceable, while pseudonymisation allows removed information to be stored separately under strict control, which still remains the possibility of tracing data subject indirectly if techniques such as decryption are used, or additional information is provided. The pseudonymisation is another important personal data protection mechanism because it makes data temporarily anonymous during a limited context such as extenuating circumstances and authentication process (Plug, R., et al., 2022).

5.2.3 Synthesizing

Apart from protection mechanisms for individual records of personal data, data synthesizing is another approach. Synthetic data can be derived from original raw data based on its statistical attributes such as correlation, distribution, average value, and variance. Statistical attributes, in this case, can be regarded as the metadata providing insights of raw datasets, which are important to reproduce research projects with statistical purposes. Data synthesizing process acquires secondary information from metadata by running algorithmic applications and leverages this information to generate new data not in relation to the owner of raw data. By extracting synthetic data from original data, VODAN-A gains benefits from this process. Firstly, synthesizing process implement data protection mechanisms from the technical perspective since accessing and processing synthetic data does not violate privacy. Secondly, data subjects are no longer traceable through synthetic data. Synthetic data does not fall under the scope of GDPR, and it can be reused and interoperated easier.

However, metadata may contain information which can indirectly trace identification of data subjects. Therefore, attention should be paid within VODAN-A to avoid statistical distributions revealing any personal identifiers, both direct and indirect. In addition, synthesizing process may affect the data quality, because outputs of this process are sampled on the basis of the distribution, other than the whole population (Plug, R., et al., 2022). As a result, the reuse of synthetic data may lead to slight errors or difference while performing the simulation among the whole population in epidemiological research. This concern should be taken into account as VODAN-A aims to sustain the high level of data quality especially those scholarly data while preserving the data ownership.

5.3 Data Processing Agreement (DPA)

Apart from technical mechanisms of protecting privacy, VODAN-A also stipulates terms such as conditions, obligations, restrictions, and duties in its data accessing agreement as a high-level guideline to facilitate the legal use of data in compliance with regulatory framework (Purnama Jati, P. H., et al., 2022). For instance, leveraging this contract, VODAN-A ensures that the purpose of processing data and the way of processing are explicitly defined and constrained at a high-level alignment with the GDPR. A complete DPA signed by stakeholders of VODAN-A can be seen in the Appendix B.

High-level policies have significant impact on data control at the country level, every data service associating with sensitive data must follow instructions of DPA in terms of the appropriate processing techniques. However, DPAs have not involved concrete rules indicating data accessibility of a specific group of users; if we want to achieve a finer-grained governance of data, we need to introduce an access control policy at the repository level so that a group of users could be associated with the specific sections in the repository when the data is created using templates.

5.4 Data sensitivity hierarchy and analysis

The sensitivity level or security level was defined early in the RAND Corporation report of U.S. Department of Defense, in which data resources were categorized into different levels such as confidential, secret, top secret and etc (Ware, 1979). Similarly, we classify data based on their sensitivity within VODAN-A, the following Figure 10 illustrates that highly sensitive data has dramatically low accessibility while less sensitive data possesses relatively high accessibility. Patient’s medical diagnosis is regarded as the highest sensitive data due to the close connection with patient’s privacy, which thus deserves the securest protection; in contrast, aggregated statistics are the lowest sensitive data with the highest accessibility, which can be processed and shared without the permission of their owners.

5.4.1 Clinical healthcare data

In VODAN-A, the data production pipeline starts initially from the collection of patient’s healthcare data. These sorts of data are located above the dash line of Figure 10, such as patient registration records and clinical diagonosis which contain biometrics which can directly or indirectly track identities and disease history which might associate with patient’s privacy, therefore they are identified as the highest senstitive data and are reposited internally with the encryption protection approach (Van Reisen, 2022). These sort of data have limited visibility, only data owners and delegated data processors can visit them. After anonymization and pseudomization process, healthcare data whose sensitive parts are removed can be used for research purpose without the restriction of GDPR.

Healthcare data is described by metadata like patient’s name, blood pressure, heart rate, and other medical indicators. Some of them are terminology with domain-specific knowledge, in this case, metadata is not sensitive by nature. The remaining parts of them are linked to personal biometrics of patients, like the test results of COVID and blood pressure, in this regard, metadata is sensitive. All metadata and relevant controlled vocabularies construct the ontology of healthcare data. This ontology may have interrelations with the other ontologies, such as research data's, patient's, doctor's, researcher's, repository's, and etcetera. (see the E-R diagram in Figure 9)

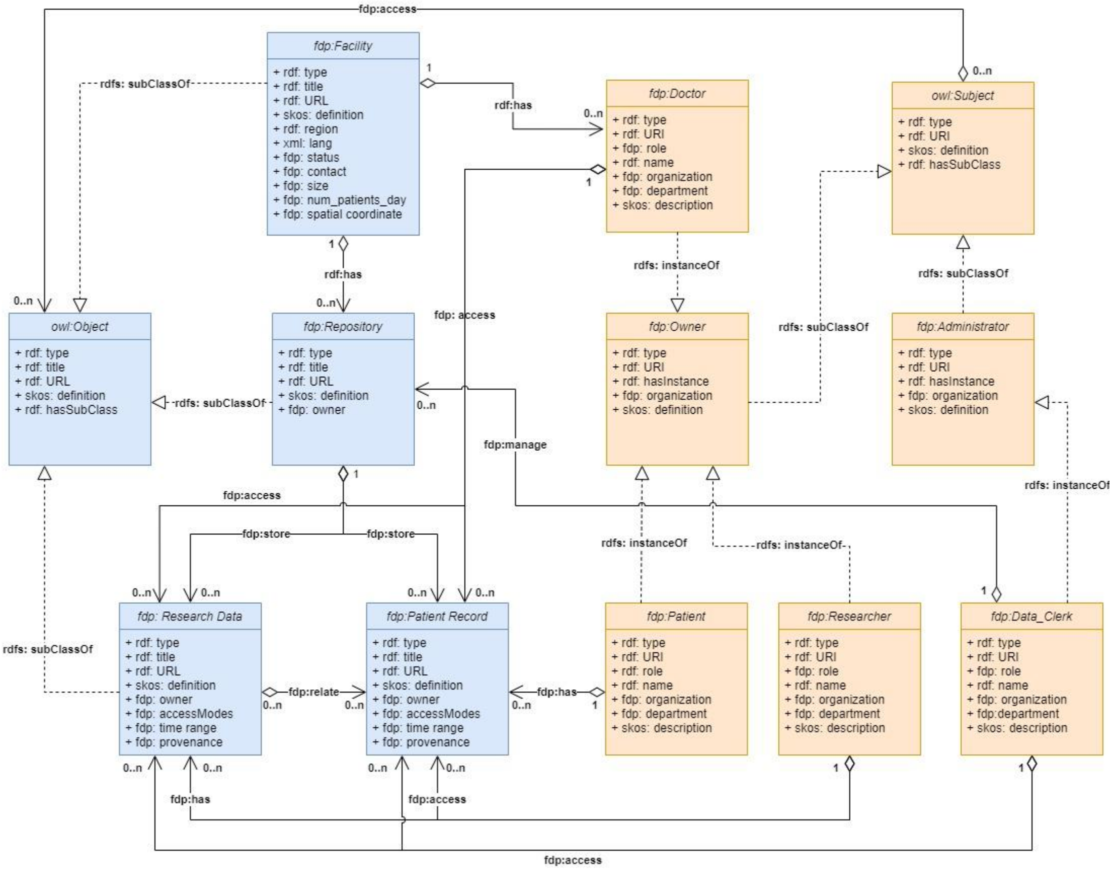


Figure 9: The partial entity-relation(E-R) graph with underlying metadata

5.4.2 Research data

Research data in scholarly publications is categorized as the non-sensitive data which removes sensitive identifiers through synthesizing process as discussed in the previous section, it thus has relatively high accessibility. These kinds of data are located under the dash line of the Figure 10, namely anonymous data, synthetic data, and aggregated results (i.e. statistical distribution of the whole population). In addition, data sensitivity also depends on the choice of owners. For example, given the fact that this research data is disclosed in publications, we can argue that owners are willing to make it open, which means that the permission to these aggregated data is approved by their owners by default.

Research data is described by metadata such as the context of underlying experiments, the purpose, methodologies of reproducing experiments, and the interpretation of results. The ontology of research data is formed by combining these metadata, which have connections with other ontologies, such as localized repository's, researcher's, patient record's, and etcetera. (see the E-R diagram in Figure 9)

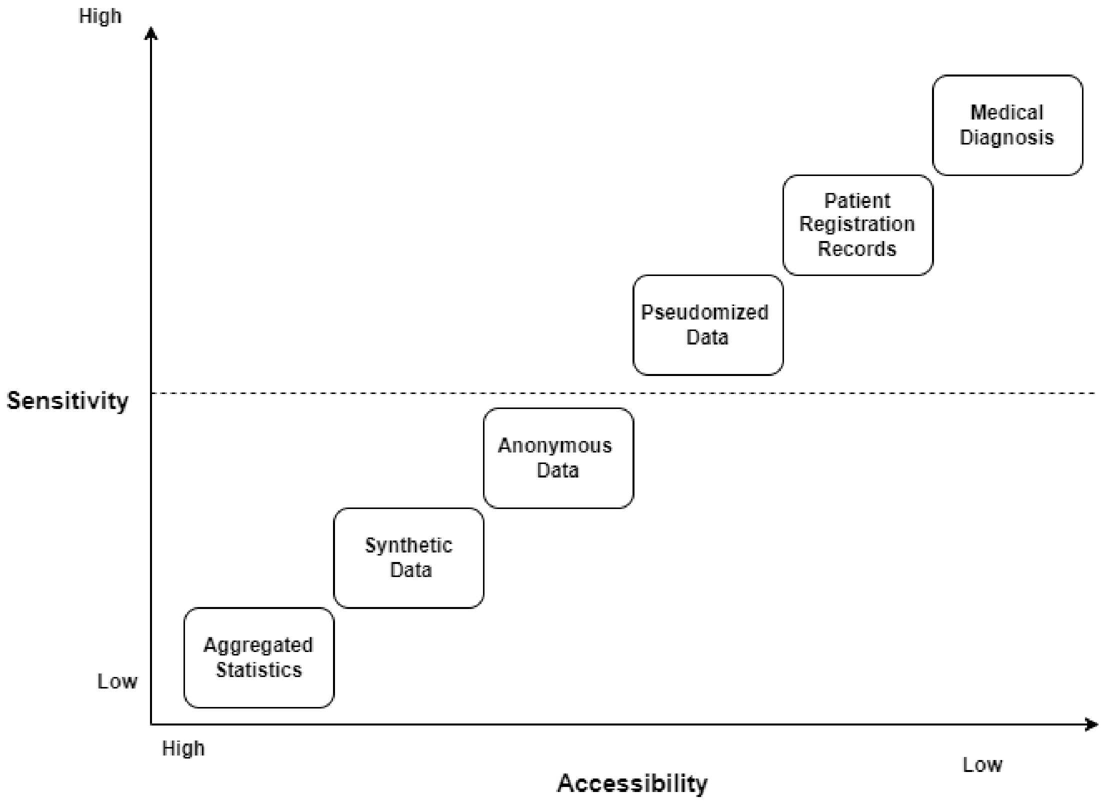


Figure 10: Data sensitivity and accessibility level within VODAN-A. Source: (Van Reisen, 2022)

5.5 Conclusion

In this chapter, investigation firstly dives into access control mechanisms implemented in functional components of VODAN-A-MVP architecture. We conclude that a role-based access control is being utilized in the internal dashboard to restrict the visibility of different web pages. Then, other data protection strategies are also identified, including data pseudonymisation, anonymization, and synthesizing. These approaches aim to remove partial or complete sensitive personal identities from original data so that the use of data could comply with regulatory frameworks such as GDPR. Finally, the data taxonomy is analyzed. We conclude that VODAN-A team has established a comprehensive scheme to classify data on the basis of sensitivity and accessibility ranging from clinical records with greater secrecy (therefore) relatively low accessibility to aggregated data with the least privacy but the highest accessibility. All approaches mentioned above sustain the data sovereignty in VODAN-A to some extent. However, access control mechanisms over bio-semantic metadata has not been identified in the current architecture so far, leading to the exploration of feasible solutions for governing the access to these metadata in this research.

6. Access Control Model and Mechanism

In this chapter, we start with the elaboration of conventional access control models from which important characteristics and concepts are still derived to form the foundation of modern access control mechanisms. Then, RBAC and ABAC mechanism are further demonstrated in terms of their features, use scenarios, mathematical expressions, and limitations which facilitate the exploration of access control mechanism which can be applied to the ontological metadata in the section 6.5.

6.1 Bell-LaPadula security model

Before the overall definition of Bell-LaPadula model is delivered, a conceptual framework describing the secure state of an access control system and several significant principals are introduced in section 6.1.1 and section 6.1.2 respectively as the foundation of this model.

6.1.1 Definition of an access control system

The initial version of security control model was proposed by David Elliot Bell and Leonard J. LaPadula. Designers attempted to model the access control system by the total sequences of relation among system state, inputs, and decisions, which is denoted by $\sum W\,(\mathrm{R}, \mathrm{D}, \mathrm{z})$, where triple $\mathrm{W}(\mathrm{R}, \mathrm{D}, \mathrm{z})$ is one single relation, R is the request from users, D is the decision of each request, and z is the system state.

The core concept is "security level" in level function. Permissions of accessing data resources are determined by comparing the security level of both users and data resources. Security level is a classification scheme applied to both users and data objects, it can be expressed in a pair which consists of clearance level of users (i.e. confidential, secret, top secret) and the category set of data resources or an organization they belong to (i.e. healthcare, LUMC), which is denoted by: security level ← (clearance + category set). According to their definition,

system state(z) = current access set(b) + access matrix(M) + level function(f) + hierarchy(H).

- **current access set** is a triple: $(subject, object, access\ mode)$, it defines access operations allowed to perform on the object, including execute, read, append, and write mode.
- **hierarchy** is a tree-like structure specifying the inheritance pattern of data objects (Bell, D. E., & La Padula, L. J., 1976) in order to achieve the control at different levels (i.e. a file repository has a root-parent-child structure, it thus requires access control rules to be defined at multiple levels).
- **access matrix** is a two-dimensional table with subject individuals as rows and object individuals as columns. It records and visualizes the access permissions which stipulate subjects can perform which access modes on object (see Figure 11).
- **level function** is a relation-mapping operation assigning the security level to subjects and objects. Level function $(f_S(S_i), f_O(O_j), f_C(S_i))$ is denoted f individually or F collectively, whereof $f_S(S_i)$ represents the (maximal) security level of subjects, $f_O(O_j)$ represents the security level of objects, $f_C(S_i)$ is the **current** security level.

In summary, the definition of an access control system containing all secure states is concisely denoted by:

$$\sum (R, D, (b, M, f, H))$$

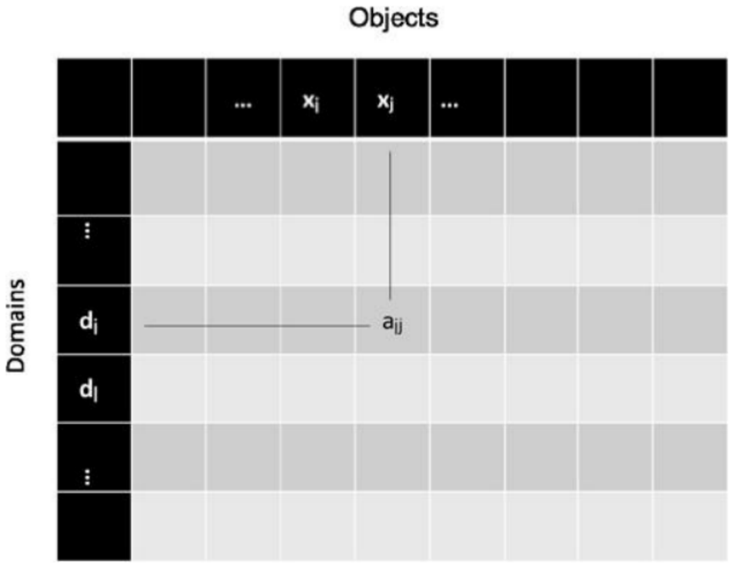


Figure 11: Access Control Matrix. Source: (Vincent C. Hu, David F. Ferraiolo, Ramaswamy Chandramouli, D. Richard Kuhn, 2017)

6.1.2 Security principal

(1) simple security property (ss-property) principal

Researchers at that time widely agreed that “access” means not only physical ownership of objects, but also the capability of extracting information from objects (Bell, D. E., & La Padula, L. J., 1976) by more operations (i.e. append, write). The ss-property principal stipulates that:

- if access mode is **read** in the current access set b(subject, object, access mode), then $f_S(S_i) > f_O(O_j)$.

However, ss-property with only one mode of manipulations does not reflect the capability of extracting information despite the fact that it played a significant role in controlling the “observation” of physical documents in the early stages.

(2) *-property principal

Alternatively, the *-property principal extended access modes from the single read mode of ss-property to three modes including append, write, and read. In any system state (z), a current access set b(subject, object, access modes) stipulates that:

- $f_O(O_j) > f_C(S_i)$ if access mode is **append**;
- $f_O\left(O_j\right) = f_C(S_i)$ if access mode is **write**;
- $f_O(O_j) < f_C(O_j)$ if access mode is **read**.

More concisely, the logic of the determination of access modes is that subjects can simultaneously read data object-1 and modify data object-2, if security level of object-1 and object-2 is higher than or equal to the **current** security level of subject. In addition, **current** security level of any objects should dominate the security level of objects.

(3) discretionary security property (ds-property) principal

Enforcement of both clearance-based and category-based standards is mandatory to all user subjects or data objects, namely that it does not allow any entities to define rules based on their discretions (Bell, D. E., & La Padula, L. J., 1976). Both ss-property and *-property principals thus fall into this so-called Mandatory Access Control (MAC). In contrast, Discretionary Access Control (DAC) permits an authorized user to grant the permission to data resources to other users who comply with MAC policies by customizing rules in a DAC policy.

The ds-property principal is satisfied if the access mode: x is recorded at the intersection of s_i row and o_j column in the access matrix M_{ij}, where s_i, o_j, x are defined in the current access set b (s_i, o_j, x). It is also worthwhile to note that access modes (x) are not necessarily required to assign to all entries of subject and object in an access matrix(M).

In summary, by combining ss-property and *-property principal, they are also termed "write up, read down" principal according to this up-to-down pattern as depicted in Figure 12. That is, users can read data if data is assigned with "upper" or equal security level than users', whilst users can modify data if data has lower security level than users.

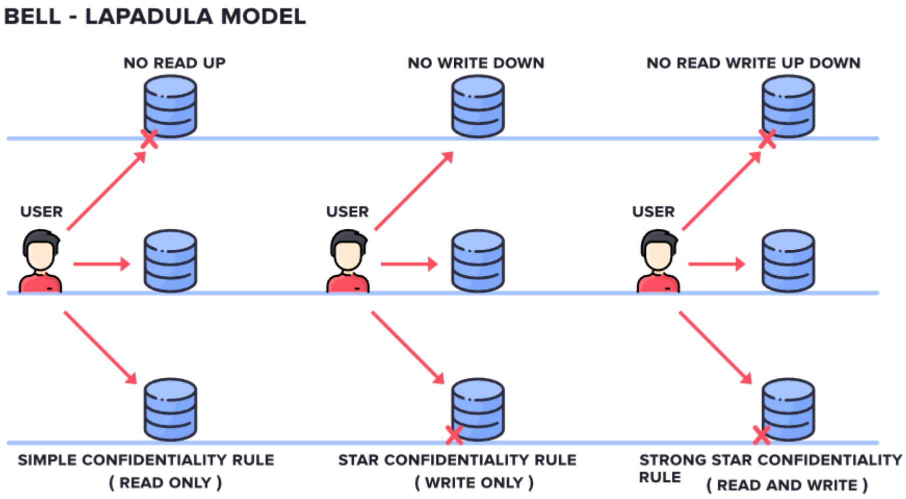


Figure 12: "write up, read down" principal in Bell-LaPadula Model. Source: https://www.geeksforgeeks.org/introduction-to-classic-security-models/

6.1.3 Bell-LaPadula model

Based on the prerequisites mentioned above, Bell-LaPadula model is represented by the following theorem:

Let ρ be a principal and satisfy $\rho(\mathrm{R}, \mathrm{v}) = (D, v^*)$, where $\mathrm{v} = (\mathrm{b}, \mathrm{M}, \mathrm{f}, \mathrm{H})$ and $v^* = (b^*, M^*, f^*, H^*)$.

- ρ is simple security property principal, if $b^* \subseteq b \ and \ f^* = f$;
- ρ is *-property principal, if $b^* \subseteq b \ and \ f^* = f$;
- ρ is discretionary security property principal, if $b^* \subseteq b \ and \ M_{ij}^* = M_{ij}$

6.2 Biba integrity model & Clark-Wilson integrity model

Traditional security models focused on the secrecy capability of U.S. Department of Defense. Clark and Wilson clarified distinctions between business and military security requirements in 1987 (Clark, D. D., Wilson, D. R., 1987), they concluded that the **integrity** issue should be urgently solved within the commercial sector. Follow-up models such as Biba model (Biba, 1977), Clark-Wilson model, and Chinese Wall model (Brewer, D. F., Nash, M. J., 1989) were proposed. Compared with prior Bell-LaPadula model in which permissions are granted based on security level, these models used **integrity level** to determine whether or not a subject (i.e.

user, application) can perform actions (i.e. read/write/append) on an object (i.e. file, relational database table) in a computer system. The integrity refers to any performances conducted on data, which may cause the change in the state of data. Integrity levels reflect the validity and consistency of data in its whole lifecycle (Boritz, 2005). Integrity control policies (which is equivalent to access control policies) are defined on the basis of integrity levels for the determination of permissible integrity or alteration accesses (Biba, 1977).

Given the definite set of subjects S, objects O, integrity levels I, we can define the following notions:

- $i\colon \mathrm{S} \cup \mathrm{O} \ \rightarrow I$. This operation assigns the integrity level to subjects and objects;
- $r\colon \mathrm{S} \times \mathrm{O}$. This means the subject set: $\mathrm{s} \in \mathrm{S}$ can read the object set $\mathrm{o} \in \mathrm{O}$;
- $w\colon \mathrm{S} \times \mathrm{O}$. This means the subject set: $\mathrm{s} \in \mathrm{S}$ can write data to the object set $\mathrm{o} \in \mathrm{O}$;
- $x\colon \mathrm{S} \times \mathrm{O}$. This means the subject set: $\mathrm{s} \in \mathrm{S}$ can execute actions on the object set $\mathrm{o} \in \mathrm{O}$.

6.2.1 Biba integrity model

Like the "write up, read down" principal of Bell-LaPadula model, Biba integrity model determines access permissions according to integrity level defined above, this model can be represented using the following theorem:

- $\exists s \in S$, s can **read** $\forall o \in O$ only if $i(s) \leq i(o)$
- $\exists s \in S$, s can **write** $\forall o \in O$ only if $i(o) \leq i(s)$
- $\exists s_1, s_2 \in S$, $s_1$can **execute** s_2 only if $i(s_2) \leq i(s_1)$

It is important to note that the integrity level is different from the security level which is used in Bell-LaPadula security model. As revealed by the first principle, subjects are allowed to read resources only if the integrity level of subjects is lower than or equal to that of objects. In contrast, subjects are permitted to read resources and write data to resources only if the security level of subjects is higher than or equal to that of objects in a security level based model. In a word, these sorts of mechanisms require a user's clearance to dominate data resources' classification in order to obtain the accessing permission (Vincent C. Hu, et al., 2017).

6.2.2 Clark-Wilson integrity model

Compared with Bell-LaPadula security model and Biba's integrity control addressing requirements of military system, the one for commercial information system emphasized the importance of fraud and error prevention (Clark, D. D., Wilson, D. R., 1987). For example, users of accounting system who lack familiarity with business process or financial knowledge may be granted the right to change data in such a manner that error of balance and other accounting items exist, even though their access is authorized. The military's security control model which aims to manage simple operations such as read and write is not sufficient to handle dynamic requirements in the commercial world. Therefore, a more constrained mechanism which takes business process into consideration is required to sustain high data integrity level for the sake of fraud and error control.

Clark and Wilson thus extended integrity actions from initial three access modes (read/write/execute) to a more flexible one called transactions, which are defined by a set of constraints (Clark, D. D., Wilson, D. R., 1987). A transaction procedure (TP) consisting of read, write, and execute actions with certain sequence is granted after integrity constraints are satisfied. They categorized data into constrained data item (CDI) and unconstrained data item (UDI) based on data's compliance with integrity control policies. In this model, unconstrained data can be accessed by subjects without any restrictions, whilst constrained data cannot be directly accessed by subjects, these data can be accessed only if valid transformation process and integration verification process are performed on them.

In this model, other important principles enhancing the data integrity level include well-informed transactions and separation of duty (Clark, D. D., Wilson, D. R., 1987). The former controls how data is manipulated by users in order to guarantee that a valid state of data would be consistent, whilst the latter requires users with various functions to mutually constrain operations performed on data in a business process in order to decrease the possibility of fraud.

6.3 Role-based access control

The history of role-based access control (RBAC) dates back to the 1990s, with the inheritance relationship with previous integrity models, it was proposed to deal with security requirements existed outside the defense domain, such as private and public sectors. According to (David F. Ferraiolo, D. Richard Kuhn, 1992), RBAC does not simply just constrain the read and write operations over data resources, it instead imposes control over a transaction flow which comprises these fundamental operations with certain sequence. Data transaction can also be regarded as the dataset binding with transformation procedures, which change the data from an available state to another one according to requirements of an actual business process (Clark, D. D., Wilson, D. R., 1987).

RBAC grants data accessing decisions on the basis of functions a user is allowed to perform within an institute (David F, et al., 1992), therefore it belongs to non-discretionary mechanism in which access constraints are applied to all users within the organization. Basic RBAC model has the following permission mapping pattern:

$$U \to R \to Perm$$

where *U, R, Perm* represents the user set, role set, and permission set respectively. This pattern reveals that user's identifications determine roles it has, whilst roles decide access permissions. Like Clark-Wilson's integrity model, the core concept of RBAC is separation of duty which creates mapping relations between users and data objects. These relations can be defined as one-to-many or many-to-many. This characteristic solves the delegation chain problem which exists in prior traditional models.

One of advantages of RBAC is its capability to centrally manage and maintain data accessing policies without data owners specifying the permission to visit the data individually as achieved by discretionary access control lists. ICT administrator in an organization is responsible for distributing approved operations performed on datasets within certain contexts to roles of a group of users. As to the use case within a hospital, for example, actions for a doctor involve reading patients' records, adding, or altering medical treatment diagnosis while operations for a pharmacist include writing the prescription to a record of patients. It is important to note that operations here are not solely read or write a record, it is instead a workflow combining read and write actions with a certain sequence derived from a practical business/activity process.

However, the central management of policies, as the advantage of basic RBAC model, makes it cumbersome in a distributed context today. Large organizations and their subsidiaries face the challenge of role exploration resulting in the difficulty of defining complex RBAC policies because different branches or even functional units within the same branch vary in the role architecture, the number of roles is very likely to exceed that of actual employees in the complicated distributed environment. The massive personnel mobility within these organizations would also lead to frequent modifications of mapping relations between users and roles, which poses a threat to the RBAC policy maintenance.

Another deficiency of RBAC is its inadequate capability to integrate dynamic environment characteristics such as a fixed time frame for accessing the data. The sole reliance on the use of role may violate principals and does not leave any solutions for emergent situations in the healthcare domain. For instance, if a patient suffers from serious disease, doctors on duty (whereas they are not delegated to this patient) should own the permission to visit this patient's diagnosis history within a period (Khan, M. F. F., Sakamura, K., 2012). Both limitations mentioned above can be resolved by ABAC as discussed in the section 6.4.

6.4 Attribute Based Access Control

Attribute based access control (ABAC) associate attributes of both data resources and users with permissions. Basic ABAC mechanism has the following mapping pattern:

$$A_1, \dots, A_n \rightarrow Perm$$

where A_i is the set of attributes. In addition to this classical model, there are several derivatives of original ABAC. Dynamic ABAC model combining the feature of RBAC has the following permission mapping pattern:

$$U, A_1, \dots, A_n \rightarrow R \rightarrow Perm$$

In contrast, the role is solely treated as one of attributes assigned to users in the attribute-centric model without the role as the medium to determine permissions:

$$U, R, A_1, \dots, A_n \rightarrow Perm$$

A practical implementation of this pattern is proposed by (Kuhn, D. R., Coyne, E. J., & Weil, T. R., 2010) in which attributes of either subjects or data resources to RBAC.

Attributes on data resources can be assigned by either its owner when created or automatic approach such as populating (meta)data template with controlled vocabularies. Attributes on users are defined by the personnel management system within organizations. Vincent C. Hu used the term – **security states** to list all possible permission entries between users and data resources which each user is able to access at a specific period. The number of security states is equal to $U \times R$, where U is the size of users while R stands for the size of rules.

Generally speaking, there are two approaches explicitly expressing ABAC policies: logical-formula and enumeration (Biswas, P., et al., 2016).

Logical-formula technique utilizes relational operators (e.g. $\neq, \leq, >, \cap, \vee$) on attribute values of numerical (meta)data and binary operators (e.g. AND, OR, NOT, ANY, ALL) on those of textual (meta)data. This technique can flexibly define any elements of ABAC policies. For instance, an OBAC policy can be defined as the following tuple:

$$\mathrm{policy} \leftarrow (\mathrm{sa}, \mathrm{rule}, \mathrm{oa})\ (\mathrm{Shen}, \mathrm{H.B.}, 2006)$$

where sa is the definite set of subject's attributes, $rule \leftarrow (date \geq 01/05/2022 \cap date \leq 01/08/2022 \cap role = "practitioner" \cap institute = \text{"LUMC"} \cap division = "respiratory\ " \cap country = "Netherlands" \cap action = \text{"write"OR "read"})$ is a series of constraints that subject's attributes need to satisfy, and $\mathrm{oa} \leftarrow (\mathrm{ID} = 12345, \mathrm{address} = 192.168.0.1, \mathrm{location} = \text{"LUMC"})$ is the set of properties describing the resources that subjects desire to access. The role above specifies that subjects who work as a practitioner in respiratory division of LUMC in the Netherlands can write new records to a data object from 01 May to 01 August.

Another common method is to enumerate or to configure the relationships between entities (Vincent C. Hu, 2017). In Next Generation Access Control (NGAC) model, as a standard implementation of enumeration-based expression, four types of relations are defined: *assignment*, *association*, *prohibition*, and *obligation*.

Like the policy triple in the logical-formula model, similar notation is defined to represent the *assignment* and *association* relations in enumerated approach: $ua - ar - oa$, where ua is the set of user's attributes, ar is the set of access right on objects (i.e., read, write, append), oa is the set of object's attributes. All users complying with the ua set can perform manipulations in the ar set on data objects. Figure 13 visualizes the enumeration graph where solid arrows and dash lines denote the *assignment* relation and *association* relation respectively. As depicted in this graph, it enumerates all possible access privileges of users who are contained in the set ua.

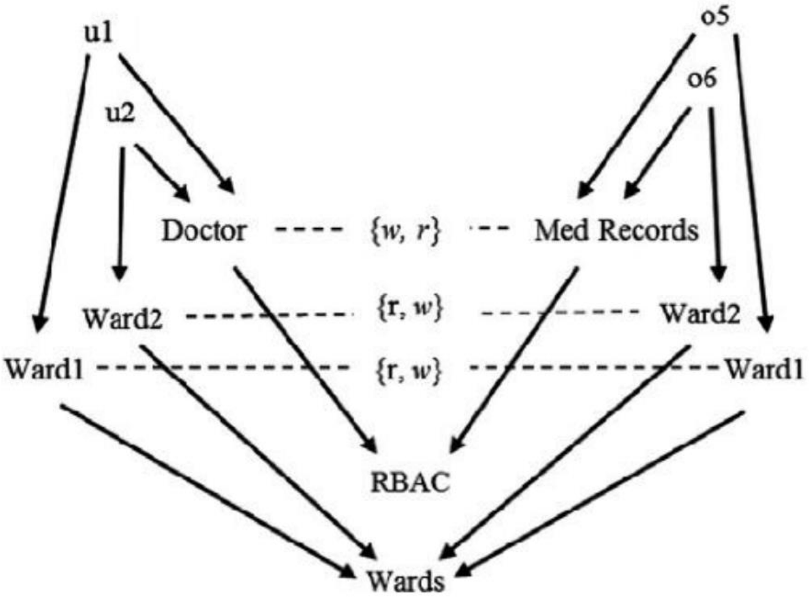


Figure 13: Enumeration graph. Source: (Vincent C. Hu, David F. Ferraiolo, Ramaswamy Chandramouli, D. Richard Kuhn, 2017)

Both approaches have similar expression capability. Even though the number of association relations defined in enumerated ABAC model is exceed the one of logical-formula approach, they both require the same amount of operations which assign value to specific attributes.

After policy logics are defined, the next step is to convert these human-readable policies into machine-actionable ones by policy description languages. Existing policy specification languages include (W3C, 2009): Open Digital Rights Language (ODRL) by ODRL Initiative, A P3P Preference Exchange Language (APPEL) by W3C, Autonomic Computing Policy Language (ACPL) by IBM, Web Services Policy Framework by W3C. To be specific, XACML is chosen in this research to generate ontological access control policies with more detail revealed in the next section.

On one hand, policymaking should satisfy various authorization requirements of applications in a distributed environment. On the other hand, it is essential to use a commonly adopted language for the consolidated perspective of access control policies to facilitate the communication of access control information across various organizations (Vincent C. Hu, 2017). Compared with other policy languages, the eXtensible Access Control Markup Language (XACML) standard is a mature framework designed by OASIS for the guidance of ABAC implementation and policy representation (OASIS, 2013). Its syntax using XML and

JSON schema makes it be capable of supporting the basic ABAC model as well as its different derivatives, such as OBAC.

One of the limitations of ABAC is its significant time spent on computation of permission sets accessible to all users (Vincent C. Hu, et al., 2017). ABAC allows a limited part of resources to be accessed by users according to the least privilege rule, this feature requires a regularly daily evaluation, it is thus challenging for ABAC to achieve the real-time permission authorization.

The key to reduce the calculation cost is the decrease in security states by limiting either the size of U or R. First, we can reduce the number of R by merging data resources into groups or hierarchies in accordance to the taxonomy. Bio-semantic ontologies defined in VODAN-A project can be used to categorize data. In this way, an individual rule can apply to all associated data entities which fall under a particular group (Vincent C. Hu, et al., 2017). But in some special contexts, data resources cannot be combined because they do not have any hierarchical relationship. Second, it is also significant to decrease the size of U. The most common approach is to define users' privilege hierarchy based on roles a user is allowed to play within an organization, as what RBAC functions. The joint use of features from ABAC and RBAC effectively reduces the computation cost. For example, if 10,000 users are assigned to 100 roles, the size of security states ($U \times R$) declines by a factor of 100. Another strategy is to use geographical characteristics to separate user groups, this is efficient especially in the distributed system such as VODAN-A. Data processors can be distinguished based on countries as criteria when assigned the permission to access data. If more precise control is required, data processors can also be classified at province/region/community geographical level. This approach not only satisfies the requirement of localization in which data ownership is controlled by locales, but also solves the computational challenges of ABAC.

However, the more classification of both resources and users a system has, the less ability of fine-grained access control it owns. Therefore, we should make the trade-off between the computation efficiency and the granularity level of access control.

6.5 Access control over ontology

On one hand, as explained in previous sections, a feasible option for reducing the computation cost is to classify either user entities or data entities by the use of ontology, which is efficient especially in a distributed environment. On the other hand, accessing to ontological metadata avoid direct manipulation of actual data, thus decreases the possibility of violating data protection regulation especially involving sensitive personal identities.

The joint use of ontologies and existing access control mechanisms is motivated by the convenience of modelling an AC policy in the open circumstance in which the volume of data is huge. Initially, as pointed out by (Heath, T., & Bizer, C., 2011), linked metadata published in Linked Open Data Cloud did not include any metadata which explicitly stipulate access constraints and define the scope of data accessibility. To address this issue, (Liu, et al., 2016) proposed a fined-grained context-aware access control to model the ontologies including subject, object, operation, and context, which belong to components of a standard ABAC policy. In addition, (Khan, M. F. F., Sakamura, K., 2012) used Web Ontology Language (OWL) to represent the role's and permission's ontologies of RBAC under the architecture of eTRON. These permissions are defined in the Health Level Seven technical standard (HL7, 2011) in the healthcare domain, in order to comply with the Health Insurance Portability and Accountability Act (HIPAA) of U.S. and similar regulatory frameworks of other countries.

Another purpose is to prevent (sensitive) raw data being accessed by only allowing the permission to semantic metadata. Permission to metadata only is widely used in the defense domain to protect extremely sensitive documents, as pointed out by one of the interviewees in this research. This requirement can be satisfied by the design of appropriate data accessing level. For example, (Brewster, 2020) proposed an accessing hierarchy in which metadata is located in the upper position, whilst raw data is situated in the bottom level (see Figure 14). In this way, data access decisions are made by ontological metadata without the involvement of actual raw data; only if all conditions are fulfilled, users can ultimately visit the actual data at the bottom level. It is also worthwhile to note that this descending access pattern is similar to the "read down, write up" principal from Bell-LaPadula model: users can view a resource only if their current secure level is lower than that one of resource. They proved their concept by applying this model to simulate the process of issuing a warrant for drug dealers in the police station. As depicted in Figure 14, narcotic police officers with certain attributes (i.e. role, nationality, rank, investigation case) can initially access a limited amount of objects' metadata (i.e. category, name of drug), they are able to access further details of a drug transaction (raw data) after a broader permission is granted based on the status of investigation they are responsible for. Similar implementation can be seen in the work of (Kiran, G. M., & Nalini, N., 2020), they proposed a ontology graph specifying various operations can be performed on medical metadata by users with roles in corresponding levels.

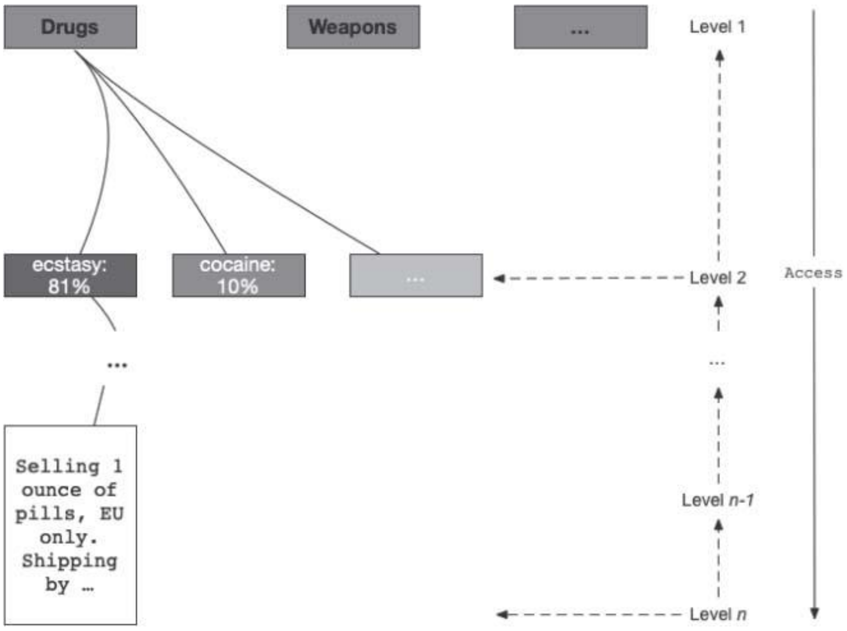


Figure 14: metadata accessing hierarchy. Source: (Brewster, 2020)

The implementation detail of OBAC tailored to VODAN-A will be further elaborated in the chapter 8, including the design of ontological structure, policy expression, mathematical foundations, as well as the overarching architecture of the system.

6.6 Conclusion

In this chapter, access control mechanisms and models are introduced in detail especially RBAC and ABAC, since most access control models at present fall into these two categories. Conventional secure models are also elaborated because they are foundation of modern access control mechanisms. For instance, security level based models (i.e. Bell-LaPadula model) were primarily applied in the military domain to constrain the read, append, and write operations on secret files, whilst integrity level based mechanisms were widely adopted in the commercial environments such as banking system, accounting system, and medical application for fraud and error prevention in accordance to relevant business process or logic within domains.

Important concepts and features derived from traditional integrity models are also adapted to the requirements of modern information systems. For example, the creation and modification of both clinical and scientific research data in VODAN-A project are equivalent to the concept of data integrity originated from early integrity models. Moreover, data collection approach in VODAN-A involves commonly approval FAIRification process in which raw data is transformed into machine-actionable FAIR data using standard templates. This is consistent with the Certification Rule No.5 of Clark-Wilson Model: Any transaction procedures that take UDI as the input perform valid transformation steps on all possible values of UDI to obtain CDI (Clark, D. D., Wilson, D. R., 1987). Data is persistently reposited in local facilities with robust access control mechanisms which prevent any unauthorized data manipulations and ensure that data is consistent across facilities. As a consequence, integrity level in terms of validity and consistency within VODAN-A sustains at a high standard.

As to the selection of access control mechanisms, there is no one access control being perfect in all systems, it thus requires system architects to analyze requirements of specific domains and make trade-off among existing control models. Similarly, single access control mechanism cannot fit into multiple components within VODAN-A architecture, these components vary in terms of both the data usage and functionality. The hybrid use of access control models can satisfy both security and functional requirements for the community. In VODAN-A, semi-structured data triples encoded using RDF are stored in the Allegrograph repository, this requires a special language – SPARQL which can recognize the triple patterns to query data. Given this fact, conventional access control models cannot satisfy such a requirement because these models are only applicable to structured data in a relational database system. Moreover, on one hand, simple protection mechanisms such as allowing users to define the accessibility (public/private) of these linked ontologies limit the availability of open data even though they enhance the security level of the system. On the other hand, some public triple stores in each FDP endpoint may still indirectly link to the sensitive information without being recognized by data owner. In this case, creating access control policies over these triple stores by default contributes to achieve a robust data protection scheme in VODAN-A. However, some concepts such as role distinctions, properties of entities from these classic models and mechanisms are critical to the design of access control policies. Therefore, we argue that ontology-based access control together with these concepts is a feasible mechanism for controlling the data triples in the Allegrograph.

7. Research design – Requirement analysis

In-depth interviews were performed mainly with VODAN-A stakeholders to validate our assumptions regarding the lack of an access control mechanism and to figure out their prospects related to the research design of our proposed ontology-based access control. OBAC system refers to an access control mechanism which poses constraints on ontologies in a open data network. In general, one ontology consists of the main entity (also known as raw data) it represents and corresponding metadata. In our design, raw data is hidden behind the metadata without disclosure to data requestors before a permission decision is grant. Currently, technical team of VODAN-A is responsible for designing and maintaining biomedical ontologies on BioPortal (Basajja, M., 2022). However, ontologies of access control policies together with relevant ontologies pertaining to role distinctions, accessing obligations, and availability scopes have not been created within VODAN-A. Therefore, the design of these policy-related ontologies is conducted in this chapter.

As shown in the Table 5, seven interviewees from five participant groups were selected in our interviews with consideration of participants' diverse expertise. For example, firstly, Interviewee A, as the specialist with expertise of privacy preservation, can offer insights about access control in both VODAN-A and biomedical domain. Secondly, Interviewee B, one of team leads in VODAN-A, is able to share comprehensive knowledge about the security scheme within VODAN-A and data classification in relation to sensitivity. Thirdly, Interviewee C, one of engineers in the team, can offer data visiting architecture and current access control mechanisms of VODAN-A in detail. Fourthly, Interviewee D who works as the data steward and data scientist in Kenya can offer feedback regarding privacy/security concerns from the perspective of the VODAN system user. Finally, Interviewee E is selected as the participant because of his expertise of privacy protection in relation to the high-level policy-making process. In addition, two interviews with Dutch government officers were performed to confirm that data ownership and security challenge is a global issue, rather than an issue in developing countries.

Table 5: The overview of interview participants

Participant	**Role**	**Group**
Interviewee A	Researcher in LUMC, Member of VODAN-A	Relevant stakeholders
Interviewee B	Data scientist in the Netherlands, Member of VODAN-A	Technical team
Interviewee C	Member of VODAN-A	Technical team
Interviewee D	Data scientist in Kenya, Member of VODAN-A	Data steward/processor
Interviewee E	Privacy expert, Member of VODAN-A	Security/Privacy officer
Interviewee F	Data governance officer in Dutch government	Digitalization conference
Interviewee G	Digitalization officer in Dutch government	Digitalization conference

Table 6 lists all relevant facets and corresponding answers extracted from the transcript of each interviewee. Labels such as sensitive data, non-sensitive data, and protection approach are designed to identify the data classification as well as corresponding protection strategies within VODAN-A and to figure out participants' prospect of access control to be implemented in the future. Moreover, the data governance level and secure score are utilized to understand the necessity of data stewardship at different levels and to diagnose the secure status of current architecture. Finally, we aim to identify factors related to the elements of proposed OBAC policy through labels including data operation, prerequisites, and AC policy elements. Several remarkable findings regarding the requirements raised by interviewees are summarized and discussed in detail in the following sections.

Table 6: The overview of interview results

Interviewee	**Sensitive data**	**Non-sensitive data**	**Governance level**	**Protection approach**	**Data operation**	**Secure score**	**Prerequisites to process data**	**AC policy**	**AC policy elements**
A	Biometrics	Isolated metadata	Multiple levels	Encryption, RBAC	Download, copy	8 (design)	Legal and ethical clearance	Defined by users	Role
B	Structured knowledge	Isolated controlled vocabulary, open data	Multiple levels	Agreement, ABAC, anonymization	Download, view	7	User expertise, background (no data offensive record)	Depend on facilities	Time range, user group
C	Outpatient data	Aggregated statistics	Service (API)	Aggregation, anonymization	Integrity processing	8 (Implement)	Token-based authentication	Depend on facilities	Role-based privileges
D	Biometrics	-	Data owner	RBAC, VPN	Extract, view, modify	5	Signed contract	Group user	User group, location (proxy)
E	-	Basic info of facility	Facility	Localized storage, RBAC, PIA	View	7(design& implement)	Identity validation process	Cultural & legal practice	Role, Users' identity
DPA	Personal Data	-	Country	Legal compliance	-	-	Have it singed	Comply with local regulations	Time range

7.1 Requirement 1: well-documented data sensitivity specification

Participants share different opinions towards the sensitivity of semantic metadata or CEDAR templates. Interviewee A commented that he did **not** think metadata is sensitive because metadata in essence is the description of raw data. This is partially true in specific cases, especially when we refer to isolated metadata, as pointed out by Interviewee B. In this case, controlled vocabularies, as the metadata of CEDAR-based data input templates, are the common knowledge widely adopted within domains, which are not sensitive by nature. For instance, it could be the classification of disease in the medical terminology; it could also be the biometric samples such as the blood pressure. The sensitivity also depends on owner's prospects pertaining to the way how the metadata is utilized. For example, structured metadata in a publicly funded scientific project is not sensitive since government patrons want it to be open and shared.

Meanwhile, interviewee B also argued that semantic metadata may result in privacy concerns depending on various contexts. In his opinion, domain-specific knowledge, as a part of intellectual properties of institutions, is formed when metadata is combined together, which requires the investment of resources (i.e. time and money), it thus associates with secrecy concerns. For example, researchers do not expect other parties to directly know what they are actually researching. In this case, if unauthorized users retrieve knowledge from knowledge graph (or semantic web) with intentions, their behaviors are regarded as the theft of intellectual properties of an organization. The leakage of incomplete research data may cause misinterpretation of research results, which may indirectly damage the reputation of organizations. In contrast, if we randomly query knowledge without any purposes, it is less likely for us to derive sensitive information from the graph.

The reason why they hold different points of view lies in aspects or internal layers of VODAN-A they involved. Interviewee A is a privacy expert who has expertise in the generalized access control and the regulatory framework so he does not identify sensitive parts of metadata from a macro perspective, whilst Interviewee B is a member of technical team who dives into more details of the issue so he can recognize the potential concerns regarding the sensitivity of metadata from a micro perspective.

Therefore, the specification of data sensitivity hierarchy should be well-documented to enhance the transparency among VODAN-A community and make sure that everyone understands the context.

7.2 Requirement 2: consistent goal towards the high security standard

Furthermore, it is also worthwhile to mention that perceptions of VODAN-A's secure state vary between internal stakeholders and external beneficiaries. Internal engineers who are responsible for the design and the implementation of VODAN-A consistently think the secure level of current architecture regarding the privacy preservation is high (scores are higher than 7), whilst users of VODAN-A express their concerns towards the secure state (score that Interviewee D ranks is only 5).

For example, from the perspectives of internal engineers, Interviewee B and C who are responsible for the implementation from scratch showed positive attitudes towards the secure state of the overarching architecture. They pointed out several data protection approaches existing within the current architecture, such as (1) encryption which protects the data communication, (2) anonymization which protect users' privacy by removing sensitive identities, (3) VPN which controls data accessibility based on locations, last but not least, (4) data processing agreement as the legal approach. From the perspectives of privacy specialists, Interviewee E commented that the privacy by design in the localized architecture with limited permissions to data is safe, which allows the data to be stored in residence and be aggregated internally. Similarly, Interview A as an important stakeholder argued that the ethic and security specifications are well documented in compliance with GDPR in the design phase.

However, Interviewee E still thought privacy impact assessment (PIA) is required to achieve quicker response to privacy risks in the next phase of VODAN-A despite his positive attitude towards the privacy architecture. Through this assessment, people's mindset of security, potential risks, responsive processes are identified and further documented in order to convince stakeholders how secure and mature VODAN-A system is. Apart from that, the concern from Interviewee D also indicates that VODAN-A needs to engage more its beneficiaries or users in collaboration during the designing process in the upcoming phases. Moreover, most participants also expressed that data accessibility management can be improved by the role configuration at different governance levels in the future.

To conclude, a technical access control mechanism at multiple levels and the human involvement with the understanding the organizational process towards the high-level security are required in order to enhance the trust among existing institutes as well as attract more worldwide partners, such as the governments, research institutes, and private companies in the industry. This requirement can be realized by the mentorship programs.

7.3 Requirement 3: more granular access control mechanism

As analyzed in the former chapter, we assume that VODAN-A is currently using authorization (i.e. assigning users to specific groups) and authentication (i.e. logging in) processes to restrict the visibility of aggregated statistics in the internal dashboard without the authentic access control mechanism. This hypothesis is proved by three participants: Interviewee E who works as the privacy specialist in VODAN-A, Interviewee C who works as an engineer within VODAN-A, and Interviewee D who plays a role as the data scientist in healthcare domain as well as the user of VODAN-A dashboard. Specifically, Interviewee C pointed out that statistics displayed on the internal dashboard vary between clinics, this localization feature allows each locale determines the scope of the data accessibility at the facility level. Moreover, he added that data owners can authorize permissions of an individual user or a group of users to access metadata templates in the CEDAR system, so the structure of internal templates which are identified to be sensitive data relating to intellectual properties as we discuss in the previous sections are protected. However, there is no more granular unit of accessibility, such as, at the 'data' or 'repository' level, which means that everyone within the facility is able to view all the information on the dashboard after logging in. In this case, the whole dashboard is regarded as an individual resource which can be accessed by everyone, but there should be more accurate distinction between data to be visited within this dashboard. Meanwhile, Interviewee D commented that there is no authentic access control implemented in VODAN-A and other information systems in the medical sector even though Virtual Private Network (VPN) is used to restrict data resources within organizations to be accessed by external parties, and thus that possible approach could be assigning users to different groups according to factors such as roles, geographic boundaries.

Therefore, we argue that this management approach is still not robust since there is more precise distinction between roles and other properties within the facility. For example, entities such as doctors, data clerks, and board members have slightly different privileges in accordance to their responsibilities. Furthermore, the control over more data manipulation modes (i.e. modify, copy, download, delete) should be implemented in the current system. The combination use of them, defined as the transformation process based on the domain-specific requirements, should also be considered. For example, 'FAIRification' transformation process can be separated into individual operations, including loading a CEDAR template, loading controlled vocabularies, adding metadata, saving data in the repository. Last but not least, both roles and manipulation modes of each locale even deserve our attention when we aggregate data across facilities or countries, this requires access control policies to be defined at different levels because data owners do not want to share anything by default To conclude, a finer-grained access control requires all context factors to be stipulated unambiguously in the access control policies in both human and machine readable format.

7.4 Requirement 4: a standardized template of access control policy

All participants pointed out that access control policies should align with cultural practices, administrative procedures, and legal requirements of local facilities. For example, as pointed out by Interviewee E, data owners cannot view the data they published until relevant publications are ready in Nigeria. This special process requires elements such as published time, users' role, status of publication to be considered when we design access control policies for Nigeria.

In addition, participants mentioned several important facets that can be included in an access control policy to define factors (i.e. roles, the time and locations) or accessibility modes (i.e. view, download, copy, modify). However, few participants are aware of perceptions of the data access control policy or assert that the approach to implement it. Therefore – in this research, a standardized template of access control policy will be conceptualized (see the following Section 8.3) based on the facets that participants mention, and elements extracted from the data processing agreement. The aim is to deliver a foundation of the access control policy initially, then to inspire VODAN-A members to extend and sustain the controlled vocabularies in this template hereafter.

Another reason why we need a standardized access control policy is to construct the ontology of all relevant policies, which is the basis of automated access control in the future. Leveraging the reasoning capability of semantic web, it is feasible to track the changes in the metadata which might lead to any updates on the security level of data, and automatically modify the terms in access control policies.

8. Research design - Ontology Based Access Control

Data is produced and maintained at its source location. Data templates vary between facility to facility. The stewardship of heterogeneous data requires the access control over SPARQL queries to be dynamic and flexible on the basis of the data sensitivity level. In this chapter, we propose an Ontology-Based Access Control architecture which is integrated into the current VODAN-A system.

8.1 Comparison of OBAC Implementation Approaches

In addition to VODAN-A project, similar scientific data opening initiatives such as Personal Health Train (PHT), smashHit project, and Social Linked Data (SOLID) project (Inrupt Group, 2022) also face the access control challenge of ontological data. (Beyan, O., et al., 2020) argued that XACML framework is one of ideal solutions for managing the access to ontologies in PHT. In smashHit project, XACML, as the external tool, which is embedded into their existing security mechanism, is also used to define policy based on the consideration of the flexibility to represent and modify policies in a machine-readable fashion (Chhetri, T. R., et al., 2022). Differently, in SOLID project, they implement the access control over linked ontologies by Open Digital Right Language (ODRL).

8.1.1 Open Digital Right Language (ODRL)

ODRL is a descriptive standard to define permission control policies of digital resources (Steyskal, S., et al., 2014). It was applied in the SOLID project as the approach to protect linked ontologies in this project. This language specifies facets in a digital policy, such as (1) Action – the manipulation allowed to perform on data; (2)Assets – data to be protected; (3) Constraint – a series of limitations when accessing data; (4) Duty – obligations required to fulfil before accessing data; (5) Party – data owners or data processors; (6) Agreement – an overview of all rules defined in the policy. However, (Pandit, H. J., et al., 2022) argued that ODRL is not an appropriate method to track any changes associated with the local clinics and states of ontologies specified in a policy. In contrast, XACML contains a timestamp in its policy, therefore it can perform well when detecting the changes of the policy.

8.1.2 Consent & Data Management Model (CDMM)

CDMM is one of the policy specification languages. However, there is only one property - "permission" in CDMM to describe the states of policies (Pandit, H. J., et al., 2022). According to GDPR, the purpose and the approaches of processing data should be explicitly defined in a consent, in both human and machine-readable format. Therefore, CDPM lacks the ability to model the policy which is align with GDPR. In contrast, we can extend ontologies of policies by adding metadata called "purpose" or "processing" and specify these updates in XACML policies for the sake of GDPR compliance.

8.1.3 Privacy Ontologies for Legal Reasoning (PrOnto)

PrOnto is the ontology designed to represent the knowledge of GDPR, including privacy agents, data classification, data manipulation types, responsibilities, and rights. Corresponding components are thus defined, namely 'Data', 'Purpose', 'Processing', 'Agent', and 'Right'. The PrOnto has clear sensitivity hierarchy of data (see Figure 15) as what we define within VODAN-A (see Section 5.4 and Figure 10).

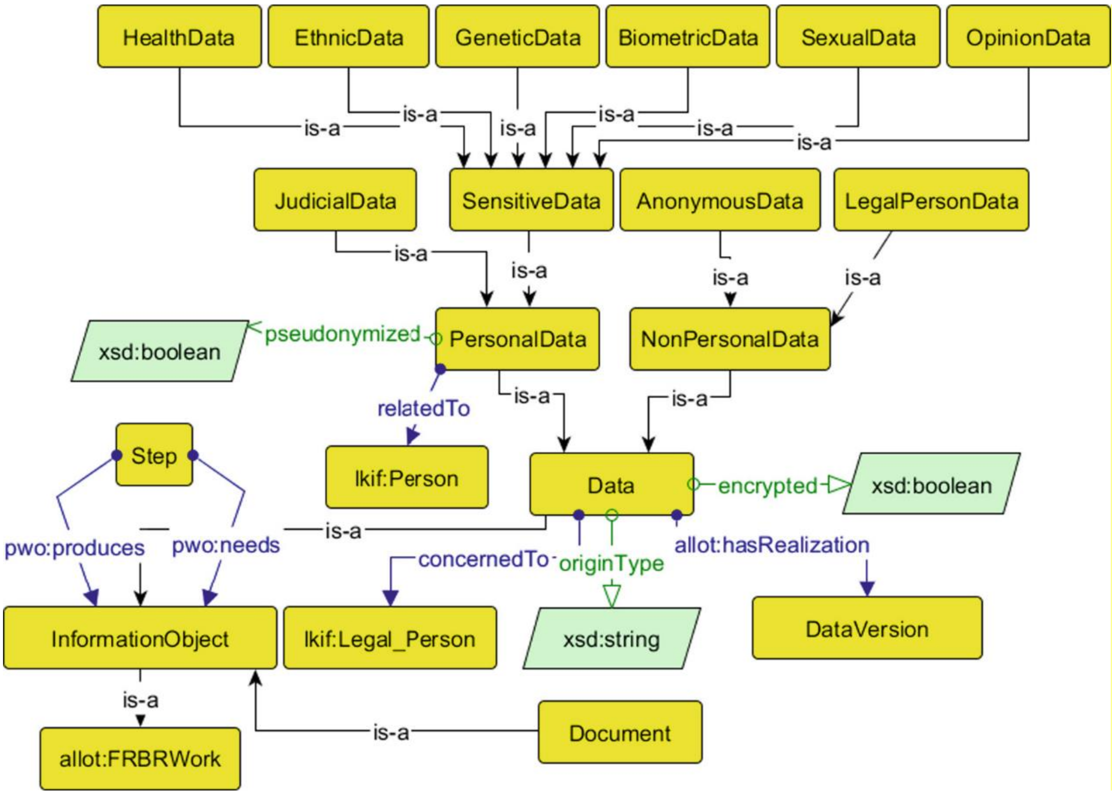


Figure 15: Ontologies of PrOnto revealing the classification of data. Source: (Pandit, H. J., et al., 2022)

Furthermore, PrOnto distinguishes operations allowed to perform on data (see Figure 16). In contrast, we do not create a specific ontology to model data actions in our work; instead, we clearly define data operations (i.e., access) as the relationship (see solid lines in Figure 9) between subjects and objects in our ontology graph. However, on one hand, the lack of sufficient documentations cannot satisfy our requirement of high-standard specifications as discussed in Section7.2. On the other hand, the inefficiency of references to other ontologies or metadata disobey the Interoperability principle of FAIR. In contrast, XACML maintained by OASIS Consortium is a widely adopted standard with sufficient documentations and rich linkage to other ontology library such as Wikidata. Therefore, we do not select PrOnto to implement OBAC within VODAN-A.

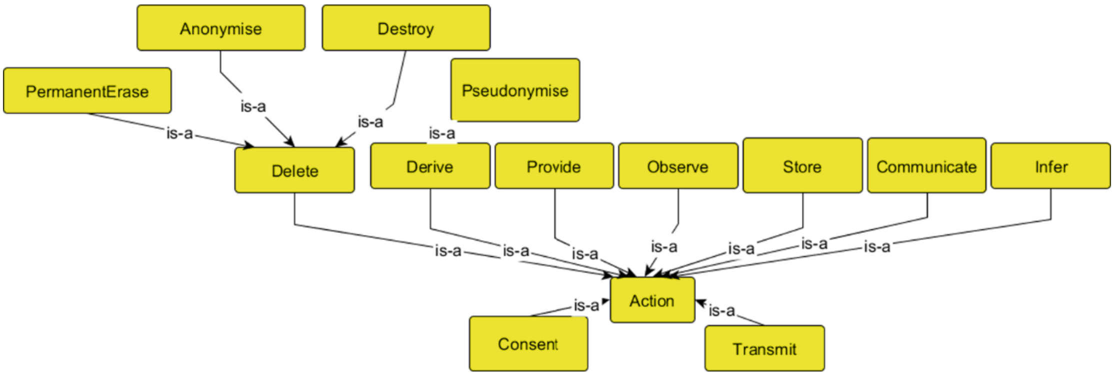


Figure 16: Ontology of data actions defined in PrOnto. Source: (Pandit, H. J., et al., 2022)

To conclude, all frameworks mentioned above are feasible approaches to implement ontology-based access control in a flexible way. However, based on diverse considerations such as the compliance with GDPR, the ability of detecting changes, and the alignment with requirements of VODAN-A and FAIR principle, these approaches cannot appropriately fit into the design of

OBAC in VODAN-A compared with XACML. Moreover, both Personal Health Train and smashHit which have similar localization structure as VODAN-A leverage XACML to implement their access control mechanism over ontologies. Therefore, we argue that XACML is a good fit for VODAN-A.

8.2 OBAC Architecture

The design of OBAC follows the eXtensible Access Control Markup Language (XACML) standard, so terminology in this design aligns with terms defined in official XACML specification. XACML is a standardized way to express Attributed-based access control policy in XML and JSON schema. The OBAC architecture integrates existing VODAN-A MVP with additional functional modules proposed in accordance with the findings of this research. The components of this architecture consist of the VODAN-A MVP, Allegrograph platform and Policy Validation Engine (see Figure 17).

Policy Validation Engine is designed to evaluate the SPARQL queries against machine-readable policy files with the XACML syntax. In this engine, all relevant information including (meta)data, requests, and policies are collected as the input and processed before authorization decisions as the output will be made. Policy Validation Engine is built upon AuthzForce, an open-source framework following the XACML standard, which provides a series of APIs to implement functions such as policy management, request-policy validation, and decision authorization. The specification of terms is elaborated below:

- Policy Enforcement Point (PEP) is the component where XACML request is automatically generated based on the data request and authorization decision is enforced.
- Policy Administration Point (PAP) is the component in which data accessing policies and policy sets are created and stored. It exposes an entry to external policy editor, which allows privacy-related clerks to compose new accessing policies.
- Policy Decision Point (PDP) is the component where authorization decisions are issued based on data accessing policies and XACML request from PEP.
- Policy Information Point (PIP) is the component where metadata of data resources, data requestors, and the context are stored.

Compared with original architecture of XACML data-flow model, several components such as context handler and obligations service are removed for the ease of integration with Allegrograph platform. It is also worthwhile to note that an interface for context attributes allows the engine to import other environment-related attributes which are not included in the ontological metadata. As a result, the scalability of this engine is enhanced. As depicted in the Figure 17, following steps are operated in this policy validation workflow:

- [1] PIP derives ontological metadata from Allegrograph Repository. In this step, attributes (metadata) of data requestors and target data resources are also converted into XACML files that can be executed by PDP for evaluation procedure.
- [2] Data requestors send SPARQL requests to PEP through Allegrograph's Query portal.
- [3a] SPARQL requests will be sent to PDP after being converted to XACML request files wrapped with XACML syntax.
- [3b] Metadata files with XACML syntax will be delivered to PDP.
- [3c] Executable XACML policy files are retrieved from PAP to PDP so that data authorization decisions could be issued based on the multilateral information (i.e. data

requests, data's metadata, requestors' metadata, accessing policies).

- [4] PDP returns authorization decisions to the PEP.
- [5] If decisions are positive, the PEP will grant data requestors the permission to data resources; otherwise, it denies the access.

VODAN-A MVP converts non-FAIRified data into RDF-format data resources in compliance with FAIR principle and stores them in Allegrograph (Van Reisen, et al., 2022). Within this component, first of all, CEDAR templates with controlled vocabularies will be used to collect original data such as outpatient data generated from local facilities in a standardized manner. Afterwards, imported data will be further transformed into either triples stored in Allegrograph repositories or structured data stored in relational database respectively depending on the categories of data. After these transforming processes, FAIRified data is made and is available for the exposure and further processing. Even though RBAC is currently utilized in the internal dashboard, data transforming is still taken into consideration for consistency of access control workflow in the future. Moreover, heterogeneous datasets are transformed into triple stores for simplification of the design in this research despite the fact that OBAC does not constrain data formats under its management.

Allegrograph Platform functions as both data repository which exposes (meta)data to Policy Validation Engine and data query interface which allows data requestors to send SPARQL queries to Policy Validation Engine to obtain accessing permissions to specific datasets. The copy of data together with its metadata will firstly be delivered to PIP component for temporary storage in a secure environment, then will be sent back to Allegrograph query portal through PEP once access is approved. If the access is denied, this copy will be released from PIP. In addition, data communication between Allegrograph and Policy Validation Engine is fully encrypted. In this case, data safety is assured: (meta)data remains at the source and cannot be altered.

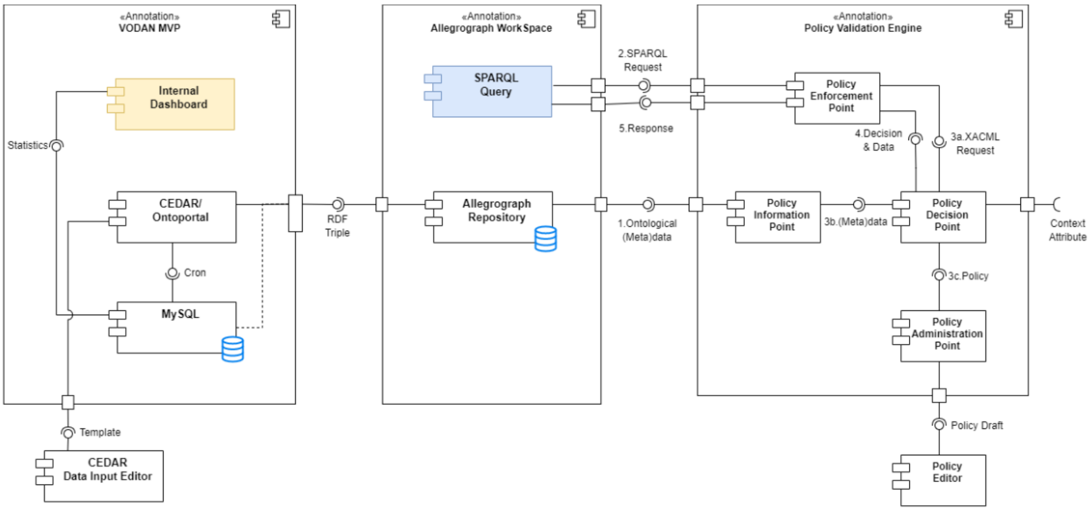


Figure 17: OBAC Architecture in relation to the VODAN-A architecture. The design of Policy Validation Engine takes XACML specification as references. This figure does not reflect the whole architecture of both VODAN-A MVP and Allegrograph platform, only components related to this research are displayed.

As illustrated by the Figure 18, data access control policies are created manually based on the requirements of facilities and converted to machine readable format with XACML notations, which are stored at the PAP. According to our design, privacy specialists are responsible for creating the general access control policies, whilst data clerks in local facilities are supposed to maintain those customized policies. However, according to the findings of interviews, there have not been qualified experts in each facility; this issue is supposed to be solved by appropriate training programs in the future. At the same time, data resources with metadata are transformed to RDF and deposited in the Allegrograph triple store. When users send a SPARQL triple request, elements including the SPARQL query, XACML policies, and metadata of resources will be put together to PDP for the evaluation. If the decision is approved, users will receive a Http response with URI specifying the location of data resources that they can access in the repository; otherwise, users need to fulfill other obligations defined in a policy and their new request will repeatedly go through the steps mentioned above until the permission is approved.

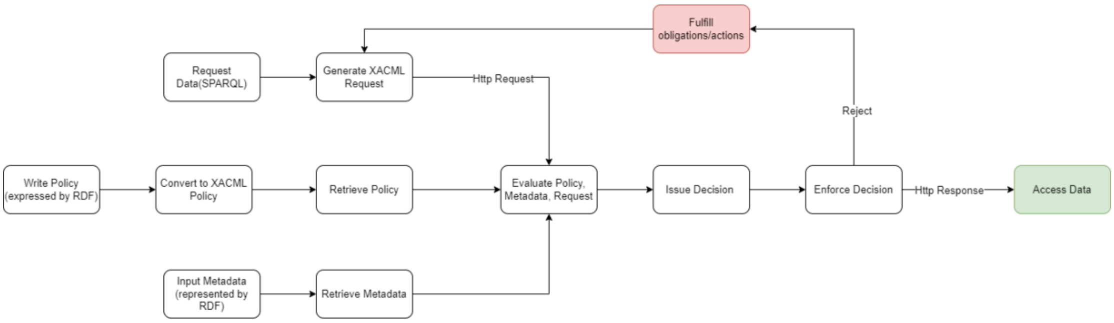


Figure 18: The process of accessing data under the control of OBAC

8.3 The design of access control policy

After understanding functions of the OBAC architecture and the interactions between components, the next consideration is the way to design OBAC policies. Data access control policy is defined as a series of constraints and rules which offer the approach to protect the data when they are applied (Vincent C. Hu, et al., 2017). Data to be preserved in this research refers to ontology triples stored in Allegrograph repository. According to VODAN-A stakeholders, data requestor’s common responsibilities to fulfill and conditions are stipulated in data processing agreement. Therefore, corresponding rules and terms will be written in data accessing control policies with either natural language or machine-resolvable format.

Possible conditions can be the envisaged time range for which the data will be stored (Article 15 (1d), GDPR). These common conditional requirements are generally provided by domain experts within VODAN-A community, who play a role as the data controller to define the purpose or processing approaches of data. Data controller can not only be single entity, but also an ad hoc committee consisting of multiple institutes with medical professionals. They work jointly to define those conditional requirements in common and the scope of data visibility, which comply with GDPR and local regulations.

In addition, the role, as one of attributes of data processors, should also be taken into consideration when defining the scope of data visibility. This is because the scope varies from role to role. (Priebe, T., et al., 2006) proved that different access control policies can exist simultaneously. TCSEC also requires that data resources are regulated by joint access control policies if a system is evaluated as Division B1 or higher trusty level (TCSEC, 1983). Certificated public service such as public security, defense, and jurisdiction requires full permission for accessing broader range of data. According to GDPR, this sort of service appoints municipal authorities with prioritized role and allows them to process data under strict regulations stipulated by local legislative body without acquiring informed consent (Article 9(3), GDPR). For example, VODAN-A MVP exposes its data to DHIS2 through which ministries of health can access these internal data at the government level (Lin, 2021), therefore, a data accessing policy with full data accessibility for officially certificated data processors is produced and stored in policy repository of each locale. In this regard, facilities can custom data access control policy based on any metadata including the role of data processor. In the Policy Validation Engine module, one external Policy Editor component is introduced as an entry point for the sake of policy composition, which allows privacy officers to create rules in access control policies.

The following five examples integrate requirements extracted from interviews with VODAN-A stakeholders. As discussed in the previous chapter, most stakeholders mention that the user's role, location are important factors to configure the permissions to data resources and that possible data manipulations are read, write actions. Therefore - in these five rules, aforementioned roles, locations, and actions are defined. For example, in Rule 1, users who are identified as the researcher (role) are allowed to read (action) patient records in some facility (location).

Rule 1 - Researcher can read patient records which is related to researcher's research data and is stored in facility's FDP.

Ontology: Researcher's, patient record's, research data's

Action: read

Request Pattern:

select ?subject ?predicate ?patient_record

Conditions:

?subject rdfs:instanceof fdp:researcher

?predicate rdf:type rdf:read

?patient_record fdp:relate_to fdp:research_data

fdp:repository fdp:store ?patient_record

Rule 2 - Doctors can read and modify records of their patients.

Ontology: Doctor's, patient's, patient record's

Action: read, write

Request Pattern:

select ?subject ?predicate ?patient_record

Conditions:

?subject rdfs:instanceOf fdp:doctor

?predicate rdf:type rdf: read

?predicate rdf:type rdf:write

?subject fdp:treat fdp:patient

fdp: patient fdp:has ?patient_record

Rule 3 - Doctors can read the research data which is related to their patients' records.

Ontology: Doctor's, patient's, research data's

Action: read

Request Pattern:

select ?subject ?predicate ?research_data

Conditions:

?subject rdfs:instanceOf fdp:doctor

?subject fdp:treat fdp:patient

fdp:patient fdp:relate ?research_data

Rule 4 - Owner can perform all actions on their data.

Ontology: Owner's, object's

Action: read, write

Request Pattern:

select ?subject ?predicate ?object

Conditions:

?subject rdfs:subClassOf fdp:owner

?predicate rdf:type rdf:read

?predicate rdf:type rdf:write

Rule5 - Data clerks can view research data and patient records which are stored in the FDP repository they manage.

Ontology: Data clerk's, researcher data's,

Action: read

Request Pattern:

select ?subject ?action ?object

Conditions:

?subject rdfs:instanceOf fdp:data_clerk

?predicate rdf:type rdf:read

fdp: repository fdp:store ?object

The next step would be translating the data accessing agreement into machine-readable policy files with XACML syntax, which will be stored in the policy library upon approval.

8.4 Data Access Control Policy with XACML syntax

The representation of data access control policies across African geographies should be compatible with local regulations. The XACML language allows, on one hand, commonly agreed-upon access control policies to be inherited by each facility across nations; on the other hand, data controllers in various geographies can define customized policies in compliance with local laws. The XACML language supports administration of multiple domains which are separate from each other, data controllers are domain administrators in fact, because they are responsible for the policy management in specific domain and the maintenance of the policy repository in residence in compliance with local legal framework.

XACML language is an embodiment of RDF which includes XML's property-value entries. In this way, it offers a flexible way to represent metadata's values in access control policies, which include all related subjects, data resources, permissions, and external contexts. XACML syntax can be imbedded in normative RDF triples (subject, predicate, object) by creating a subset relation with RDF syntax, for example:

xacml:Resource, rdfs:subClassOf, rdfs:Resource

xacml:Subject, rdfs:subClassOf, rdfs:Resource

xacml:Action, rdfs:subClassOf, rdfs:Resource

xacml:Environment, rdfs:subClassof, rdfs:Resource

By combining this schema and all rules together, we obtain a data access control policy. A complete XACML policy is shown in the Appendix A. The advantage of this OBAC policy is its flexibility to express rules depending on requirements of local facilities at different levels. However, the customization tends to bring the redundancy problem. As shown in the Appendix A, there are hundreds of lines defined in a machine-readable format, which might increase the difficulty of maintenance.

8.5 Conclusion

In this chapter, the architecture of proposed OBAC is presented, including all associated components and information communication interfaces of each component. And XACML is selected to express OBAC policies because this standard not only theoretically addresses access control challenges in the localized environment where both data and function heterogeneity exist, but also it is technically supported by a wide range of open-source communities like OASIS, Authzforce and mainstream platform vendors such as Apache Jena Fuseki. Five roles with XACML notation are drafted based on the requirements of VODAN stakeholders.

9. Discussion

In this chapter, implications of the ontology-based access control by design regarding research questions are explained. Afterwards, the limitations of proposed OBAC model and the prospects of automated access control in the future are also discussed.

9.1 RQ1: What are data protection techniques implemented in all components respectively within VODAN architecture?

It would be better to classify data produced in VODAN-A before we discuss corresponding protection strategies. The data can be roughly divided into research data and patient records, they can be identified as sensitive data depending on different scenarios or owners' proposition pertaining to how they want their data to be dealt with.

Patient's record which contains biometric or disease history is classified as sensitive data since these information directly or indirectly links to personal privacy of patients. In contrast, depersonalized patient's record using the aggregation technique is exempted from the control of GDPR, thereby being non-sensitive data.

Research data can contain sensitive information when it involves collaborative knowledge in an organization since these intellectual properties are parts of organizations' assets. The illegal or intentional access to these data would cause the loss of profits for a company. On the contrary, research data in a publicly funded project is not sensitive because the owners are willing to make it open.

Another non-sensitive data that VODAN-A handles is isolated controlled vocabulary which defines the domain-specific terminology from external parties. These vocabularies are utilized to explicitly describe a concept or a knowledge in the healthcare domain.

In VODAN-A, role-based access control has been implemented to authorize the permission of logging in to the internal dashboard based on pre-defined user groups, which means that users falling under authorized groups can view all statistics after they successfully log in. Statistics shown on internal dashboard vary between facilities in accordance to actual requirements of locales, whilst statistics on external dashboard are consistent, which aggregate individual statistics from all internal dashboard. Through this data localization technique, data protection is achieved at facility level: each facility can customize the structure of their own dashboard and the details of internal dashboard are hidden. However, a robust access control mechanism should achieve more accurate distinction between permissions to different data categories rather than the performance of the current case.

Other data protection techniques such as anonymization, aggregation, and synthesizing remove sensitive identifiers from original data in compliance with GDPR, which allows to process non-sensitive data without restrictions after these depersonalization processes. With data transformation techniques applied to different data, data sensitivity hierarchy has been established, data can thus be precisely categorized into aggregated statistics, synthetic data, anonymous data, pseudomized data, patient records, and medical diagnosis records. Among all types of data identified, aggregated statistics contain the least sensitive information, thereby having the highest accessibility; in contrast, medical diagnosis records closely relate to patients' personal privacy and thus have the lowest accessibility. The permission to those highly sensitive resources should be constrained by more conditions specified in a data access control policy (as depicted in the Rule 2).

9.2 RQ2: What would be the feasible access control mechanisms for Allegrograph-based semantic data queries?

Ontology-based access control is regarded as the feasible mechanism for managing permissions to RDF triples in the Allegrograph triple store. First, generally speaking, the ontology is broadly used in medical domain to model the structure of concepts. Second, in VODAN-A, massive FAIR data described by rich metadata is linked together to form a semantic web. Retrieving ontologies in this web requires special querying language which can recognize the triple graph patterns (subject-predicate-object), that language is called SPARQL language. Third, using ontology to model and represent data access control policies is able to satisfy the requirement of data localization where data control approaches are flexibly customized by facilities. Fourth, OBAC prevent (sensitive) raw data being accessed by only allowing the permission to semantic metadata.

The old generation of access control models such as Bell-LaPadulla security model, Clark-Wilson integrity model, and role-based access control mechanism were widely applied to information systems of individual organizations in both military and commercial domains. However, on one hand, they only focus on single aspect to separate data accessibility levels, namely security level, integrity level, and role distinction respectively, without considering multiple properties of entities. On the other hand, these conventional models appeared prior to the World-Wide Web, which means that they are not capable to handle the cross-organization data access control. Moreover, these models are only applicable to relational database systems where data is structured. Therefore, these conventional models are not eligible for controlling semi-structured data triples in the semantic web.

However, ontology-based access control may lead to redundancy of permission rules and inconsistency. This is because ontology is formed depending on the cognition of human-beings and their knowledge within domains (Kiran, G. M., & Nalini, N., 2020). The possible evaluation approach can leverage execution time and the number of policies as the indicators to compare the performance of different access control mechanisms. But VODAN-A has not implemented any access control mechanism, so it is difficult to conduct evaluation experiments currently, which leaves the room for future discussion.

Based on the analysis above, we argue that OBAC is a good fit for the Allegrograph triple store to control SPARQL queries despite the deficiency of redundancy.

9.3 RQ3: What is the difference between stakeholders' opinions discovered in terms of the requirement of semantic data queries?

Stakeholders expressed different opinions towards the secure status of VODAN-A architecture, internal engineers and privacy specialist consistently thought current architecture is secure in the design and implementation phase. Nevertheless, external data stewards argued that this architecture is not safe due to the lack of a robust access control policy.

Regarding the structure of a data access control policy, stakeholders didn't have prospects of it when asked, but they mentioned various factors related to elements in a data access control policy. Elements can be prerequisites such as (1) identity validation, (2) user expertise validation, and (3) legal and ethical clearance. Elements can also be constraints such as the time range, role distinctions, and locations. By comparing points of view from stakeholders, we summarize the following key requirements that should be considered in our design.

Firstly, a finer-grained access control is required given the fact that current data protection approaches still remain at the facility level. Data resources reposited in the facility are heterogeneous based on sensitivity level; similarly, privileges of staff in the same facility vary depending on their functions. Therefore, we need to associate the sensitivity hierarchy of data to accessibility of users based on properties or metadata of both users and data modelled in ontologies, namely a finer-grained access control below the facility level.

Secondly, a standardized template of access control policies is required to help stakeholders understand the outline of access control policies. Elements in an access control policy could be: (1) prerequisites such as the match between user's role and data's category, (2) constraints such as the time range and location. Ideally, inspired by this preliminary template, stakeholders are able to expand and customize it based on their own requirements.

9.4 RQ4: How does the design of Ontology-Based Access Control (OBAC) over semantic data queries look like?

In our design, the overall OBAC architecture consists of Allegrograph Workspace, Validation Engine and existing VODAN-A MVP. To be specific, Validation Engine is further decoupled into sub-components: PEP, PIP, PDP, PAP, and Policy Editor. In addition to the design of OBAC architecture, another important design is the template of OBAC policy tailoring to requirements of VODAN-A stakeholders.

As discussed in the previous section, elements in OBAC policies are identified through interviews of VODAN-A stakeholders. Important factors such as roles, the time range, and locations are gathered from the requirements of interviewees, these factors are the basis of rules in an OBAC policy. The detailed rules with these factors are shown in Role 1 – 5 from section 8.3. Another way to explore elements in OBAC policy is the DPA. Several relevant factors such as the time range, role classification are identified. These factors from DPA overlap with ones of stakeholders' points of view.

Regarding the method of representation, OBAC policy is encoded using XACML syntax in a machine-readable format. XACML policy can be generated using open-source tools such as *WSO2 Identity Sever* and *Security Policy Tool.* These tools follow standard communication protocol in the industry, which align with the FAIR principle.

9.5 Limitation

As analyzed in the chapter 8, ontology-based access control may lead to redundancy of permission rules and inconsistency. This is because ontology is formed depending on the cognition of human-beings and their knowledge within domains (Kiran, G. M., & Nalini, N., 2020). Feasible approaches such as combined algorithms of XACML standards and conflict-detection algorithms. Due to the time constraint, these technical topics would not be covered in this research, which leaves the room for improvement of proposed OBAC model and has aforementioned problems solved in the future work. Another limitation of this study is the deficiency of evaluation mechanisms to examine the performance of OBAC over SPARQL queries, this part will also be considered in the future work.

9.6 Prospect

The trend from manual permission authorization towards the automated access control requires the assistance of Artificial Intelligence. However, due to the lack of standardized access control policies in the current phase of VODAN-A, it is challenging to achieve automated permission authorization at the current phase. Alternatively, manual creation of data access control policies is still required to collect sufficient samples. Therefore, New research gap has been identified: on one hand, mathematical model regarding knowledge reasoning problem in the semantic web has not been formed; on the other hand, the sample of access control policies is not sufficient to train the model. Therefore, new research direction in the future can be divided into two branches: first, to research knowledge-graph-based reasoning algorithm which enables identification of metadata's update and automatic modification of access control policies; second, to conceptualize a knowledge reasoning problem in order to obtain a standard machine learning model (mathematic theorems, the structure of neural network, a pipeline of model training, performance validation mechanism, etcetera).

10. Conclusion

In this research, nationwide data leakage in the healthcare domain firstly raises our attention to discover its potential reasons and feasible solutions. The biggest challenge the Netherlands faces is the lack of security/privacy experts who can identify the vulnerabilities and understand the consequences in both private and public sectors, which is also the main root resulting in the outflow of data. Several critical consequences underlying this problem are identified, including the loss of value of data in residence and the violation of citizens' privacy. In this context, global collaboration project – Virus Outbreak Data Network-Africa provides a feasible solution to tackle this challenge, its data localization concept allows data to be stored at its source securely without being accessed by unauthorized entities. This initiative effectively avoids the outflow of the valuable data collected in residence while facilitates the reuse of scientific data under the guidance of FAIR principal and local regulatory frameworks, which contributes to the immediate response to global pandemic and the improvement of local healthcare ecosystem. However, VODAN-A currently remains at the emerging phase, which still requires a multidimensional data protection scheme, especially a more robust access control mechanism. The exploration of an appropriate access control mechanism is the primary objective of this research.

Relevant research approaches are conducted to fully understand the current situation. Firstly, preliminary literature review was carried out to diagnose the problem of VODAN-A regarding its data protection mechanisms from a general scope. Several important VODAN-A related articles published in Data Intelligence Journal consistently demonstrates that the current emergence is to enhance data sovereignty by the appropriate strategies in different layers. Furthermore, we seek to understand the context broadly by exploring literature about security and privacy concerns in different domains, especially the bio-semantic field. Next, the investigation of data protection approaches was under way into all components of VODAN-A architecture. Even though some privacy protection techniques such as anonymization and aggregation are identified, an authentic access control which achieves data governance at a finer-grained level has not been implemented yet. Given this scenario, further literature review was carried out to select one eligible access control mechanism by scanning and comparing all relevant models proposed by other researchers in different data sharing initiatives. For example, conventional integrity models prior to the birth of world-wide web cannot satisfy the requirement of data governance in a distributed context; the policy expression performance of SBAC and RBAC is less than that of OBAC. Therefore, ontology-based access control (OBAC) is regarded as a good fit for VODAN-A due to its requirements of multilevel data governance. Moreover – among other implementation approaches, XACML is regarded as the most appropriate approach to implement OBAC within VODAN-A based on multiple considerations such as the distributed structure of VODAN-A, the compliance with GDPR and FAIR principle. Last but not least, in-depth interviews were conducted with stakeholders ranging from technical engineers to privacy specialists in order to figure out their prospects of an access control model. The contributions of this research are the OBAC architecture by design and the access control policy with important factors extracted from stakeholders' inputs.

In the trust capacity building phase of VODAN-A, security level can be enhanced by both administrative approaches like the privacy impact assessment (PIA) as the early preventive evaluation and technical mechanisms like the access control models at the moment when data is exposed to end users. From operational perspective, first, mentorship programs should be introduced to ensure that all members understand the organizational process and have consistent goal towards the high standard of both security and privacy preservation. Second, PIA should also be included to help data clerks build up the mindset of critical consequences of specific risks. Third, both ethical and secure specifications should be well-documented by a broadly understandable language or description to strengthen the trust among all participants so that new partners could be attracted as well. From legal perspective, current data processing agreement should be kept updated according to any changes in local regulations about the data protection. From technical perspective, an access control model based on the ontologies should be designed to assist the legal approach in terms of flexible alignment with data processing agreement and other local legal frameworks.

References

[1] Akindele, A. T., Tayo, A. O., Taye, G. T., Amare, S. Y., Van Reisen, M., Berhe, K. F., Gusite, B., Edozie, E. (2022). The impact of COVID-19 and FAIR data innovation on distance education in Africa. *Data Intelligence Vol. 4*.

[2] Amare, Y. S, Taye, G. T., Van Stam, G., Van Reisen, M. (2022). Conundrum of diversity: The FAIR experience. *Data Intelligence Vol. 4*.

[3] Basajja, M., Nambobi, M. (2022). Information streams in health facilities: The case of Uganda. *Data Intelligence Vol. 4*.

[4] Basajja, M., Nambobi, M., Wolstencroft, K. (2022). Possibility of enhancing digital health interoperability in Uganda through FAIR data. *Data Intelligence Vol. 4*.

[5] Bell, D. E. (2005). Looking back at the Bell-La Padula model. *21st Annual Computer Security Applications Conference (ACSAC'05)* (p. 15). IEEE.

[6] Bell, D. E., & La Padula, L. J. (1976). Secure computer system: Unified exposition and multics interpretation. *MITRE CORP BEDFORD MA*.

[7] Beyan, O., Choudhury, A., Van Soest, J., Kohlbacher, O., Zimmermann, L., Stenzhorn, H., Dekker, A. (2020). Distributed analytics on sensitive medical data: the personal health train. . *Data Intelligence, 2(1-2)*, 96-107.

[8] Biba, K. J. (1977). Integrity considerations for secure computer systems. *MITRE CORP BEDFORD MA*.

[9] Biswas, P., Sandhu, R., Krishnan, R. (2016). Label-based access control: An ABAC model with enumerated authorization policy. *ACM International Workshop on Attribute Based Access Control*, (pp. 1-12).

[10] Björklund, T. A. (2013). Initial mental representations of design problems: Differences between experts and novices. *Design Studies*, 34(2), 135-160.

[11] Bless, C., Dötlinger, L., Kaltschmid, M., Reiter, M., Kurteva, A., Roa-Valverde, A. J., & Fensel, A. (2021). Raising Awareness of Data Sharing Consent Through Knowledge Graph Visualisation. *Knowledge Graphs IOS Press.*, 44-57.

[12] Boritz, J. E. (2005). IS practitioners' views on core concepts of information integrity. *International Journal of Accounting Information Systems*, 260-279.

[13] Brewer, D. F., Nash, M. J. (1989). The Chinese Wall Security Policy. *IEEE symposium on security and privacy (Vol. 1989)*, 206.

[14] Brewster, C. N. (2020). Ontology-based access control for FAIR data. *Data Intelligence*.

[15] Candan, K. S., Liu, H., & Suvarna, R. (2001). Resource description framework: metadata and its applications. *Acm Sigkdd Explorations Newsletter, 3(1), 6-19.*

[16] Chhetri, T. R., Kurteva, A., DeLong, R. J., Hilscher, R., Korte, K., & Fensel, A. (2022). Data Protection by Design Tool for Automated GDPR Compliance Verification Based on Semantically Modeled Informed Consent. *Sensors, 22(7), 2763.*

[17] Chindoza, K., Van Stam,G., Mulingwa,A.,Mawere,G. E., Mukute, S., Winji,

L., Marima, I. J. (2022). eHealth implementation in Zimbabwe: An exploration into the usability of FAIR in data integration. *Data Intelligence Vol. 4*.

[18] Clark, D. D., Wilson, D. R. (1987). A comparison of commercial and military computer security policies. *IEEE Symposium on Security and Privacy* (pp. 184-184). IEEE.

[19] David F. Ferraiolo, D. Richard Kuhn. (1992). Role-Based Access Controls. *15th National Computer Security Conference*, (pp. 554 - 563). Baltimore.

[20] Defense, D. o. (1986). *Trusted computer system evaluation criteria.*

[21] Elger, B. S., Iavindrasana, J., Iacono, L. L., Müller, H., Roduit, N., Summers, P., & Wright, J. (2010). Strategies for health data exchange for secondary, cross-institutional clinical research. *Computer methods and programs in biomedicine, 99(3)*, 230-251.

[22] Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PloS one, 10(2).*

[23] Fernández-Alemán, J. L., Señor, I. C., Lozoya, P. Á. O., & Toval, A. (2013). Security and privacy in electronic health records: A systematic literature review. *Journal of biomedical informatics, 46(3)*, 541-562.

[24] Folorunso, S., Ogundepo, E. A., Basajja, M., Awotunde, J. B., Kawu, A. A., Oladipo, F. O., Abdullahi, I. (2022). FAIR machine learning model pipeline implementation of COVID-19 data. *Data Intelligence Vol. 4*.

[25] Group, I. (2022, 10 12). *About Solid*. Opgehaald van solidproject: https://solidproject.org/about

[26] Harshvardhan J. Pandit, Christophe Debruyne, Declan O'Sullivan, Dave Lewis1. (2022, 09). *GConsent - A Consent Ontology based on the GDPR.* Opgehaald van GComsent: https://openscience.adaptcentre.ie/ontologies/gconsent/main.html

[27] Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology 1(1)*, (pp. 1-136).

[28] HIPAA. (sd). *Health Insurance Portability and Accountability Act.* Opgehaald van https://www.hhs.gov/hipaa/index.html

[29] HL7. (2011). Opgehaald van Health Level Seven International: http://www.hl7.org/

[30] *ibestuurcongres 2022*. (2022, 09 14). Opgehaald van ibestuurcongres: https://www.ibestuurcongres.nl/2022

[31] Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L. O. B., Mons, B., Schultes, E., Roos, M., & Thompson, M. . (2020). A generic workflow for the data FAIRification process. . *Data Intelligence,* 2(1-2), 56-65.

[32] Khan, M. F. F., Sakamura, K. (2012). Context-aware access control for clinical information systems. *International Conference on Innovations in Information Technology* (pp. 123-128). IEEE.

[33] Kiran, G. M., & Nalini, N. (2020). Enhanced security-aware technique and ontology data access control in cloud computing. *International Journal of*

Communication Systems, 33(15), e4554.

[34] Kuhn, D. R., Coyne, E. J., & Weil, T. R. (2010). Adding attributes to role-based access control. *Computer, 43(6)*, 79-81.

[35] Lin, Y. (2021). *A FAIR Data Based BI Framework within the Healthcare Domain in Africa.*

[36] Liu, Z., & Wang, J. (2016). A fine-grained context-aware access control model for health care and life science linked data. *Multimedia Tools and Applications, 75(22)*, 14263-14280.

[37] Luiz Olavo Bonino da Silva Santos, Kees Burger, Rajaram Kaliyaperumal, Mark D. Wilkinson . (2022). FAIR Data Point: A FAIR-oriented approach for metadata publication. *Data Intelligence*.

[38] Masoumzadeh, A., & Joshi, J. B. (2011). Ontology-based access control for social network systems. *International Journal of Information Privacy, Security and Integrity (IJIPSI), 1(1)*, 59-78.

[39] Mawere, M., & van Stam, G. (2020). Data Sovereignty: A Perspective From Zimbabwe. *12th ACM Conference on Web Science Companion* , (pp. pp. 13-19).

[40] Motta, G. H., & Furuie, S. S. (2003). A contextual role-based access control authorization model for electronic patient record. *IEEE Transactions on information technology in biomedicine, 7(3)*, 202-207.

[41] Musen MA, Bean CA, Cheung K-H, Dumontier M, Durante KA, Gevaert O, Gonzalez-Beltran A, Khatri P, Kleinstein SH, O'Connor MJ. (2015). The Center for Expanded Data Annotation and Retrieval. *Journal of the American Medical Informatics Association, JAMIA.*

[42] OASIS. (2013, January). *eXtensible Access Control Markup Language (XACML) Version 3.0. 22*. Opgehaald van OASIS Standard: http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html

[43] Oladipo, F. O., Folorunso, S., Ogundepo, E. A., Osigwe, E., Akindele, A.T. (2022). Curriculum development for FAIR data stewardship. *Data Intelligence Vol. 4*.

[44] Padia, A., Finin, T., & Joshi, A. (2015). Attribute-based fine grained access control for triple stores. *14th International Semantic Web Conference.*

[45] Park, M. A. (2011). Embedding security into visual programming courses. *2011 Information Security Curriculum Development Conference*, (pp. 84-93).

[46] Peleg, M., Beimel, D., Dori, D., & Denekamp, Y. (2008). Situation-based access control: Privacy management via modeling of patient data access scenarios. *Journal of Biomedical Informatics, 41(6)*, 1028-1040.

[47] Plug, R., Liang, Y., Aktau, A., Basajja, M., Oladipo, F., Van Reisen, M. (2022). Terminology for a FAIR framework for the Virus Outbreak Data Network-Africa. *Data Intelligence Vol. 4.*

[48] Priebe, T., Dobmeier, W., Kamprath, N. (2006). Supporting attribute-based access control with ontologies. *First International conference on availability, reliability and security.* IEEE.

[49] Purnama Jati, P. H., Van Reisen, M., Flikkenschild, E., Oladipo, F. O., Meerman, B., Plug, R., Nodehi, S. (2022). Data access, control, and privacy protection in the VODAN-Africa architecture. *Data Intelligence Vol. 4*.

[50] Quantin, C., Jaquet-Chiffelle, D. O., Coatrieux, G., Benzenine, E., & Allaert, F. A. (2011). Medical record search engines, using pseudonymised patient identity: An alternative to centralised medical records. *international journal of medical informatics*.

[51] *Regulation (EU) 2016/679*. (2016, 05 04). Opgehaald van EUR-Lex: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[52] Røstad, L. (2008). An initial model and a discussion of access control in patient controlled health records. *Third International Conference on Availability, Reliability and Security* (pp. 935-942). IEEE.

[53] Shen, H. B., & Hong, F. (2006). An attribute-based access control model for web services. *Seventh International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'06)* (pp. 74-79). IEEE.

[54] Tauqeer, A., Kurteva, A., Chhetri, T. R., Ahmeti, A., & Fensel, A. (2022). Automated GDPR Contract Compliance Verification Using Knowledge Graphs. *Information, 13(10), 447.*

[55] Van Reisen, e. a. (2022). Incomplete COVID-19 data: The curation of medical health data by the Virus Outbreak Data Network-Africa. *Data Intelligence*.

[56] van Reisen, M., Oladipo, F., Stokmans, M., Mpezamihgo, M., Folorunso, S., Schultes, E., ... & Musen, M. A. (2021). Design of a FAIR digital data health infrastructure in Africa for COVID-19 reporting and research. *Advanced Genetics*, 2(2).

[57] Vincent C. Hu, David F. Ferraiolo, Ramaswamy Chandramouli, D. Richard Kuhn. (2017). *Attribute-Based Access Control.* Boston: ARTECH HOUSE.

[58] W3C. (2009, May 20). *PolicyLangReview*. Opgehaald van World Wide Web Consortium: https://www.w3.org/Policy/pling/wiki/PolicyLangReview

[59] Ware, W. H. (1979). *Security controls for computer systems. report of defense science board task force on computer security.* SANTA MONICA: RAND CORP.

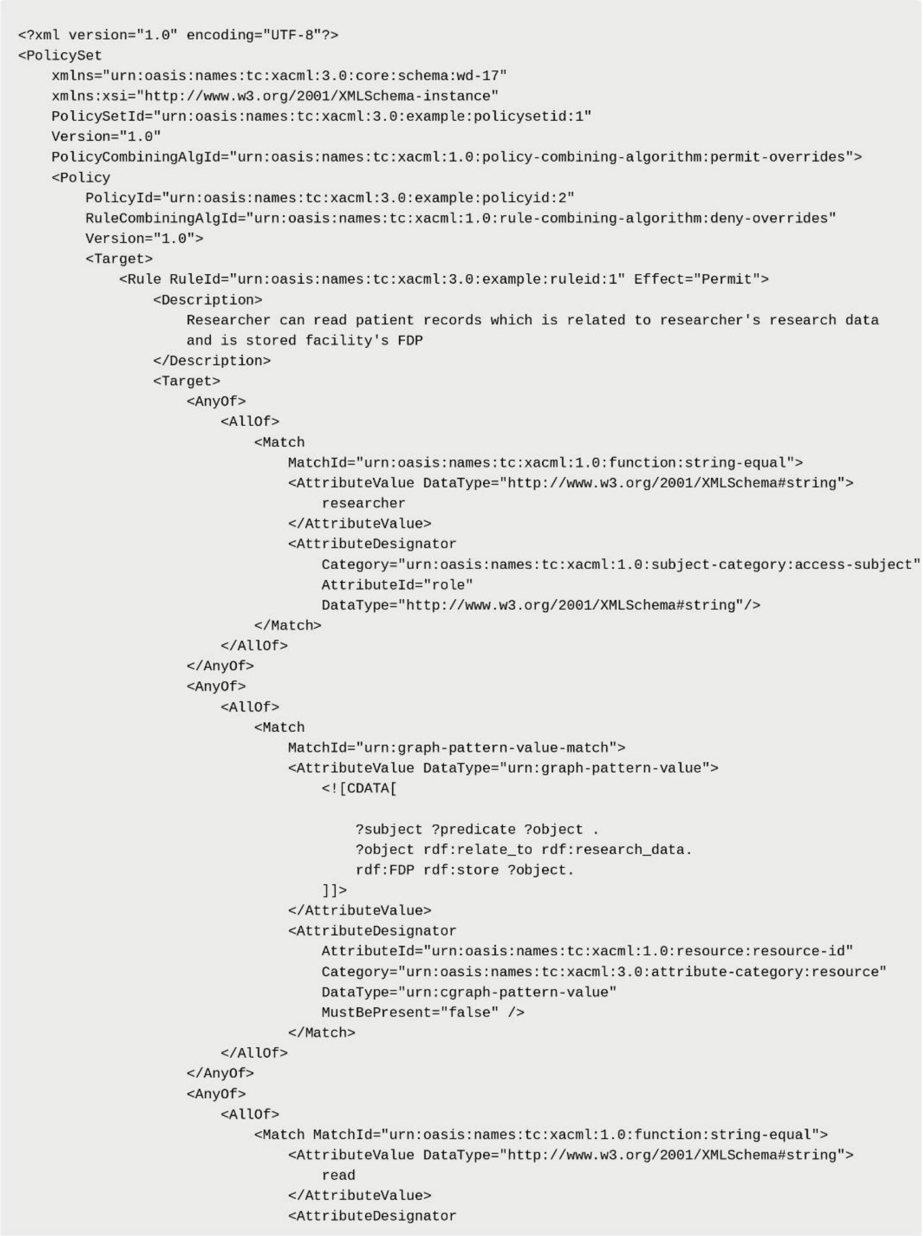
[60] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. . (2016). The FAIR Guiding Principles for scientific data management and stewardship. . *Scientific data*, 3(1), 1-9.

[61] Wolford, B. (sd). *What is GDPR, the EU's new data protection law?* Opgehaald van GDPR.EU: https://gdpr.eu/what-is-gdpr/

[62] Zhang, R., & Liu, L. (2010). Security models and requirements for healthcare application clouds. *IEEE 3rd International Conference on cloud Computing* (pp. 268-275). IEEE.

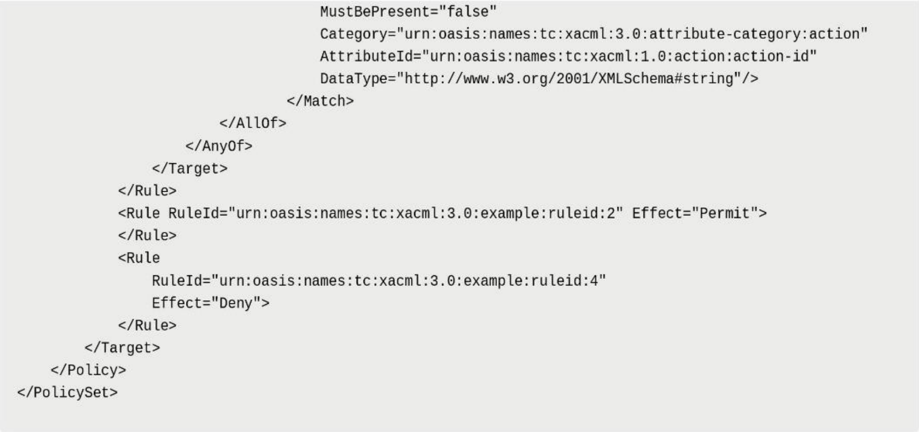
Appendix

A. Data Access Control Policy



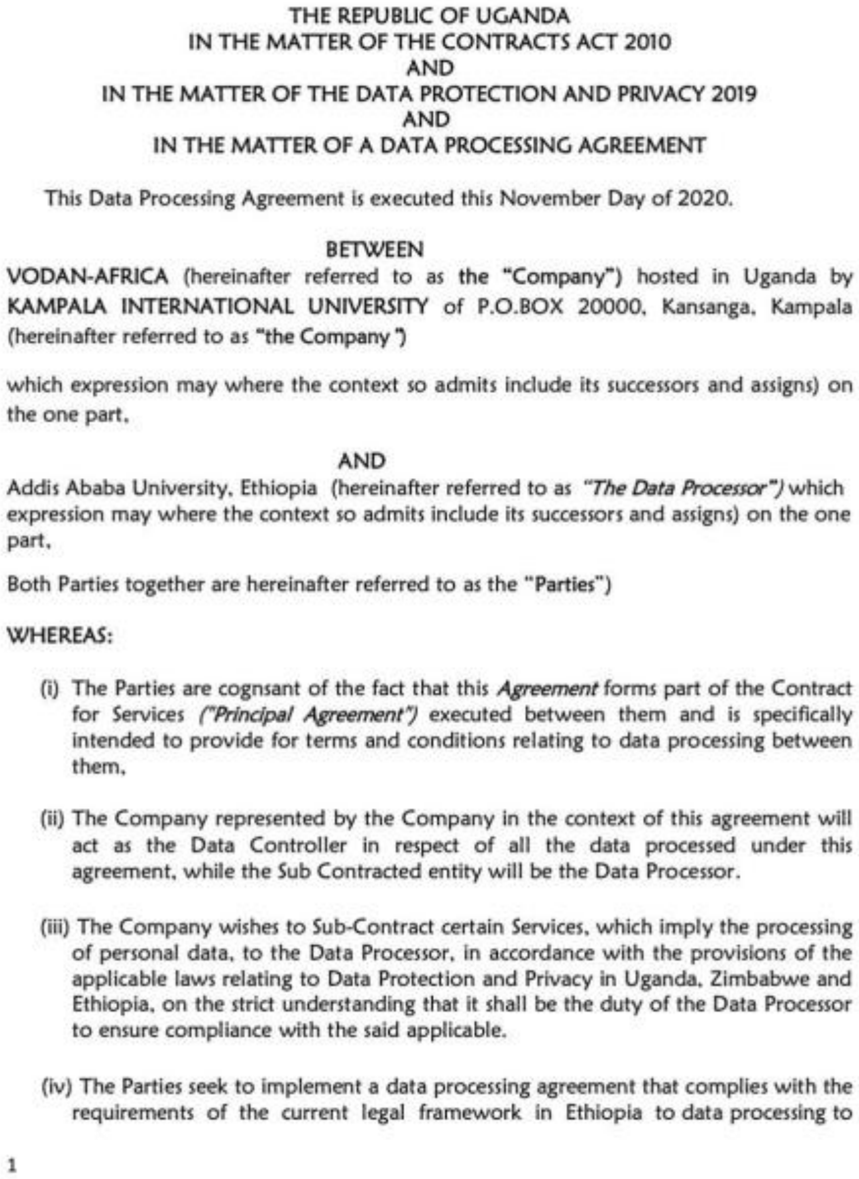
```
<?xml version="1.0" encoding="UTF-8"?>
<PolicySet
    xmlns="urn:oasis:names:tc:xacml:3.0:core:schema:wd-17"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    PolicySetId="urn:oasis:names:tc:xacml:3.0:example:policysetid:1"
    Version="1.0"
    PolicyCombiningAlgId="urn:oasis:names:tc:xacml:1.0:policy-combining-algorithm:permit-overrides">
    <Policy
        PolicyId="urn:oasis:names:tc:xacml:3.0:example:policyid:2"
        RuleCombiningAlgId="urn:oasis:names:tc:xacml:1.0:rule-combining-algorithm:deny-overrides"
        Version="1.0">
        <Target>
            <Rule RuleId="urn:oasis:names:tc:xacml:3.0:example:ruleid:1" Effect="Permit">
                <Description>
                    Researcher can read patient records which is related to researcher's research data
                    and is stored facility's FDP
                </Description>
                <Target>
                    <AnyOf>
                        <AllOf>
                            <Match
                                MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
                                <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">
                                    researcher
                                </AttributeValue>
                                <AttributeDesignator
                                    Category="urn:oasis:names:tc:xacml:1.0:subject-category:access-subject"
                                    AttributeId="role"
                                    DataType="http://www.w3.org/2001/XMLSchema#string"/>
                            </Match>
                        </AllOf>
                    </AnyOf>
                    <AnyOf>
                        <AllOf>
                            <Match
                                MatchId="urn:graph-pattern-value-match">
                                <AttributeValue DataType="urn:graph-pattern-value">
                                    <![CDATA[

                                        ?subject ?predicate ?object .
                                        ?object rdf:relate_to rdf:research_data.
                                        rdf:FDP rdf:store ?object.
                                    ]]>
                                </AttributeValue>
                                <AttributeDesignator
                                    AttributeId="urn:oasis:names:tc:xacml:1.0:resource:resource-id"
                                    Category="urn:oasis:names:tc:xacml:3.0:attribute-category:resource"
                                    DataType="urn:cgraph-pattern-value"
                                    MustBePresent="false" />
                                </Match>
                        </AllOf>
                    </AnyOf>
                    <AnyOf>
                        <AllOf>
                            <Match MatchId="urn:oasis:names:tc:xacml:1.0:function:string-equal">
                                <AttributeValue DataType="http://www.w3.org/2001/XMLSchema#string">
                                    read
                                </AttributeValue>
                                <AttributeDesignator
```



```
                                    MustBePresent="false"
                                    Category="urn:oasis:names:tc:xacml:3.0:attribute-category:action"
                                    AttributeId="urn:oasis:names:tc:xacml:1.0:action:action-id"
                                    DataType="http://www.w3.org/2001/XMLSchema#string"/>
                                </Match>
                        </AllOf>
                    </AnyOf>
                </Target>
            </Rule>
            <Rule RuleId="urn:oasis:names:tc:xacml:3.0:example:ruleid:2" Effect="Permit">
            </Rule>
            <Rule
                RuleId="urn:oasis:names:tc:xacml:3.0:example:ruleid:4"
                Effect="Deny">
            </Rule>
        </Target>
    </Policy>
</PolicySet>
```

B. Data Processing Agreement (DPA)



THE REPUBLIC OF UGANDA
IN THE MATTER OF THE CONTRACTS ACT 2010
AND
IN THE MATTER OF THE DATA PROTECTION AND PRIVACY 2019
AND
IN THE MATTER OF A DATA PROCESSING AGREEMENT

This Data Processing Agreement is executed this November Day of 2020.

BETWEEN

VODAN-AFRICA (hereinafter referred to as **the "Company")** hosted in Uganda by **KAMPALA INTERNATIONAL UNIVERSITY** of P.O.BOX 20000, Kansanga, Kampala (hereinafter referred to as **"the Company"**)

which expression may where the context so admits include its successors and assigns) on the one part,

AND

Addis Ababa University, Ethiopia (hereinafter referred to as *"The Data Processor")* which expression may where the context so admits include its successors and assigns) on the one part,

Both Parties together are hereinafter referred to as the **"Parties"**)

WHEREAS:

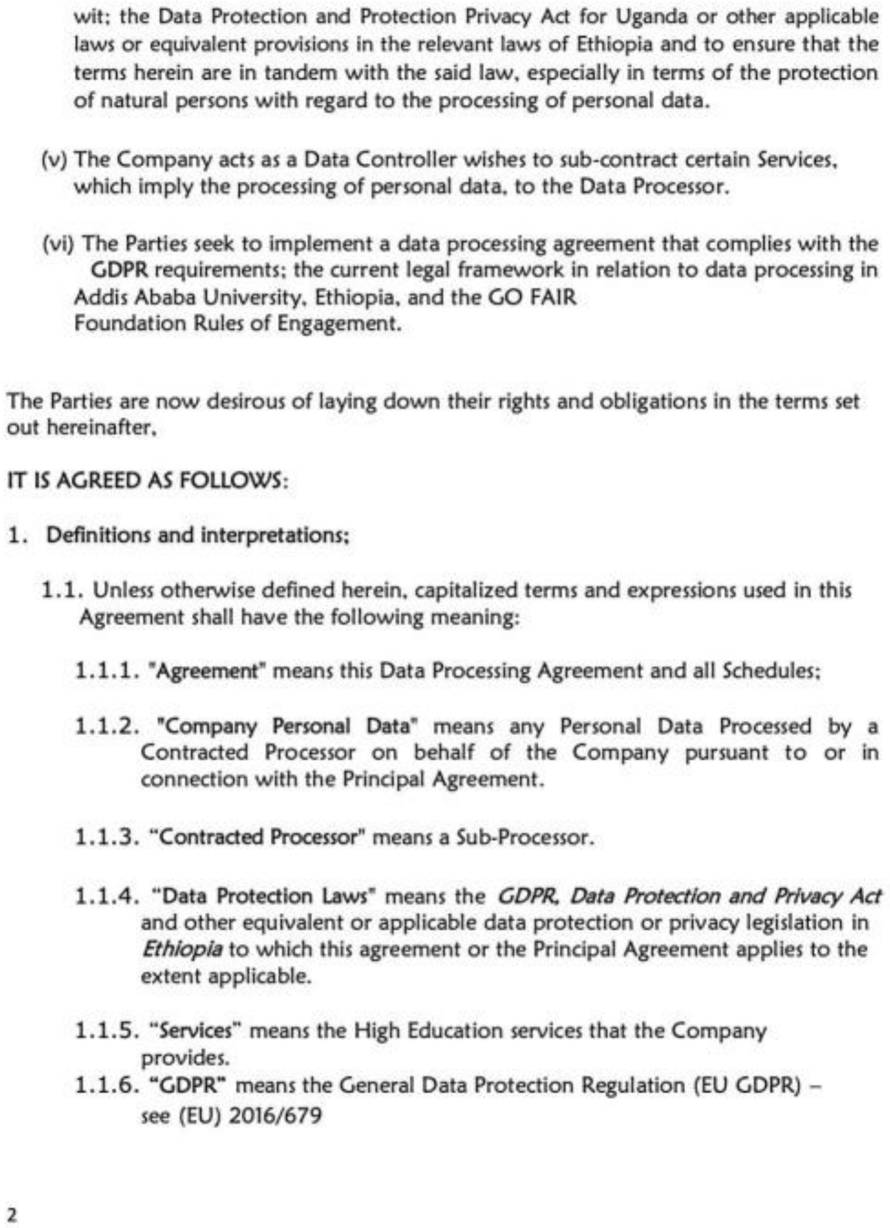
(i) The Parties are cognsant of the fact that this *Agreement* forms part of the Contract for Services *("Principal Agreement")* executed between them and is specifically intended to provide for terms and conditions relating to data processing between them,

(ii) The Company represented by the Company in the context of this agreement will act as the Data Controller in respect of all the data processed under this agreement, while the Sub Contracted entity will be the Data Processor.

(iii) The Company wishes to Sub-Contract certain Services, which imply the processing of personal data, to the Data Processor, in accordance with the provisions of the applicable laws relating to Data Protection and Privacy in Uganda, Zimbabwe and Ethiopia, on the strict understanding that it shall be the duty of the Data Processor to ensure compliance with the said applicable.

(iv) The Parties seek to implement a data processing agreement that complies with the requirements of the current legal framework in Ethiopia to data processing to

1



wit; the Data Protection and Protection Privacy Act for Uganda or other applicable laws or equivalent provisions in the relevant laws of Ethiopia and to ensure that the terms herein are in tandem with the said law, especially in terms of the protection of natural persons with regard to the processing of personal data.

(v) The Company acts as a Data Controller wishes to sub-contract certain Services, which imply the processing of personal data, to the Data Processor.

(vi) The Parties seek to implement a data processing agreement that complies with the **GDPR** requirements; the current legal framework in relation to data processing in Addis Ababa University, Ethiopia, and the GO FAIR
Foundation Rules of Engagement.

The Parties are now desirous of laying down their rights and obligations in the terms set out hereinafter,

IT IS AGREED AS FOLLOWS:

1. **Definitions and interpretations;**

1.1. Unless otherwise defined herein, capitalized terms and expressions used in this Agreement shall have the following meaning:

1.1.1. "**Agreement**" means this Data Processing Agreement and all Schedules;

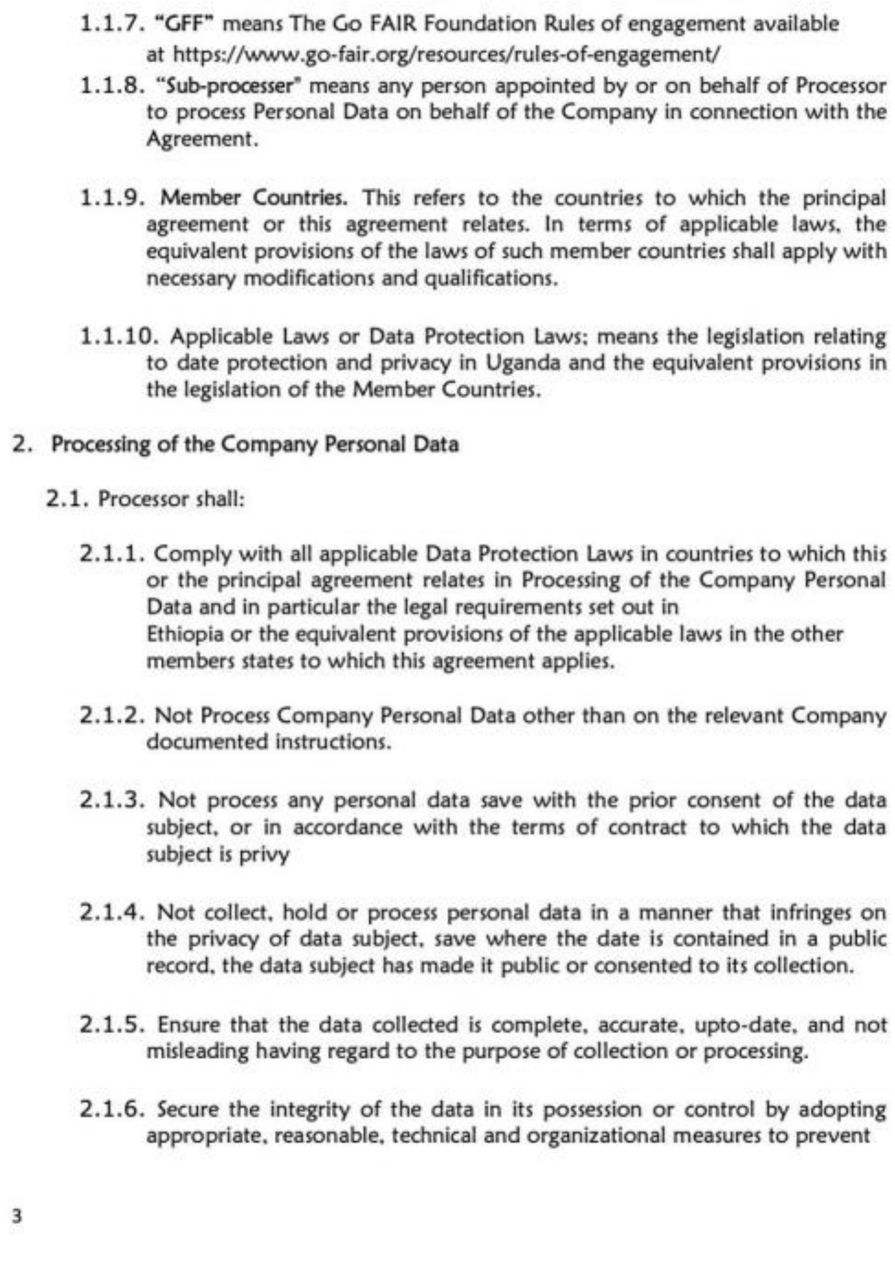
1.1.2. "**Company Personal Data**" means any Personal Data Processed by a Contracted Processor on behalf of the Company pursuant to or in connection with the Principal Agreement.

1.1.3. "**Contracted Processor**" means a Sub-Processor.

1.1.4. "**Data Protection Laws**" means the ***GDPR, Data Protection and Privacy Act*** and other equivalent or applicable data protection or privacy legislation in ***Ethiopia*** to which this agreement or the Principal Agreement applies to the extent applicable.

1.1.5. "**Services**" means the High Education services that the Company provides.

1.1.6. "**GDPR**" means the General Data Protection Regulation (EU GDPR) – see (EU) 2016/679



1.1.7. **"GFF"** means The Go FAIR Foundation Rules of engagement available at https://www.go-fair.org/resources/rules-of-engagement/

1.1.8. **"Sub-processer"** means any person appointed by or on behalf of Processor to process Personal Data on behalf of the Company in connection with the Agreement.

1.1.9. **Member Countries.** This refers to the countries to which the principal agreement or this agreement relates. In terms of applicable laws, the equivalent provisions of the laws of such member countries shall apply with necessary modifications and qualifications.

1.1.10. Applicable Laws or Data Protection Laws; means the legislation relating to date protection and privacy in Uganda and the equivalent provisions in the legislation of the Member Countries.

2. **Processing of the Company Personal Data**

2.1. Processor shall:

2.1.1. Comply with all applicable Data Protection Laws in countries to which this or the principal agreement relates in Processing of the Company Personal Data and in particular the legal requirements set out in
Ethiopia or the equivalent provisions of the applicable laws in the other members states to which this agreement applies.

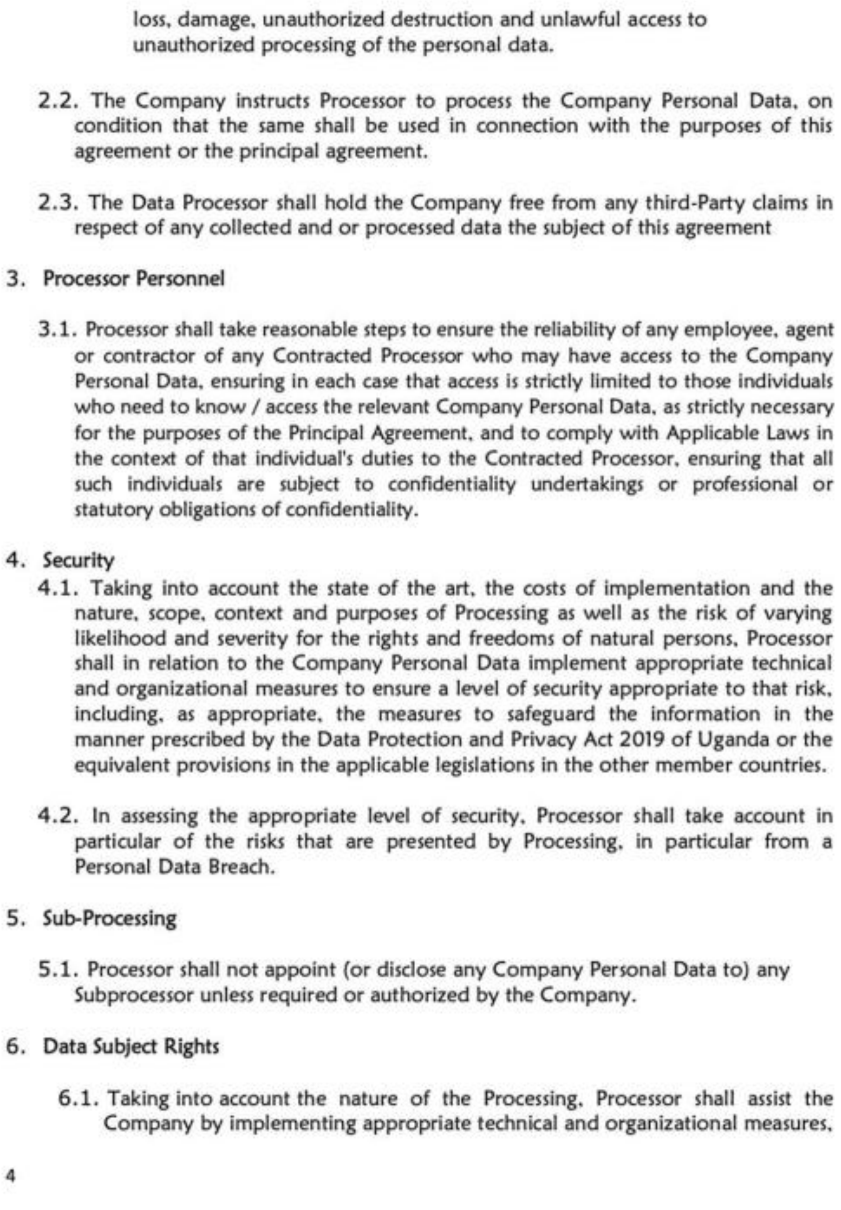
2.1.2. Not Process Company Personal Data other than on the relevant Company documented instructions.

2.1.3. Not process any personal data save with the prior consent of the data subject, or in accordance with the terms of contract to which the data subject is privy

2.1.4. Not collect, hold or process personal data in a manner that infringes on the privacy of data subject, save where the date is contained in a public record, the data subject has made it public or consented to its collection.

2.1.5. Ensure that the data collected is complete, accurate, upto-date, and not misleading having regard to the purpose of collection or processing.

2.1.6. Secure the integrity of the data in its possession or control by adopting appropriate, reasonable, technical and organizational measures to prevent



loss, damage, unauthorized destruction and unlawful access to unauthorized processing of the personal data.

2.2. The Company instructs Processor to process the Company Personal Data, on condition that the same shall be used in connection with the purposes of this agreement or the principal agreement.

2.3. The Data Processor shall hold the Company free from any third-Party claims in respect of any collected and or processed data the subject of this agreement

3. Processor Personnel

3.1. Processor shall take reasonable steps to ensure the reliability of any employee, agent or contractor of any Contracted Processor who may have access to the Company Personal Data, ensuring in each case that access is strictly limited to those individuals who need to know / access the relevant Company Personal Data, as strictly necessary for the purposes of the Principal Agreement, and to comply with Applicable Laws in the context of that individual's duties to the Contracted Processor, ensuring that all such individuals are subject to confidentiality undertakings or professional or statutory obligations of confidentiality.

4. Security

4.1. Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of Processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, Processor shall in relation to the Company Personal Data implement appropriate technical and organizational measures to ensure a level of security appropriate to that risk, including, as appropriate, the measures to safeguard the information in the manner prescribed by the Data Protection and Privacy Act 2019 of Uganda or the equivalent provisions in the applicable legislations in the other member countries.

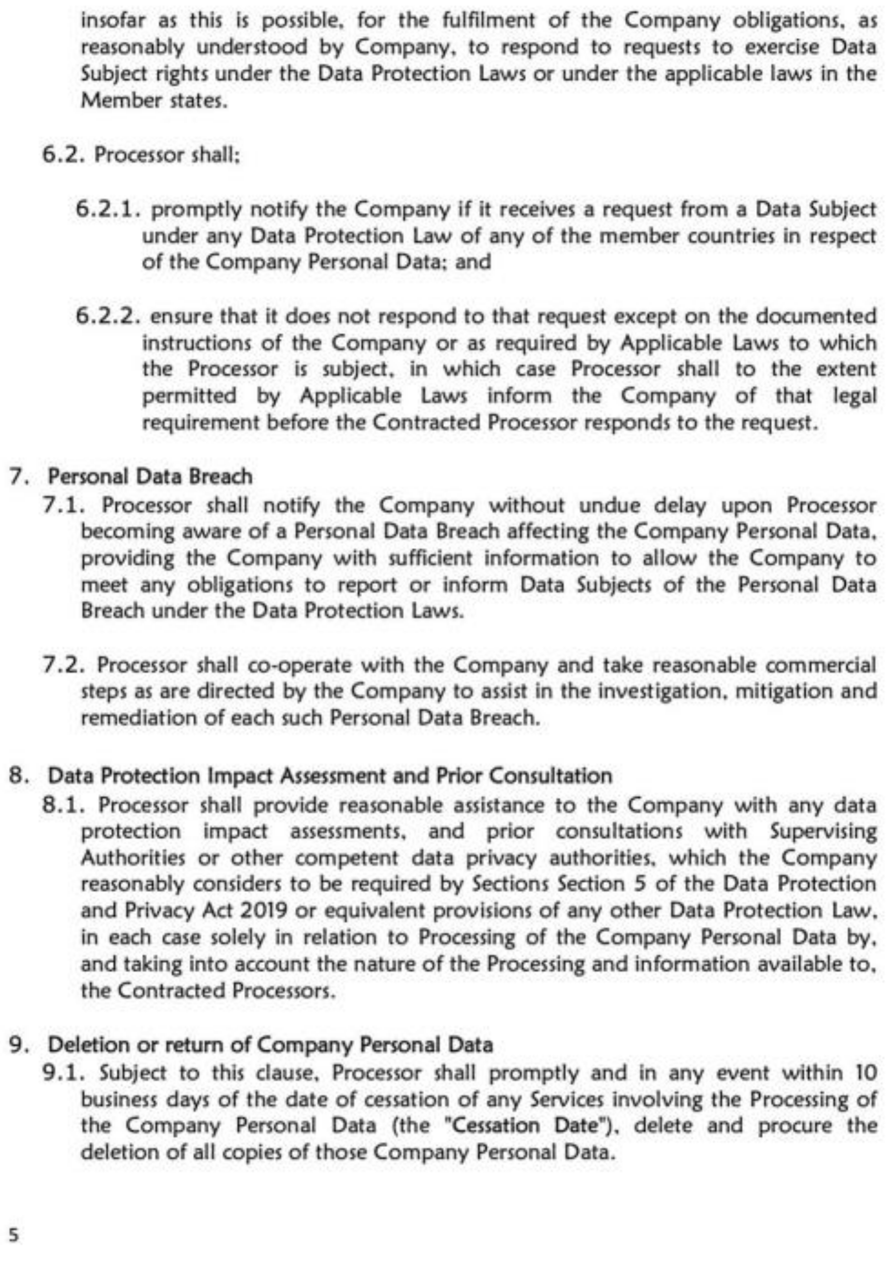
4.2. In assessing the appropriate level of security, Processor shall take account in particular of the risks that are presented by Processing, in particular from a Personal Data Breach.

5. Sub-Processing

5.1. Processor shall not appoint (or disclose any Company Personal Data to) any Subprocessor unless required or authorized by the Company.

6. Data Subject Rights

6.1. Taking into account the nature of the Processing, Processor shall assist the Company by implementing appropriate technical and organizational measures,



insofar as this is possible, for the fulfilment of the Company obligations, as reasonably understood by Company, to respond to requests to exercise Data Subject rights under the Data Protection Laws or under the applicable laws in the Member states.

6.2. Processor shall;

6.2.1. promptly notify the Company if it receives a request from a Data Subject under any Data Protection Law of any of the member countries in respect of the Company Personal Data; and

6.2.2. ensure that it does not respond to that request except on the documented instructions of the Company or as required by Applicable Laws to which the Processor is subject, in which case Processor shall to the extent permitted by Applicable Laws inform the Company of that legal requirement before the Contracted Processor responds to the request.

7. Personal Data Breach

7.1. Processor shall notify the Company without undue delay upon Processor becoming aware of a Personal Data Breach affecting the Company Personal Data, providing the Company with sufficient information to allow the Company to meet any obligations to report or inform Data Subjects of the Personal Data Breach under the Data Protection Laws.

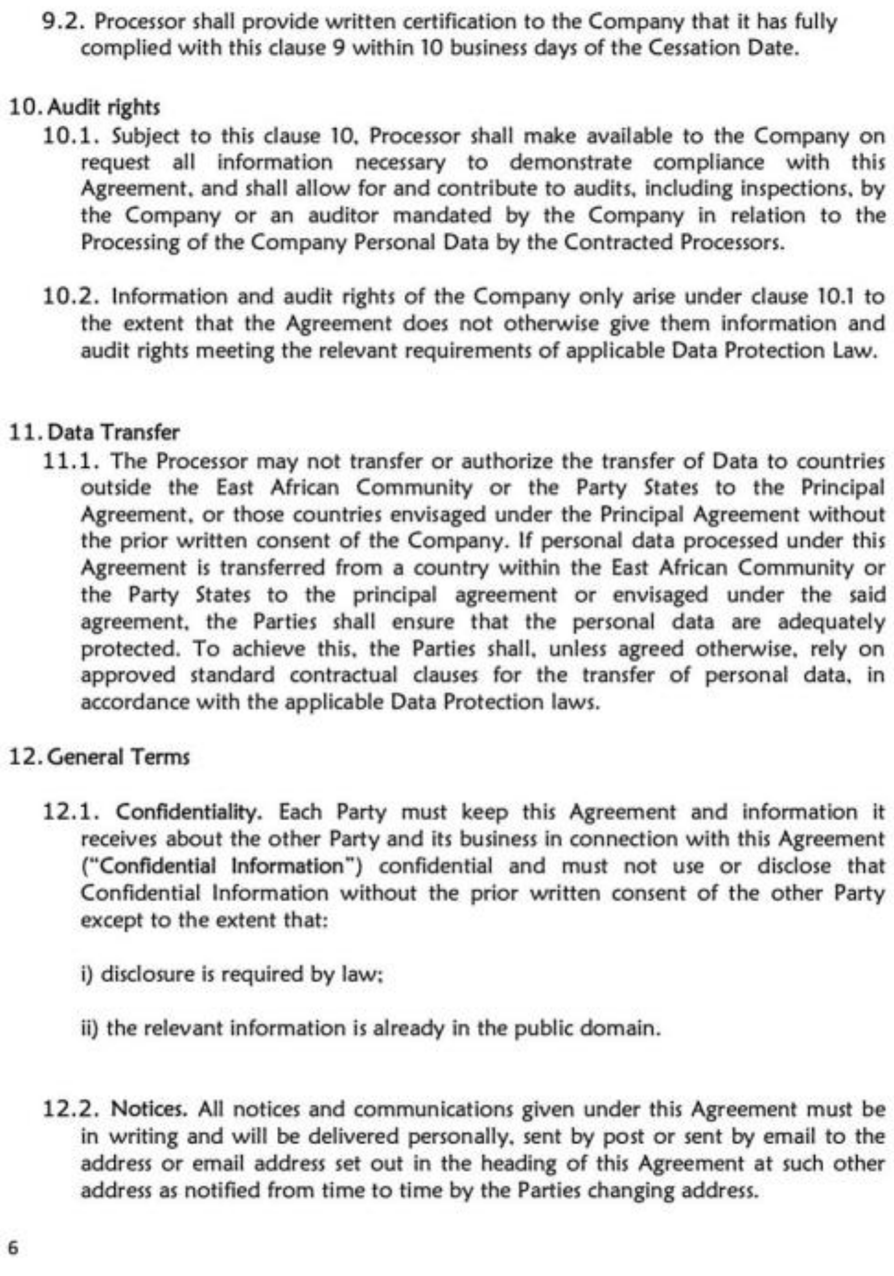
7.2. Processor shall co-operate with the Company and take reasonable commercial steps as are directed by the Company to assist in the investigation, mitigation and remediation of each such Personal Data Breach.

8. Data Protection Impact Assessment and Prior Consultation

8.1. Processor shall provide reasonable assistance to the Company with any data protection impact assessments, and prior consultations with Supervising Authorities or other competent data privacy authorities, which the Company reasonably considers to be required by Sections Section 5 of the Data Protection and Privacy Act 2019 or equivalent provisions of any other Data Protection Law, in each case solely in relation to Processing of the Company Personal Data by, and taking into account the nature of the Processing and information available to, the Contracted Processors.

9. Deletion or return of Company Personal Data

9.1. Subject to this clause, Processor shall promptly and in any event within 10 business days of the date of cessation of any Services involving the Processing of the Company Personal Data (the "**Cessation Date**"), delete and procure the deletion of all copies of those Company Personal Data.



9.2. Processor shall provide written certification to the Company that it has fully complied with this clause 9 within 10 business days of the Cessation Date.

10. Audit rights

10.1. Subject to this clause 10, Processor shall make available to the Company on request all information necessary to demonstrate compliance with this Agreement, and shall allow for and contribute to audits, including inspections, by the Company or an auditor mandated by the Company in relation to the Processing of the Company Personal Data by the Contracted Processors.

10.2. Information and audit rights of the Company only arise under clause 10.1 to the extent that the Agreement does not otherwise give them information and audit rights meeting the relevant requirements of applicable Data Protection Law.

11. Data Transfer

11.1. The Processor may not transfer or authorize the transfer of Data to countries outside the East African Community or the Party States to the Principal Agreement, or those countries envisaged under the Principal Agreement without the prior written consent of the Company. If personal data processed under this Agreement is transferred from a country within the East African Community or the Party States to the principal agreement or envisaged under the said agreement, the Parties shall ensure that the personal data are adequately protected. To achieve this, the Parties shall, unless agreed otherwise, rely on approved standard contractual clauses for the transfer of personal data, in accordance with the applicable Data Protection laws.

12. General Terms

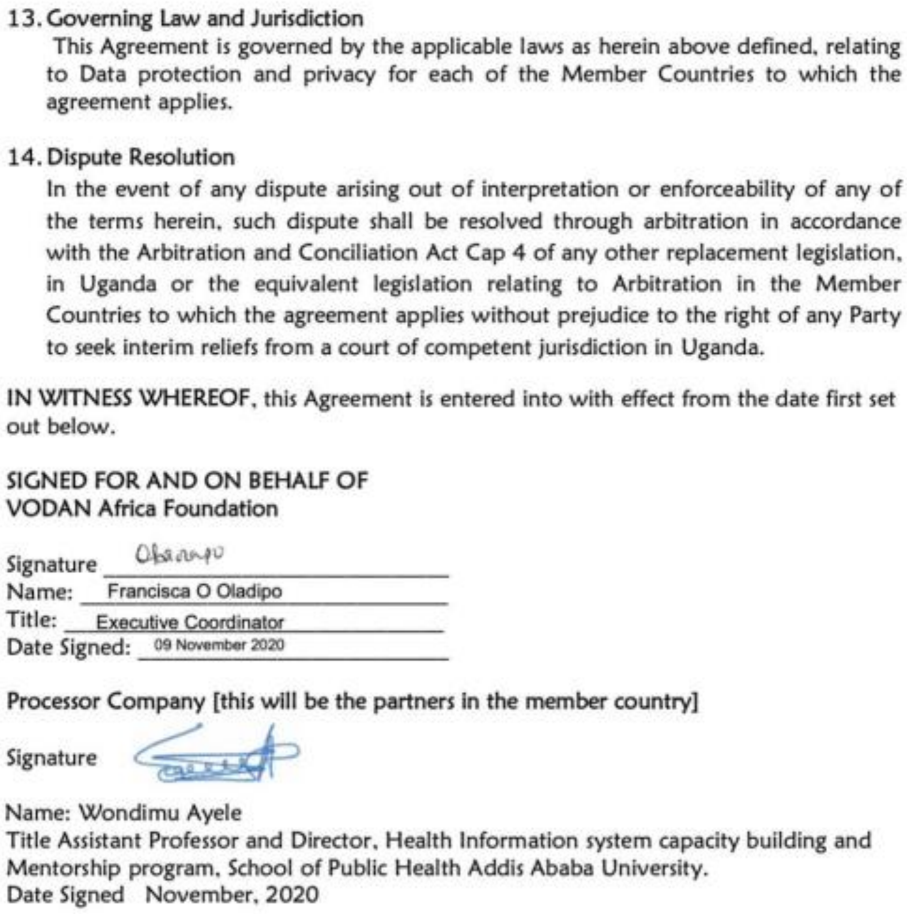
12.1. **Confidentiality.** Each Party must keep this Agreement and information it receives about the other Party and its business in connection with this Agreement ("**Confidential Information**") confidential and must not use or disclose that Confidential Information without the prior written consent of the other Party except to the extent that:

i) disclosure is required by law;

ii) the relevant information is already in the public domain.

12.2. **Notices.** All notices and communications given under this Agreement must be in writing and will be delivered personally, sent by post or sent by email to the address or email address set out in the heading of this Agreement at such other address as notified from time to time by the Parties changing address.

6



13. **Governing Law and Jurisdiction**

    This Agreement is governed by the applicable laws as herein above defined, relating to Data protection and privacy for each of the Member Countries to which the agreement applies.

14. **Dispute Resolution**

    In the event of any dispute arising out of interpretation or enforceability of any of the terms herein, such dispute shall be resolved through arbitration in accordance with the Arbitration and Conciliation Act Cap 4 of any other replacement legislation, in Uganda or the equivalent legislation relating to Arbitration in the Member Countries to which the agreement applies without prejudice to the right of any Party to seek interim reliefs from a court of competent jurisdiction in Uganda.

IN WITNESS WHEREOF, this Agreement is entered into with effect from the date first set out below.

SIGNED FOR AND ON BEHALF OF
VODAN Africa Foundation

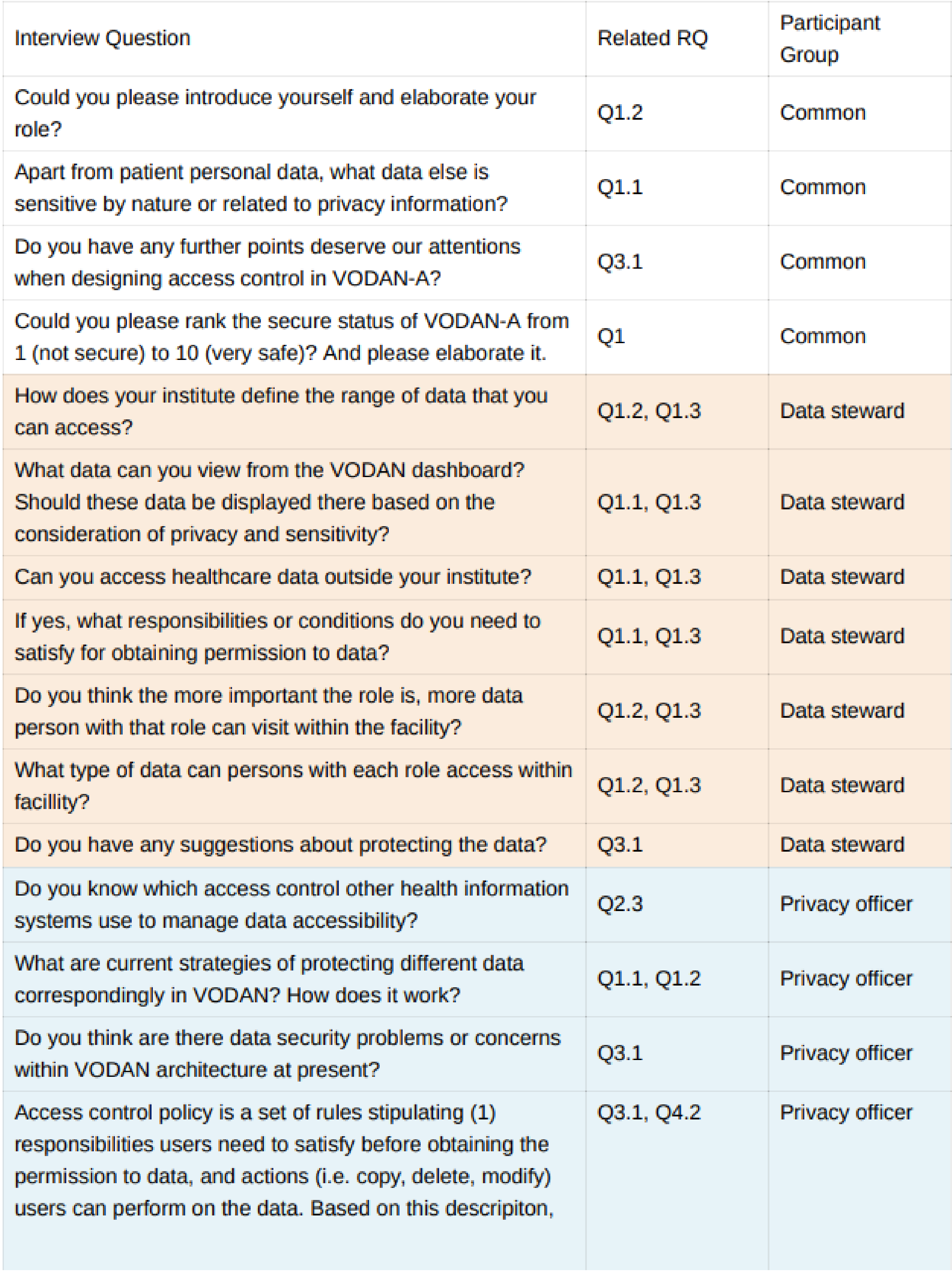
Signature Oladipo
Name: Francisca O Oladipo
Title: Executive Coordinator
Date Signed: 09 November 2020

Processor Company [this will be the partners in the member country]

Signature

Name: Wondimu Ayele
Title Assistant Professor and Director, Health Information system capacity building and Mentorship program, School of Public Health Addis Ababa University.
Date Signed November, 2020

C. Interview Questions



Interview Question	Related RQ	Participant Group
Could you please introduce yourself and elaborate your role?	Q1.2	Common
Apart from patient personal data, what data else is sensitive by nature or related to privacy information?	Q1.1	Common
Do you have any further points deserve our attentions when designing access control in VODAN-A?	Q3.1	Common
Could you please rank the secure status of VODAN-A from 1 (not secure) to 10 (very safe)? And please elaborate it.	Q1	Common
How does your institute define the range of data that you can access?	Q1.2, Q1.3	Data steward
What data can you view from the VODAN dashboard? Should these data be displayed there based on the consideration of privacy and sensitivity?	Q1.1, Q1.3	Data steward
Can you access healthcare data outside your institute?	Q1.1, Q1.3	Data steward
If yes, what responsibilities or conditions do you need to satisfy for obtaining permission to data?	Q1.1, Q1.3	Data steward
Do you think the more important the role is, more data person with that role can visit within the facility?	Q1.2, Q1.3	Data steward
What type of data can persons with each role access within facillity?	Q1.2, Q1.3	Data steward
Do you have any suggestions about protecting the data?	Q3.1	Data steward
Do you know which access control other health information systems use to manage data accessibility?	Q2.3	Privacy officer
What are current strategies of protecting different data correspondingly in VODAN? How does it work?	Q1.1, Q1.2	Privacy officer
Do you think are there data security problems or concerns within VODAN architecture at present?	Q3.1	Privacy officer
Access control policy is a set of rules stipulating (1) responsibilities users need to satisfy before obtaining the permission to data, and actions (i.e. copy, delete, modify) users can perform on the data. Based on this descripiton,	Q3.1, Q4.2	Privacy officer

what prerequisites else do you think should data processors to satisfy when accessing the data?		
Does data processing agreement (DPA) include factors to configure the permissions to data?	Q1.2, Q4.2	Privacy officer
What responsibities do you want data processors to fulfill before accessing the data	Q3.1, Q4.2	Privacy officer
What ontology or metadata may directly or indirectly link to sensitive information?	Q1.1	Technical team
What sensitive attributes may be contained in the CEDAR templates or controlled vocabularies?	Q1.1	Technical team
Which access control methods do you recommend for managing the accessibility of metadata?	Q1.2	Technical team
Which access control do you implement in the internal VODAN dashboard?	Q1.3	Technical team
What are data security problems or concerns within VODAN architecture at present?	Q3.1	Technical team
In VODAN architecture, how do you control data travelling between Allegrograph and external dynamic SPARQL queries?	Q1.1	Technical team
In VODAN architecture, how do you control the data exposed to DHIS2?	Q1.1	Technical team
What is the difference between data displayed in the internal dashboard and external one?	Q1.3	Technical team
In "Access Control" page of internal dashboard, what data resources does it control?	Q1.2	Technical team
In "Access Control" page of internal dashboard, how do you configure the permission to data?	Q1.2	Technical team
Do you have any suggestions about setting access control over SPARQL dynamic queries?	Q3.2	Technical team
Apart from patient personal data, what (meta)data else is sensitive by nature or related to privacy information?	Q1.1	Relevant stakeholders
Do you know which accss control models do other (health) information systems (i.e. anDREa, DHIS2) use to manage data accessibility	Q2.3	Relevant stakeholders
What are current strategies of protecting different	Q1.1, Q1.2	Relevant

(meta)data correspondingly in VODAN-A? How does it work?		stakeholders
What are data security problems or concerns within VODAN architecture at present?	Q3.1	Relevant stakeholders
Are there new updates in the data protection regulations recently? To what extent are these updates associated with FAIR guideline?	Q4.3	Relevant stakeholders