Exploring Strategies to Establish a Big Data Integration and Harmonization Framework for National Identity Databases

Dissertation Manuscript

Submitted to Northcentral University

School of Technology

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

by

DENEKEW ABERA JEMBERE

La Jolla, California

June 2022

Approval Page

Exploring Strategies to Establish a Big Data Integration and Harmonization Framework for National Identity Databases

By

DENEKEW ABERA JEMBERE

Approved by the Doctoral Committee:

DocuSigned by: Will Tribbey 1240E67C9457406...	Ph.D.	08/18/2022 | 10:00:26 MST
Dissertation Chair: Will Tribbey	Degree Held	Date
DocuSigned by: Christopher Weppler 08612569C0724F6...	DCS	08/22/2022 | 04:18:42 MST
Committee Member: Christopher Weppler	Degree Held	Date
DocuSigned by: Milton Kabia BA3218B4929B445...	Ph.D., DHA	08/18/2022 | 10:15:20 MST
Committee Member: Milton Kabia	Degree Held	Date

Abstract

Integrating data from various standalone data sources into a centrally managed data storage has been recognized by data integration practitioners and authors as a slow and daunting operation. This study addressed the lack of a standardized data harmonization framework for integrating a national identity database with disparate digital identity data sources to implement a national digital identity system in developing countries. Components of a design theory are used as the study's theoretical framework to explore and establish the data harmonization integration framework through the coalition of big data theories, ISO standards, and the principles of digital identification for sustainable development. An exploratory qualitative research design was utilized with open-ended semi-structured questions through face-to-face and virtual interviews. Data were collected from 9 subject matter experts and analyzed to explore in-depth information, explore new ideas, gain insights, and outline the proposed data harmonization and integration framework details. This study established an organized data integration approach and contributed toward developing a standardized data integration and harmonization framework for a national digital identity system. The data harmonization and integration framework facilitate collaboration among the digital identity solution stakeholders and accelerate the digital identity system's implementation to enable countries to gain the benefits of a digital identity system. This study also identified data elements and components for establishing a national digital identity-related big data analytics system. The identified data elements and components of big data analytics can facilitate the practical implementation of national digital identity-related big data analytics. Since the suggested big data analytics components are identified mainly based on participants' theoretical responses, expanding and verifying the findings using practical experiences in national digital identity systems is suggested to maximize the study's contribution.

Acknowledgements

I would like to thank the Lord for the guidance, wisdom, and strength He gave me throughout my study. Without His blessings and enriching spirit, I would not have been able to get to this point in life and withstand the rigors of this study. I would also like to thank the Holy Mother, Virgin Mary, for keeping my soul at peace throughout my life and this study. My deepest gratitude also goes to my parents, my father, R.D. Abera Jembere (in heaven), and my mother, Wro Etaferahu Gichamo, for instilling in me the desire to achieve the highest education.

I would like to express my deepest gratitude to my wife, Wro Misrak, without whom this study would not have been conceivable. Misreye, thank you for always being there for me, giving me the love I needed, listening to me, and counseling me on my study details. To my kids, Afomia, Henok, Ephratah, and Mahlet, sorry for the family time I have taken away from you throughout my study, and thank you for your understanding. Kids, if I, the "kid from nowhere," can get this far, I know you could keep the impetus and go even further. Furthermore, I would like to thank my siblings: Wro Debrework, Ato Abayneh, Ato Ayalew, Wro Serkalem, Kesis Kebede, Wro Tsehay, Wro Berhane, Ato Endale; my nephew Ato Fasika, my brother-in-law Ato Netsanet, and my sister-in-law Wrt Ywubdar for their continual support and motivation to complete what I started.

I would like to thank my religious fathers and community at the Seattle St. Teklehaymanot Church and my friends Dr. Bezalem, Wro Tsedale, Dr. Habtamu, Wro Frehiwot, Dn Tegene, Ato Mastewal, Ato Tadesse, Ato Astatke, and Ato Kebede for their prayers and support. Thanks to my colleagues at Microsoft, Snigdha, Jeff, Christopher, and all my teammates for your pep talks, and encouragement when I most needed them. Especially for Snigdha, thank you for nudging me to start this study in time soon after I shared my plans with you.

I would like to thank all the study participants for providing me with the required expert details and data needed for this study. Finally, I would like to thank my Chair Dr. Will Tribbey, who provided me with guidance and invaluable feedback throughout this study. I would also like to thank Dr. Christopher Weppler, my Subject Matter Expert, and Dr. Milton Kabia, my Academic Reader, for their invaluable feedback and assistance in completing my dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The rapid growth and accrual of big data from heterogeneous data sources with varying data formats intensified the existing data integration and interoperability challenges, which have become the subject of multiple studies (Brown et al., 2017; Kadadi et al., 2014; Nashipudimath et al., 2020). Despite the approaches proposed in the mentioned studies, data integration and interoperability are still a challenge for both public and private organizations (Festus et al., 2017; World Bank Group [WBG], 2017). The focus of this study is the lack of a standardized data integration framework for integrating a centrally managed digital identity national database with siloed identity data sources for the implementation of digital identity in developing countries.

Since there are an estimated one billion people in the World lacking a legally recognized form of identification (World Bank Group [WBG], 2018), individuals' lack of legal identification is a global issue that requires a global initiative. To address this issue and enable all people to access public and private services with an authorized form of identification, the WBG initiated the identification for development (ID4D) project in 2014 (WBG, 2018).

In line with WBG's initiative, countries have implemented digital identity systems with varying success levels and benefited from the system, enabling their citizens and local and global residents to access e-services from anywhere globally (Mamani & Shriram, 2018; Tampon & Masso, 2019). However, while countries like Estonia have become the World's most digital countries, with 99% of public e-services to their citizens (Martinson, 2019), other countries like Jamaica and Nigeria have faced issues when implementing the digital identity system (Dunn, 2020; Uzonite & Onapajo, 2019).

Inclusive and trusted digital identification unlocks opportunities for the World's most vulnerable (White et al., 2019). However, the implementation of national digital ID is very slow

in several countries (White et al., 2019; World Bank Group [WBG], 2018). Based on a survey in 17 African countries, the WBG (2017) reported that most countries have highly fragmented identity systems maintained in parallel and siloed databases for civil registration, identification, and functional uses. Furthermore, the highly siloed and decentralized digital identity and civil registration systems lack data standardization processes between the sources. In contrast, the highly centralized systems have difficulty functioning at the local level (Wenz et al., 2018).

The status of digital identity in Ethiopia, which is among the 17 African countries surveyed by the WBD (2017), is unique. In Ethiopia, various functional ID schemes are based on a single foundational ID. The foundational ID in Ethiopia, called *kebele ID,* is administered by about 17,000 kebeles (kebele – in Amharic, the official Ethiopian language, meaning neighborhoods) on a decentralized basis (Gelb & Palacios, 2016). According to Gelb and Palacios, the kebele ID system is a significant and widespread paper-based ID held by 40 million adults, about 30% of the population of Ethiopia.

Each of the 17,000 kebeles autonomously designs the data details on the Ethiopian kebele ID and the ID format, making it the most complex data integration challenge for creating a national digital ID based on the kebele ID (Gelb & Palacios, 2016). In addition, the vital events registration agency (VERA) of Ethiopia is the autonomous national agency to oversee, support, register, and maintain vital events and records. According to Gelb and Palacios, the regional VERA (RVERA) follows Ethiopia's federal structure. Therefore, each RVERA is autonomous in deciding the type and format of data collection from the vital events, which causes data integration challenges with the target national digital ID database. Therefore, this study proposed a standard data harmonization framework for integrating data from siloed data sources.

Statement of the Problem

The problem addressed by this study is the lack of a standardized data harmonization framework for integrating a centrally managed digital identity national database with disparate standalone identity data sources for the implementation of digital identity in developing countries (Festus et al., 2017; World Bank Group [WBG], 2017). Integrating data sets from multiple standalone data sources into a centrally managed identity database has been a slow and daunting process for many countries struggling to implement the national identity databases. This problem has existed in 17 African countries surveyed by the WBG (2017). Most of these countries have highly fragmented identity systems maintained in different databases for civil registration, identification, and functional uses.

The highly fragmented identity databases observed in most of the 17 African countries (WBG, 2017) and the ad hoc data harmonization process from such fragmented databases into a central digital identity database played a significant role in slowing down the rollout of digital identity (Aziz, 2017; WBG, 2017). Unfortunately, the existing body of literature does not adequately address the data harmonization issue that requires significant time and labor for mapping and moving data from heterogeneous sources into a centrally managed digital identity database (Festus et al., 2017).

The data harmonization issue hinders digital identity solution providers from effectively implementing a digital identity database. Unless digital identity solution providers are equipped with a standardized framework for efficiently harmonizing data from different data sources into the target centrally managed digital identity database, they may be disabled from timely enabling the countries with digital identity (Aziz, 2017). As a result, the countries could fail to realize the 3 to 13 percent economic benefit from the timely implementation of the digital identity (White et

al., 2019), and they could fail to deliver on their commitment to strengthen their national identity system by 2030 (Esquivel-Korsiak & Mittal, 2018).

Purpose of the Study

In response to the absence of a standardized data harmonization framework, this qualitative exploratory study aimed to explore, understand, and establish a standardized data harmonization framework used for integrating a national identity database with siloed data sources. The proposed study's framework helps digital identity solution developers and technology leaders effectively communicate the integration details and accelerate the implementation of national digital identity (NDID) databases for countries. Furthermore, the underlying details of the data harmonization framework could also be used to define big data analytics elements and components used to create hindsight, insights, and data movement visualizations to make data-driven decisions.

Introduction to Theoretical Framework

The theoretical framework of this study was the amalgamation of components of design theory (DT) (Agogué & Kazakçi, 2014; Gregor & Jones, 2007) with concepts from three sources. The first source comprises the theories of big data and data analytics used to organize and describe the varying data types from heterogeneous sources (Daniel, 2018; Guerra et al., 2019; Sazontev & Stupnikov, 2019; Yu & Wu, 2020). The second and third sources are the set of digital identification principles published by the World Bank Group (2021) and the related standards and frameworks defined by the international organization for standardization (ISO).

The World Bank Group published ten *principles of the identification for sustainable development*, which more than thirty organizations endorse and are categorized into inclusion, design, and governance principles (World Bank Group [WBG], 2021). The principles focused on

designing trusted, secure, interoperable, and operationally sustainable identity systems were used to assess the data integration framework's security, interoperability, and stainability.

In this study, the ISO standards and frameworks were used for assessing data integration needs while complying with the privacy, security, and format of the data. First, the specific biometric-related ISO standards (ISO, 2009a; ISO, 2014; ISO, 2018a) were checked to assess the standardized biometric data exchange formats, enrolment requirements, and security evaluation needs. For accessing the general data security and privacy-related details needed for data integration needs, related data privacy ISO standards (ISO, 2011; ISO, 2015; ISO, 2018b; ISO, 2019b; ISO, 2020) were used. Finally, the ISO standard framework for authentication assurance (ISO, 2013) was consulted for assessing the scalability of the data integration at the target identity system.

The theoretical framework of this study combined DT with the principles of digital identity, standards defined by the international organization for standardization (ISO), and theories of data science with a focus on the latest technological advancement in data integration, big data, and data analytics studies. Furthermore, in the context of national digital identity (NDID), this study ventured to fill the knowledge gap in big data integration and analytics to investigate and recommend a harmonized data integration framework that could accelerate the implementation of digital identity solutions in different countries.

Establishing the theoretical constructs of this study on global principles and ISO standards would make the study's contribution scope broader than the study's target country. Apart from accelerating countries' ability to implement an NDID system to realize the 3 to 13 percent economic benefit from the solution (White et al., 2019), this study would advance knowledge in data integration and big data analytics fields. Furthermore, the output of this study

would also facilitate the design and implementation of big data analytics solutions that would, in turn, assist in the formulation of citizens-based data-driven policies and programs in the target countries (Blazquez & Domenech, 2018; Hilbert, 2016).

Introduction to Research Methodology and Design

The research method that seems appropriate for this study is the qualitative method research. The qualitative method involves a data collection approach whereby the researcher plays a significant role with inductive and recursive analysis of data to accrue a complex picture of the research question or questions under investigation and reflect the researcher's role and responsibility in the study (Yilmaz, 2013). The selection of the qualitative method research over the quantitative (Barczak, 2015) and mixed-method (Heyvaert et al., 2013) research designs are briefly summarized in this section.

The research questions of this study do not involve any dependent or independent variables that are amenable to quantitative data collection and statistical analysis. Since there is no theory or hypothesis testing with numerical data collection and statistical data analysis, quantitative method research and, by extension, the mixed research methods are not appropriate for this study. Therefore, in line with the statement of the problem, purpose, and the associated research questions defined in this study, the qualitative research method is the most appropriate method for the nature of this study.

For selecting a specific qualitative research design, qualitative designs such as the ethnographic study (Khoo et al., 2012; Wakimoto, 2013), the phenomenological study (Williams, 2021), the narrative study (Papakitsou, 2020), and the multiple case study (Tsang, 2014) designs are evaluated. The ethnographic, phenomenological, and narrative qualitative study designs explore the human experience, social behaviors, and settings of an informant-based story,

respectively (Papakitsou, 2020; Wakimoto, 2013; Williams, 2021). On the other hand, the multiple-case design enables the investigator to gain a comprehensive view of the phenomenon under investigation based on diverse sources of evidence (Tsang, 2014).

This study explores and establishes a standardized data harmonization and integration framework based on data collected from experts involved in NDID solution implementations. For developing a comprehensive explanation of a problem under investigation, a qualitative case study improves and expands the researcher's awareness of potential complex scenarios in a real-world setting by answering the how and why questions (Cronin, 2014; Snyder, 20212; Suryani, 2013). Therefore, of the qualitative research designs (i.e., ethnographic, phenomenological, narrative, and multiple cases), the multiple case design (Tsang, 2014) is the most appropriate qualitative design for this study.

Research Questions

The research questions for this qualitative study were:

RQ1

How can national digital identity solution providers and stakeholders be supported in integrating national digital identity data from multiple standalone data sources?

RQ2

What are the national digital identity and related components that big data analysts need to establish for a digital identity big data analytics system?

Significance of the Study

This study explored approaches to establish a standardized big data integration and harmonization framework. The proposed standardized data harmonization framework would enable digital identity solution providers to effectively and efficiently implement national

identity databases integrated from different standalone digital identity data sources. Efficient implementation of national identity databases with the digital ID programs prioritized for use cases that generate meaningful value for both individuals and institutions that entail frequent use benefit individuals in cost and time savings or more accessible products and services (White et al., 2019).

According to White et al. (2019), extended and full digital ID coverage with multiple high-value use cases to attain high levels of adoption could unlock economic benefits equivalent to up to 13 percent of GDP in 2030, especially for developing countries. Furthermore, timely implementation of the national identity database enables countries to deliver on their sustainable development commitment to strengthen their national identity system by 2030 (Esquivel-Korsiak & Mittal, 2018). The stakeholders considered for this study are information technology (IT) solution developers, big data analysts, and IT leaders in governmental and private organizations involved in implementing a digital national identity solution.

Various frameworks were proposed to facilitate the big data solution implementation and integration processes (Chalaemwongwan & Kurutach, 2018; Festus et al., 2017). However, approaches proposed in the existing literature do not adequately address the big data harmonization issue that requires significant time and labor to integrate different data sets into a central digital identity database. The digital national identity solution stakeholders recognize the importance of a data integration framework to accelerate the digital identity system's implementation to enable countries to achieve their digital identity commitment by 2030 (Esquivel-Korsiak & Mittal, 2018; Martin, 2021, Saputro et al., 2020). Therefore, the intent to explore approaches to establish a standardized big data integration and harmonization framework to respond to the problem makes this study unique.

Developing a standardized big data integration and harmonization framework addresses the time-consuming and tedious integration of diverse data sets into a central identity database. Moreover, the framework fills the knowledge gap in computer science by advancing the field of big data processing and integration (Hilbert, 2016). The outcome of this study benefits academics, researchers, and practitioners in identifying approaches needed to create a well-harmonized big data integration solution.

Definitions of Key Terms

Big data

Big data refers to a large and complex data set that is made up of an assortment of structured, semi-structured, and unstructured data and described in the following characters: high volume, velocity, variety, veracity, and value (Heripracoyo & Kurniawan, 2016; Power, 2014; Shatnawi et al., 2019).

Big data analytics

Big data analytics is the method of applying data processing, exploration, and mining to large structured and unstructured data sets to extract valuable insights from the data to enhance the corporate decision-making processes and improve operational efficiencies (Anusha et al., 2021; Ardagna et al., 2021; Waterman & Bruening, 2014; Zhang et al., 2015).

The big data analytics system

Big data analytics system is an integrated big data analytics environment with tools and services used for processing, exploration, and mining big data sets to extract valuable insights from the data to enhance the corporate decision-making processes (Anusha et al., 2021; Zhang et al., 2015). A big data analytics system involves tools and services that enable data analysts to create statistical, contextual, quantitative, predictive, cognitive, and prescriptive models to

produce real-time insights and information to enhance decision-making processes and operational efficiencies (Akter & Wamba, 2016).

Data harmonization

Data harmonization is a process that combines multiple data sets into a larger dataset to make sense of the multiple smaller data sets and to understand specific phenomena better. In addition, data harmonization (DH) corresponds to a field that unifies the representation of disparate data from heterogeneous and decentralized data sources that would otherwise affect data visualization and prediction by influencing analytical results (Kumar et al., 2021). Data harmonization requires using the same units, scales, and terminology, to ensure sensitivity and less subject to miss-classification and biased conclusions using the large data sets that are not clear from smaller, more exclusive data sets (Boffetta et al., 2014; Data Harmonization, 2021). Furthermore, data harmonization may involve combining multiple years of data or much more complex tasks such as combining data collected in multiple locations, languages, and periods or collecting many different data collectors. (Data Harmonization, 2021).

Data integration

Data integration is a process by which multiple data sets can be combined or linked from various data sources to provide a more unified picture of the data, solve problems, and make informed decisions (Jung & Chung, 2021). Data integration can also be viewed from two perspectives: merging different data sets available from various data sources; and preparing the merged data in a usable insight and information in the proper form to make it available to different end-users (Jung & Chung, 2021; Mountasser et al., 2021).

Data integration framework

A data integration framework is a real or conceptual structure (Tungpantong et al., 2021) intended to support or guide the building of a data transport process that could expand the structure into a data integration solution (Mountasser et al., 2021). A data integration framework is also a logical structure to provide a more prescriptive and comprehensive representation of the data consolidation process independent of the tools and methods used in the consolidation process (Mudadu & Zerlotini, 2020).

Digital authentication

Digital authentication is ascertaining the validity of one or more identifiers used to claim a digital identity (Grassi et al., 2017). In addition, digital authentication is a formalized process of verification that involves tests of one or more identification attributes provided by an entity to ascertain their correctness using the required level of assurance (ISO, 2019a).

Digital identity

A digital identity is an identification used or maintained in cyberspace by an individual, organization, or electronic device, with a unique representation of the subject, in the context of a digital service, which can be involved in an online transaction (Grassi et al., 2017). In addition, a digital identity is a digitally stored set of attributes related to an entity (ISO, 2019b) and used by computer systems to represent an external agent, which may be a person, organization, device, or application. Furthermore, digital identity is a property used for transactional purposes for essential online services and can be stolen and criminally damaged (Sullivan, 2011).

National digital identity

The national digital identity (NDID) is a system that enables individuals and entities to prove who they claim they are to digitally access critical information or services provided by a

nation or government (Arunwatanamongkol et al., 2021). In addition, NDID is a computer system maintained and used by governments on the premises to protect national and enhance personal security, monitor the illegal residents' activities, and guard against terrorism and illegal immigration (Lim et al., 2009).

Summary of Chapter One

This chapter started by introducing a general overview in which the primary constructs of the study were explained. Following the introduction, the statement of the problem section explained the need for addressing the data integration issue. The purpose of the study details the value of proposing a standardized data harmonization framework for addressing the data integration issue detailed in the problem statement. Furthermore, the theoretical framework section detailed the theoretical concepts used to construct the theoretical framework of this study.

The research design section outlined a general overview of different research methods and the rationale for choosing a multiple-case qualitative research design for this study. In addition, the chapter outlined the two research questions that this study is intended to address. Moreover, the chapter highlighted the results and significance of this study. This study's contribution fills the knowledge gap in computer science by advancing the field of data integration and big data processing (Hilbert, 2016) and benefits academics and practitioners in identifying approaches needed to create a well-harmonized big data integration solution. Finally, the chapter concluded by defining key terms used in this study.

Chapter 2: Literature Review

Introduction

This research aims to explore, understand, and establish a standardized data harmonization framework used for integrating a national identity database from heterogeneous digital identity data sources. The standardized data harmonization framework proposed by this research might facilitate effective communication among digital identity solution stakeholders and accelerate the implementation and adoption of the NDID database for a country (Festus et al., 2017; Hilbert, 2016). Furthermore, apart from filling the void of a standardized digital identity data harmonization framework, the output of this study defined big data analytics components. The big data analytics components can be used to create hindsight, insights, and visualizations of the digital identity data movement between sources and destinations to make data-driven decisions.

This chapter outlines the review of related literature on NDID databases (i.e., their current state, implementation, and economic significance), heterogeneous data integration approaches, big data, and big data analytics. The literature search strategy was based on the research topic, problem statement, purpose statement, and the research questions using explicit and alternative search keywords and phrases.

Search Strategy

For exhaustively assessing all potential and available recent studies, the literature review was performed by searching studies from multiple conference proceedings, bibliographic databases, forward-citation searches of seminal articles, hand searches of popular journals, and Google Scholar searches of other internet sources. Evaluation and selection of each related literature were employed based on the recency of the study, objectivity, provenance, and the

value of the opinions and justifications of concluding remarks of the authors. Except for seminal articles, the year range criteria for selecting scholarly reviewed articles and journals is within the last five years (i.e., 2016 and after). The details of the search strategy, including the search parameters, search target details, databases accessed, and country-specific websites, are outlined in Appendix A.

Chapter Organization

This chapter is organized to reveal concepts, theories, practices, design, and implementation details associated with national digital identity and data integration. First, the theoretical framework used for the study is outlined, where details related to information system design theories, big data concepts, and digital identification-related ISO standards and system design principles are presented. Then, a review of data integration approaches and considerations of data integration in digital identification systems are outlined.

The national digital identification section outlines details related to the types and models of digital identification implementation and the critical stakeholders in the implementation of digital identification. Next, details about the economic significance of having a national digital identification system are outlined. Then, Ethiopia's current state of digital identification is outlined with details related to the country's vital events registration and national id initiatives.

The section for implementation considerations for national digital identity outlines details related to the identity lifecycle, the technology assessment frameworks, and technologies related to credentials and authentication protocols. Then, the recent research and practices related to security and digital identity data protection details are outlined in which blockchain, digital identity, and distributed ledger technology are discussed. The chapter then concludes by presenting a summary of the literature review.

Theoretical Framework

General systems theory (Strauss, 2002; Von Bertalanffy, 1972), information systems theories (ISTs) (Halawi & McCarthy, 2006; Kautz et al., 2020), and design theory (DT) (Walls et al., 1992) are options that were assessed for use as a theoretical lens for this study.

The general system theory depicts a system as a self-regulating entity representing a relationship of parts and does not attempt to explain inter-system interactions and the interaction of external entities with the system (Buchanan, 2019; Von Bertalanffy, 1972; Von Bertalanffy, 2008). Alternatively, ISTs provide frameworks to explain companies' practices during start-up, growth, and technological transformations (Halawi & McCarthy, 2006; Kautz et al., 2020). In contrast, instead of focusing on the solution, information system DT assesses and examines information system design as a concept to enable system designers, prescribing the importance of realizing scenarios and engagement of stakeholders as early and often as possible to develop the information system (Liedtka et al., 2013; Markus et al., 2002; Walls et al., 1992).

DT is often applied to information system processes and products (Gregor & Jones, 2007; Markus et al., 2002). Consequently, compared to the general system and information system theories, DT is more suitable and easier to integrate with the theoretical concepts and standards required for digital identity and serves as a theoretical lens for creating the theoretical framework of this study, as explained subsequently. So, for exploring a standardized data harmonization framework through qualitative exploration of big data theories, ISO standards, and the *principles of the identification for sustainable development*, DT was used as a theoretical lens for this study.

As a baseline for the core focus of this study, information systems DT (Goldkuhl, 2004; Walls et al., 1992) was used to explore the standardized data harmonization framework, which is the target of this study. Therefore, the origin, development, and application of information

systems DT, the big data concepts, and the digital identification-related standards and design principles used to create the theoretical lens of this study are outlined hereafter.

Design Theory for Information Systems

Design theory (DT) is a widely recognized and adopted theory with a growing impact in many disciplines and academic communities (Hatchuel et al., 2017). Instead of focusing on the solution, DT prescribes the importance of realizing scenarios, engagement of stakeholders as early and often as possible, willingness to redefine the problem, and designing feedback into the potential solution (Liedtka et al., 2013). DT has been defined by different authors using different lenses. While Simon (1996) and Walls et al. (1992) defined DT as a prescriptive and dualist construct theory, other authors defined DT as a principle-based (Markus et al., 2002), the basis for action (Gregor & Jones, 2007), and practical (Goldkuhl, 2004).

Regarding the application of DT for information systems, Walls et al. (1992) provided the distinction between explanatory, predictive, and prescriptive theories, whereby defining DT as prescriptive theory. This differentiation of DT types is further classified by Gregor (2006) into five types of theories: (1) theory for analyzing, (2) theory for explaining, (3) theory for predicting, (4) theory for explaining and predicting, and (5) theory for design and action. Furthermore, Gregor provided an overview and detailed example of each of the five theories and depicted the interrelationships among these five types of theories (Gregor, 2006).

The highly influential definition of DT (Aier & Fischer, 2011) states that DT is a prescriptive theory based on theoretical foundations, which asserts how a design process can be carried out in a way that is both effective and feasible (Walls et al., 1992). Furthermore, Walls et al. divided DT into two major components, *design product,* and *design process.* The subcomponents of the *design product* involve meta-requirements, meta-design, kernel theories,

and testable design products (Walls et al., 1992). On the other hand, the subcomponents of the *design process* involve design methods, kernel theories, and testable design processes (Walls et al., 1992).

Gregor and Jones (2007) divided the components of DT into eight components classified into two major areas: *core components* and *additional components*, in a compatible definition and components of DT, outlined by Walls et al. (1992). The *core components* area of Gregor and Jones's DT has six components: (1) purpose and scope, (2) constructs, (3) principle of form and function, (4) artificial mutability, (5) testable propositions, and (6) justificatory knowledge. The *additional components* area has two additional components: (7) principles of implementation and (8) expository instantiation (Gregor & Jones, 2007).

In line with Gregor and Jones's (2007) definition and components of DT, information systems DT enables system designers to verify if the fundamentals of the DT are satisfied by an information system and the processes followed to develop the information system (Markus et al., 2002; Walls et al., 1992). So, for exploring and establishing a standardized data harmonization framework, the definition and components of DT by Gregor and Jones (2007) seem to be more apt. Therefore, this definition and components of DT outlined by Gregor and Jones, combined with the big data theories, ISO standards, and the principles of digital identification for sustainable development, were used as the theoretical framework of this study. The high-level outline of the framework is portrayed in Figure 1.

Overviews of the big data concepts, the ISO standards, and the principles of digital identification for sustainable development, which were considered an integral part of the theoretical framework for this study, are outlined hereafter.

Big Data Concepts

The concept of big data involves the production, collection, storage, and processing of structured and unstructured data characterized by high volume, velocity, variety, veracity, and vulnerability (Akhtar et al., 2019; Power, 2014; Tam & Van-Halderen, 2020). Big data analytics is the automated processing of big data in a cost-effective, efficient and innovative way to generate insights that enable enhanced decision making (Akhtar et al., 2019). The concepts of big data and data analytics were used in organizing and describing the varying data types from heterogeneous sources (Daniel, 2018; Guerra et al., 2019; Sazontev & Stupnikov, 2019; Yu & Wu, 2020) for the data harmonization and big data integration approaches pursued by this study. Furthermore, considering big data and big data analytics concepts as an integral part of the theoretical framework of this study was instrumental in understanding and describing the potential big data structures of the digital identity data types (Kharat & Singhal, 2017).

Big data analytics concepts allow the researcher to describe the techniques and infrastructure needed to analyze digital identity data and the associated big data to extract information and generate valuable insights (Horita et al., 2017; Pan et al., 2016). Therefore, the theoretical framework of this study utilized the characteristics of big data, that is, the high volume, velocity, variety, veracity, and vulnerability of data (Akhtar et al., 2019; Tam & Van-Halderen, 2020), to ensure adequate explanations of the digital identity data and their varying types. Furthermore, big data analytics concepts such as analysis, modeling, and interpretation of insights (Akhtar et al., 2019; Gandomi & Haider, 2015) were utilized to explain identity-related big data processing adequately. Therefore, the big data management and analytics concepts are included as an integral part of the theoretical framework of this study (see Appendix B).

The Digital Identity Related Technical Standards

International and national standards organizations and industry consortia are involved in setting relevant technical standards for implementing and executing digital identification systems. Furthermore, the International Organization for Standardization (ISO), World Wide Web Consortium (W3C), and Internet Engineering Task Force (IETF)/Internet Society are among the international standards organizations that define digital identification-related standards (Mittal, 2018). In addition, national and country-specific organizations such as the American National Standard Institute (ANSI); the US National Institute of Standards and Technology (NIST) have also developed technical standards for digital identification based on their specific needs and systems of measurement (Mittal, 2018). Furthermore, the US government-sponsored Biometric Consortium, the IEEE Biometrics Council, and the International Biometrics and Identification Association (IBIA) are among the industry consortia involved in either developing standards or promoting best practices on digital identification-related technologies (Mittal, 2018).

As a platform for achieving the 17 United Nations' sustainable development goals (SDGs), the international standards organization (ISO) has published more than 22000 international standards and related documents (ISO, 2018a; Zhao et al., 2020). In addition, through the identity for development (ID4D) initiative, the World Bank (WB) directly supports countries to achieve the UNs SDG target 16.9 (i.e., target 9 of the 16th SDG), which provides legal identity for all, including birth registration by 2030. To realize the SDG 16, the ISO published more than 155 standards (Zhao et al., 2020). Among the ISO standards published to help meet the legal identity SDG target 16.9, the biometrics, privacy and security, scalability, and data integration standards were used as part of the theoretical framework of this study.

For assessing the standardized biometric data exchange formats, enrolment requirements, and security evaluation needs, specific biometric-related ISO standards were employed as a theoretical lens (ISO, 2009a; ISO, 2014; ISO, 2018b). In addition, for accessing the general data security and privacy-related details needed for data integration needs, related data privacy ISO standards (ISO, 2011; ISO, 2015; ISO, 2018b; ISO, 2019b; ISO, 2020) were used. Furthermore, the ISO standard framework for authentication assurance (ISO, 2013) was consulted to assess the scalability of the data integration at the target identity system. Finally, as depicted in Figure 1, in addition to the ISO standards, applicable national and industry consortium standards were used for assessing the technical standards, best practices, and feasibility of digital identification-related technologies in this study.

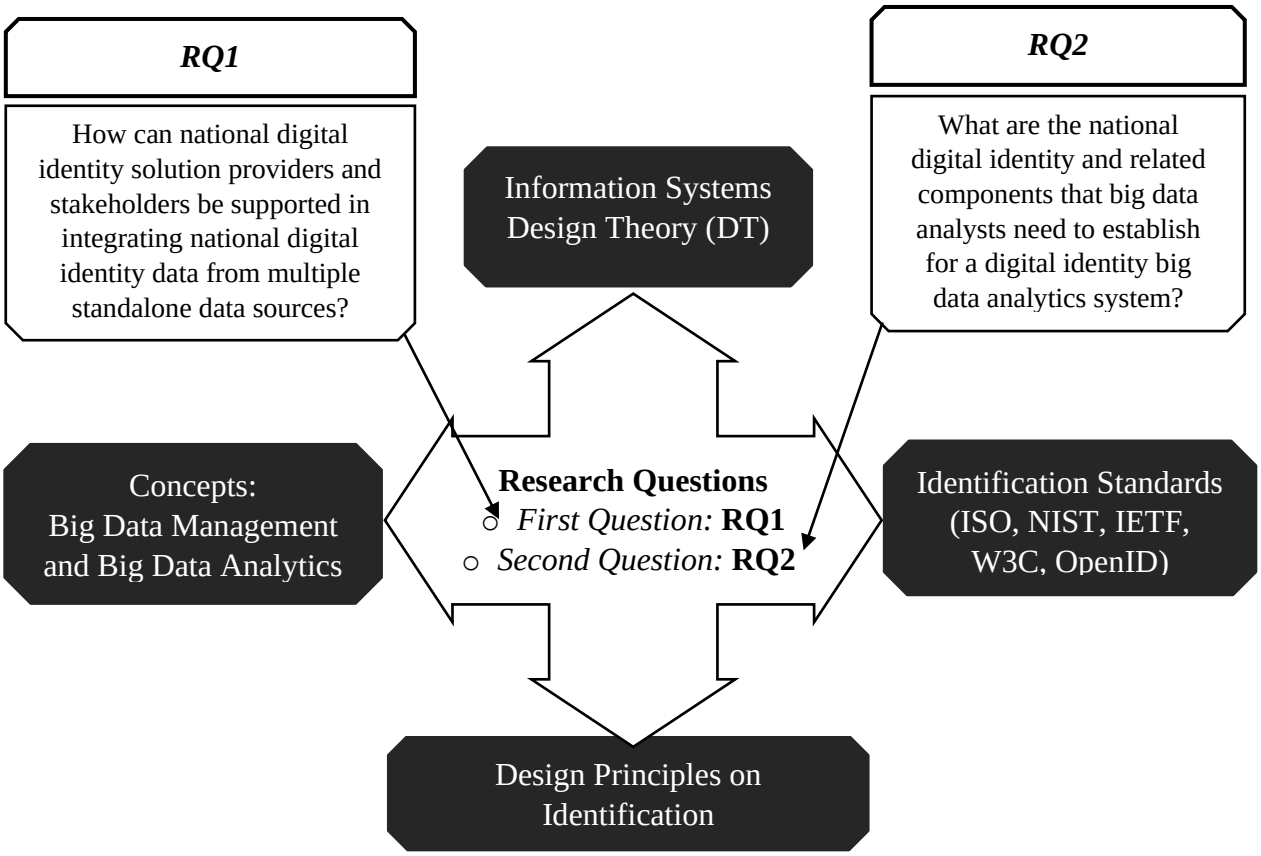
The Principles of Digital Identification

The World Bank Group published ten *principles of the identification for sustainable development*, which more than thirty organizations endorse and are categorized into three pillars: inclusion, design, and governance pillars (World Bank Group [WBG], 2021). The two principles under the inclusion pillar focus on ensuring non-discriminated universal access to legal identity for individuals and removing barriers and gaps associated with cost and technology to access legal identity. The five principles categorized under the design pillar focus on the trusted, secure, interoperable, and sustainable implementation of an identity system (WBG, 2021). Furthermore, principles under the third pillar, governance, focus on protecting personal data using a comprehensive legal and regulatory framework, establishing institutional mandates and accountabilities, and enforcing legal and trust frameworks through independent oversights (WBG, 2021).

The design pillar (WBG, 2021) was used to assess the security, interoperability, and stainability of the data integration framework explored by this study. The theoretical framework is depicted in Figure 1, with a detailed version presented in Appendix B.

Figure 1

The Theoretical Framework of the Study



Review of Data Integration Approaches

The focus of this study is on addressing the lack of a standardized data harmonization framework for integrating a centrally managed digital identity national database with disparate standalone identity data sources for the implementation of digital identity by developing countries (Festus et al., 2017; World Bank Group [WBG], 2017). Integrating data sets from

fragmented data sources into a centrally managed identity database has been a slow and daunting process for many countries struggling to implement the national identity databases (Aziz, 2017). According to a world bank survey (WBG, 2017), this problem has existed in 17 African countries. Most of these countries have highly fragmented identity systems maintained in different databases for civil registration, identification, and functional uses.

Ad hoc data harmonization processes from fragmented identity data sources into a central digital identity database played a significant role in slowing down the rollout of digital identity in the 17 African countries (Aziz, 2017; WBG, 2017). However, the existing body of literature does not adequately address the data harmonization issue that requires significant time and labor for mapping and moving data from heterogeneous sources into a centrally managed digital identity database (Festus et al., 2017). Hence, to address the lack of a standardized data harmonization framework that could be used for digital identity data sources, data integration experiences from other domains are reviewed and outlined below.

Overview of Data Integration

Data integration is a critical process to create and provide a holistic view of data stored in heterogeneous sources in different types and formats. Several data integration approaches have been proposed and implemented in different domains to address issues associated with heterogeneous sources' data integration and provide services over the integrated data. Among the approaches proposed are using distributed file systems (Sazontev & Stupnikov, 2019), using mediated schema-based queries (Festus et al., 2017), data-mapping based on standard identifiers (Faezov & Dunbrack, 2021; UniProt, 2021), using semantic queries based on linked data technique (Mihaylov et al., 2019). These approaches are closely reviewed and outlined in the following sections.

Approaches and Considerations for Data Integration

For implementing data integration solutions, defining a unified target data schema is crucial to prescribing the schema alignment needs from the different data source schemas for integration (Sazontev, 2018; Sazontev & Stupnikov, 2019). However, since schemas vary from source to source in a heterogeneous data integration setting, defining a unified target schema would be complicated. Therefore, Sazontev and Stupnikov (2019) suggested source schema profiling, scoring, and implementing probabilistic schema mappings to address this issue.

For integrating heterogeneous data sources to be queried using a mediated schema, Festus, Sunday, and Jeremiah (2017) proposed a solution that harmonizes and transforms the result from appropriate queries to the underlying data sources. Festus et al. outlined their implementation detail of a data harmonization solution that integrates data from different government databases to provide a centralized view of a web-based databank application. However, since the approach implemented by Festus et al. is focused on providing a harmonized result set for queries against a mediated schema on the fly, the approach does not address the need for integrating data from the data sources into a central digital identity database.

Several publicly available databases, associated data integration, and application platform interfaces (APIs) have been developed to support different biological information repositories for biological knowledge discovery and data-driven hypothesis generations. ConsensusPathDB is one of the public biological databases integrated from 30 other public databases and is accessed through a web interface providing a set of computational methods and data visualization tools to explore these data (Herwig et al., 2016). For integrating data into the ConsensusPathDB, entities from the source databases are mapped based on standard identifiers like UniProt, a universal protein knowledgebase identifier (UniProt, 2021), and Entrez accession numbers (Sayers et al.,

2021). Furthermore, while primary participants' data integration and interaction are mapped and grouped at different levels, using similarity, the ConsensusPathDB content gets updated from the source databases every three months (Kamburov et al., 2011).

The UniProt Knowledgebase (UniProtKB) is a protein sequences database annotated with useful information. UniProt is designed to provide users with a comprehensive, high-quality, and publicly accessible set of approximately 190 million protein sequences (UniProt, 2021). The UniProt knowledgebase combines reviewed entries added by the biocuration expert team. Unreviewed entries are annotated by automated systems and integrate, interpret, and standardize data linked to 180 other resources (Drysdale et al., 2020; Magrane & Consortium, 2010). According to Drysdale et al., for the quality of data and services it provides, the UniProt knowledgebase was recognized as an ELIXIR (the European life-sciences infrastructure for biological information) core data resource in 2017. In addition, data in the UniProt knowledgebase is available through APIs and FTP downloads in various standard formats (Garcia et al., 2019).

In response to the lack of effective integration strategies for integrating massively large-scale and heterogeneous (in types and formats) data generated by clinical tests, Mihaylov et al. (2019) proposed an intelligent data integration approach for two cancer data sets. As a result, Mihaylov et al. developed a semantically linked network-based data integration model to combine clinical and molecular data using raw data records and external knowledge sources. In their approach, Mihaylov et al. demonstrated how a linked data technique of publishing structured data, interlinked and explored via semantic queries, can be applied for patient data integration.

Established in 1971 as an archive for biological macromolecular crystal structures, the Protein Data Bank (PDB) contains about 180,000 structures solved by different methods (Faezov & Dunbrack, 2021). PDB is the single worldwide archive of biological macromolecules' structural data, serving as a repository for three-dimensional structural data of large biological molecules such as proteins and nucleic acids (wwPDB consortium, 2019). However, according to Faezov and Dunbrack (2021), authors were allowed by PDB to identify the amino acids in each protein sequence studied under different conditions in any manner they wished, causing duplicate identifiers of the same protein in the available PDB entries. As a result, reidentification, deduplication, and authors' number replacement with the proteins' UniProt sequence were proposed to fix the random PDB sequence numbers (Faezov & Dunbrack, 2021).

Digital Identification Data Integration

The distributed file systems (Sazontev & Stupnikov, 2019), mediated schema-based query (Festus et al., 2017), data-mapping based on standard identifiers (Faezov & Dunbrack, 2021; UniProt, 2021), semantic queries based on linked data technique (Mihaylov et al., 2019) are data integration approaches outlined above. However, none of these data integration approaches could address the data integration needs of the digital identification data but provide valuable lessons for creating the data integration and harmonization framework, which is the target of this study.

When designing and implementing a digital identity system, careful considerations of security, privacy, inclusion, and citizen empowerment ensure realizing significant economic values (Mir et al., 2020) from the system. In this regard, aligning the details of this study with the ten principles of identification outlined by the World Bank Group (2021) facilitates the integration of the technical and functional needs of the proposed approach. Therefore, an

overview of the national digital identification, its economic significance, implementation considerations, and security and digital identity data protection details are reviewed and outlined in the following sections.

Overview of National Digital Identification

Digital identification or electronic identity (e-ID) system is a platform of technologies, processes, and policies that enable a person to prove unambiguously and securely and assert their legal rights in a digital context (Atick, 2016). Robust, inclusive, and responsible national digital identification systems are recognized to be powerful drivers of inclusive and sustainable development (Esquivel-Korsiak & Mittal, 2018). To facilitate their public and private services and benefitting from the digital economy, countries implement different identification models (OECD, 2019) and types of identification schemes (Gelb & Clark, 2013). This section outlines the general overview of the national digital identity, the implementation approaches used in different countries, the e-ID role in different contexts, and the key stockholders for the implementation and execution of e-ID.

The Digital Identity Revolution

The proven positive impact of digital transformation on development and its financial benefits in developed and some developing countries encouraged the digital identification initiative in many countries. For example, among the developed countries, the European countries were estimated to save €50 billion by adopting digital invoicing (Billestrup & Stage, 2014). A similar digital execution of India's subsidy flow was estimated to save 1% of the country's GDP, which was equivalent to $20 billion (Gaur & Padiya, 2016; Mićić, 2017). As a result, to improve the efficiency and effectiveness of their public services, governments of both developed and developing countries are encouraged and committed to providing legal identity

for all their population, including birth registration by 2030, which is target 16.9 of the United Nation's sustainable development goals (SDGs) (Esquivel-Korsiak & Mittal, 2018).

According to Atick (2016), e-ID is a platform to enable a person to prove unambiguously and securely who they are and assert their legal rights in a digital context. Accelerated adoption of e-ID (Bughin et al., 2019) relies on the advancement of personal and professional digital skills (Kane, 2019) and the expansion of personal digital devices and big data computational platforms (Sharma et al., 2017). Furthermore, the digital transformation across businesses (Schwertner, 2017) and increased positive attitude of users towards technology such as social networks (Jahanmir & Cavadas, 2018) have a positive impact on accelerating the adoption of digital technology in both government and private organizations.

Types of Identification Schemes

Based on the breadth of the purpose they serve, identification (ID) systems are categorized into two schemes: *functional* and *foundational* identification schemes (Gelb & Clark, 2013). The ID systems that serve a specific application such as health insurance, social security, tax, or voting are *functional identities. On the other hand, foundational identities are* the ID systems developed with a universal multi-purpose application to a range of legal identities that need to attest to the identity of any individual (Gelb & Clark, 2013).

Digital Identification Systems

Several examples worldwide demonstrate robust, inclusive, and responsible national digital identification systems as powerful drivers of inclusive and sustainable development (Esquivel-Korsiak & Mittal, 2018). Digital identity systems like the e-governance of Estonia (Tammpuu & Masso, 2019), the GOV. UK Verify of the United Kingdom (Whitley, 2018), Aadhar of India (Mir et al., 2020; Sen, 2019), and Mi Argentina of Argentina (Najles, 2020)

enabled residents of the respective country to access both public and private services while enabling each country to identify their residents uniquely. In addition to country-specific residents' identification, cross-border digital identity systems interoperability is already being developed by countries like Estonia and Finland (Sullivan, 2018).

The effort to develop interoperable digital identity systems between countries has been extended to a regional scope like the EU and East Africa, involving several countries. For example, the development of the digital single market (DSM) and the establishment of a single digital identity (SDI) in the EU (Sullivan, 2018), led by Estonia, seem to inspire East African countries to develop regional mutual digital identity recognition system (Esquivel-Korsiak & Mittal, 2018). Therefore, developing a national digital identity requires considering the interoperability needs within a country's borders and beyond to enable the country to be part of regional and global mutual identity recognition systems.

Types of Digital Identity Implementation Models

To support the digital identity needs of their residents, the private sector and government services in several countries have adopted different models for implementing digital identity systems and issuing, managing and using digital identity cards (OECD, 2019). Upon exploring the relationship between the public and private sector sources and applying identities, OECD (2019) identified and outlined seven theoretical models used by different countries. Based on their coverage and the level of influence the private sector has on the digital identity models, these theoretical digital identities (DI) models can be categorized into two. The first group of theoretical DI models where the private sector has much influence is the *sector-specific DI,* and the *private DI* observed in Uruguay and Norway, respectively (OECD, 2019). The second group of theoretical DI models observed in the identity systems of the 11 out of 13 countries with

limited private sector influence are the *sector-specific DI with reusable public DI,* the *public DI,* the *shared DI,* and *the interoperable DI* (OECD, 2019).

According to OECD (2019), since the theoretical DI governance models influenced by the public sector avoid the need for operability with private systems, they are easier to manage and implement. In addition, since the private sector may already have DI solutions, reusing these private DI solutions and developing shared public and private sector applications could ensure wider adoption of DI in services of both the public and private sector.

The e-ID Roles in Different Country Contexts

With the secured use of citizens' e-ID data in public and private organizations, providing higher quality services adapted to the digital citizens' needs and expectations has become more feasible than ever (Schou & Hjelholt, 2018). Many countries have implemented the e-ID program with varying levels of success and benefited from the system enabling their citizens and local and global residents to access e-services virtually from anywhere in the World (Mhamane & Shriram, 2018; Tammpuu & Masso, 2019). However, some countries like Estonia are considered the World's most digital countries by providing 99% of public services to their citizens as e-services (Martinson, 2019). On the other hand, countries like Jamaica and Nigeria have faced issues when implementing the digital identity system (Dunn, 2020; Uzodike & Onapajo, 2019).

For the developed and early adopter nations, like Estonia, the e-ID has become an integral part of the digitized services enabling the public and private service providers to deliver effective and efficient services to residents. As a result, the state-issued e-ID is in the hands of every Estonian, irrespective of their location, the integrated services to support the e-ID, and the well-onboarded private sector for using the e-ID to obtain these services (Martinson, 2019).

Furthermore, Estonia is reported to save over 1400 years of working time and 2% of GDP annually through its digitized public services (Martinson, 2019). In addition, Estonia's digitized services implementation is so robust that the country's e-Residency program has become the first government-authenticated and operated international digital identity program for individuals who are not physical residents or citizens of Estonia (Sullivan, 2018).

Although inclusive and trusted digital identification is believed to unlock opportunities for the World's most vulnerable, the implementation of digital identification is very slow, with an estimated 1 billion people in the World lacking a legally recognized form of identification (White et al., 2019; World Bank Group [WBG], 2018). In Africa, close to 80% of the population is covered with 3G mobile service, and about 30% have access to a rapidly expanding higher-speed infrastructure such as 4G and 5G services (Choi et al., 2021). In addition to this infrastructure issue, e-ID adoption in Africa has been impacted and slowed down by the highly fragmented identity systems maintained in siloed databases for civil registration, identification, and functional uses (WBG, 2017).

Due to the level of digital transformation and economic and infrastructure contexts of countries, e-ID plays different roles in the high-, middle- and low-income countries (Atick, 2016). As Atick noted, the role of e-ID in high-income countries is motivated by convenience and e-services, and it transforms the physical and traditional identification methods into a digital system. However, in low-income countries where robust legacy IDs are typically non-existent, an e-ID is mainly used for identification purposes and plays limited or no role in e-services (Atick, 2016; Harbitz, 2016). For example, such e-ID systems with a primary purpose of identification were being developed in Bangladesh, Guinea, and Kenya (Harbitz, 2016). On the other hand, in middle-income countries like Albania, India, Moldova, and Pakistan, the

development of e-ID was serving to strengthen and progressively replace the physical ID services while supporting the emergence of e-services (Harbitz, 2016; Mir et al., 2020).

Key Stakeholders in Digital Identity

The World Bank Group (2021) acknowledged that the key stakeholders in digital identity are the individuals who are the center of the identification systems; national and local governments for providing the identification services; the private sector for playing an innovative role in developing and supplying identification systems. In addition, while international organizations could play a partnership role in setting cross-border interoperability standards, community-based nongovernmental organizations and civil societies could play vital roles in designing, implementing, and promoting identification systems (WBG, 2021). Therefore, to successfully implement an identification system in a country, the WBG recommends the coordinated and sustained efforts of the stakeholder playing their respective roles in funding, developing, overseeing, and using the identification system.

The historical individuals' concerns associated with unauthorized access to their personal information, improper data collection, and suspicion of the government's monitoring and surveillance of their activities (Lim et al., 2009) could hinder the individuals' willingness to involve in the execution of digital identity systems. So, consulting individuals and ensuring the protection of their citizens' rights are critical metrics in successfully implementing a digital identity system in a country (Dunn, 2020). Otherwise, inadequate regard for public consultation and the rights of citizens results in facing legal challenges during the collection and storage of sensitive identity-related biometric data collection, as observed in other countries like Jamaica, India, and Kenya (Dixon, 2017; Dunn, 2020; Weitzberg, 2020).

To successfully implement a digital identity system in a country, individuals' data protection and privacy rights should be in place to establish regulatory enforcement and restorative frameworks before the implementation and use of digital biometric identity data (Dixon, 2017). In addition to establishing the regulatory frameworks for the digital identity data it is recommended to clearly define the role of each stakeholder in the implementation of the digital identity systems and use of the services, using guiding principles and standards (Eaton et al., 2018). As Eaton et al. noted, clearly defined roles of digital identity solution stakeholders result in the convergence of interests and interdependency of resources ensuing in different governance solutions and facilitating cooperative actions.

Economic Significance of National Digital ID System

According to Demirgüç-Kunt et al. (2018), the primary barrier to accessing financial services cited by 26 percent of unbanked residents in low-income countries was the lack of individuals' legal documentation. As reported in a global financial inclusion and nine customer protection (FICP) survey, government-issued ID with 90 percent responses, and proof of address with 75 percent responses, were the two major types of documentation required to open an account at a commercial bank (Randall et al., 2017). Furthermore, Demirgüç-Kunt et al. reported that the absence of legal identification is the cause and effect of gender inequality in the low-income countries, with the poorest 40 percent of women staying at least 30 percent less likely to have legal identification for getting public and private services.

Digital Identity to Facilitate Financial Inclusion

According to the World Bank Group (WBG) and Global Partnership for Financial Inclusion (GPFI) (2018), the introduction of digital ID facilitates inclusive financial services making it easier to fulfill documentation requirements to open accounts for unbanked

individuals. Furthermore, digital ID enables financial sectors to remotely onboard customers using cost-effective online services that further enable them to deliver additional financial services to their customers (WBG & GPFI, 2018). In addition, governments of several countries are leveraging digital ID to substantially improve the efficiency and effectiveness of streamlined electronic cash transfers services, Government to Person (G2P) payments, and other humanitarian aid provisions (Randall et al., 2017; Reuben & Carbonari, 2017; Sen, 2019).

An individual's digital identity may be composed of biographic attributes (name, age, gender, and address) and biometric attributes (fingerprints, iris scans, and facial images) of the individual that are collected and verified to uniquely identify the individual (WBG & GPFI, 2018). According to WBG & G+PFI, there are three primary characteristics of a sound identification system that a digital ID should satisfy to facilitate both private and public services, especially for financial services: legal basis, uniqueness, and having a digital format. As WBG and GPFI explain further, for a digital ID to serve as a legal ID, it should be issued by a national or local government and is recognized as proof of legal identity following the national law. Furthermore, since the uniqueness of a digital ID maps one digital ID to only one person, it enhances a reliable view of customers' activity enabling financial service providers to offer a full range of financial services (Randall et al., 2017; WBG & GPFI, 2018).

Integration Approaches of Identification in the Financial Sector

Financial service providers (FSPs) use approaches that apply to the available identification system in the country where the service is rendered for integrating identification into the financial sector. Among the different approaches to integration are: using existing legal identifications as in Peru (Reuben & Carbonari, 2017), using identifications systems developed by private sector initiatives like the fast identity online (FIDO) alliance (FIDO, n.d.), and using

social data such as professional networks, e-commerce platforms and connected devices' sensors (IoT) for identity proofing (WBG & GPFI, 2018). Financial service providers collect specific IDs that they consider reliable and universal, and in most countries, they generally rely on existing legal IDs based on the physical exchange of documents (Cleland & Hartsink, 2019).

The use of digital identity has accelerated account opening in countries like India and Peru (Reuben & Carbonari, 2017; Sen, 2019). For example, in India, the country's universal digital ID, Aadhaar has been used to enable financial service providers to use the electronically-know-your customer (e-KYC) and customer due diligence (CDD) services to verify their customers and prevent money laundering (PML) (Srinivasan & Oreglia, 2020; WBG & GPFI, 2018). In the case of Peru, the national registry of identification and civil status (of RENIEC) of Peru has been used as a core verification database for the e-money platform, called *Modelo Peru*, for e-money transactions by millions of Peruvians (Bernal, 2017; Reuben & Carbonari, 2017). Furthermore, a new financial service platform called Billetera Movil (BiM) that used RENIEC and enabled users in Peru to conduct person-to-person (P2P) payments and top-up airtime credits has been used since 2016 (WBG & GPFI, 2018).

The Current State of Digital Identification in Ethiopia

Multiple academic studies have proposed implementing vital events registration systems to enable authorized offices to issue birth registration certificates with unique identity numbers (unique ID) to individuals in Ethiopia (Kamu, 2011; Mekonnen, 2016; Tarekegn, 2017; Worku, 2016). In addition, there are other studies focused on vital events registration systems to realize the rights of the child (Tarekegn, 2017) and for assessing the state of national digital identity in Ethiopia (Gelb & Palacios, 2016; Ministry of Information Technology [MiNT], 2020a). An overview of these academic, the World Bank-and government-initiated studies is outlined below.

Vital Events Registration in Ethiopia

Ethiopia has signed the United Nations Convention on the birth registration rights of a child (Todres, 2002, UN Human Rights Council [UNHRC], 2012). As a result, Ethiopia has enacted laws on registering vital events, including birth registration (Federal Negarit Gazeta [FNG], 2012). However, Tarekegn (2017) noted that the birth registration law is not effectively implemented in Ethiopia, and much has yet to be done to realize the child's rights in the country. Furthermore, multiple studies attempted to assess the existing state of vital registration systems in Ethiopia and proposed approaches that could be used to implement an electronic vital registration system (Kamu, 2011; Mekonnen, 2016; Worku, 2016).

Although it was carried out earlier than the official enactment of Ethiopia's vital events registration laws (FNG, 2012), the approach proposed by Kamu (2011) aimed at developing an enhanced electronic vital events registration system that could handle the registration of five vital events: birth, death, marriage, divorce, and adoption, to showcase their proposed approach, Kamu implemented a web-based vital events registration prototype system that could be used to maintain vital events records, issue associated certificates, certify vital event documents, and generate reports. Furthermore, Kamu's prototype system was implemented to support English and Amharic, Ethiopia's official and one of the eighty languages (Hudson, 2004). A related study by Mekonnen (2016) also outlined a proposed approach to implementing an integrated vital events registration system (IVERS) for Ethiopia.

Mekonnen (2016) proposed and outlined the IVERS approach for designing and implementing a federal vital event registration system to register birth, marriage, divorce, notification, death, and causes of death and certification at the national level. Mekonnen claims that the proposed IVERS system could facilitate compiling and disseminating vital events

between the federal system and IVERS integrated systems used by health service providers, care facilities, social service providers, and authorities responsible for managing vital events.

The vital events registration system-related studies (Kamu, 2011; Mekonnen, 2016) and their proposed approaches do not seem to scale to handle the vital events registration system at the national level and facilitate the integration of systems used to maintain vital events data. However, they provide pointers and approaches that could be enhanced to implement the target system, providing integrated and flexible vital events registration capabilities at Ethiopia's federal, regional, and local administrative levels. Furthermore, since implementing a successful vital events registration system is attributed to a successful implementation of national digital ID, as observed in Argentina (Najles, 2020), an integrated and flexible vital events registration system for Ethiopia could be used as a foundation for implementing the national digital ID of Ethiopia. Therefore, the following section outlines a review of the state of the national digital ID of Ethiopia.

The Current State of National Digital ID

To leverage new digital technology opportunities and address digital technology gaps, the Ethiopian Ministry of Information Technology (MiNT) created a digital strategy called digital Ethiopia 2025 (Ministry of Information Technology [MiNT], 2020a). *Digital Ethiopia 2025* is endorsed by the council of ministers (MiNT, 2020b) for implementation. In *digital Ethiopia 2025*, MiNT identified *digital Id, digital payment*, and *cybersecurity* as the key enabling systems to successfully deliver the short-term (18 months) and long-term (5 years) strategic digital technology implementation projects identified and outlined in the strategic plan (MiNT, 2020a).

According to MiNT (MiNT, 2020a), despite multiple attempts at national ID projects for over eight years, between 2010 and 2018, the government of Ethiopia has not yet firmly been

able to establish a *national ID* for Ethiopia. MiNT further noted that the current foundational and functional identification systems used by the different population segments only cover less than half of the population, and fraudulent versions of these documents are widespread. As a result, the Ministry of Peace (MoP) is setting a strategy for a National Digital ID to address issues associated with exiting foundational and functional IDs and develop a central registry to ensure uniqueness and strong credential that cannot easily be forged (MiNT, 2020a).

The primary identification systems used in Ethiopia are kebele ID, tax identification number (TIN), passports, driver's license, productive safety new program (PSNP), public sector pension (PSP), and community-based health insurance (CBHI) (Gelb, 2016; MiNT, 2020a). According to Gelb (2016), kebele ID is a highly developed and pervasive paper-based system for the personal identification of Ethiopian citizens aged 18 years and above. The kebele ID system is decentralized and administered by about 19,000 kebeles, the lowest-level administration units in Ethiopia, and is used virtually for all identification purposes and as the foundational ID to get the other functional IDs (Gelb, 2016).

According to Gelb (2016), identification using kebele ID applies to only Ethiopian citizens, while there are kebele ID entitlement issues in some border areas with pastoral populations. Furthermore, Gelb notes that separate registration and forms of identification are used for foreigners and refugees. Although the kebele ID serves as Ethiopia's de facto national ID, there is no central registry to ensure the uniqueness of the IDs, and it can easily be forged. The lack of unique support limited the usefulness of Kebele ID for a digital economy as individuals could have multiple or forged kebele IDs (Gelb, 2016; MiNT, 2020a).

Since there is a lack of coordination and communication among authorities providing identifications, multiple *unique IDs* given to individuals for different functional needs create a

challenge in introducing a national digital ID (MiNT, 2020a). As a result, there were initiatives led by the Ministry of Peace (MoP) in collaboration with MiNT and other stakeholders to introduce a *national digital ID* program to replace the kebele ID system and to facilitate the digital economy by addressing the uniqueness and security issues associated with kebele ID (MiNT, 2020a). Furthermore, in line with the World Bank's ID4D initiative and recommendations, MiNT recommended adopting the ten principles of identification for sustainable development to implement Ethiopia's national digital ID.

Implementation Considerations for National Digital Identity

The foundations for a national digital identity in a country are the tools, policies, and regulatory structures that support the implementation and execution of digital identity services in the country (OECD, 2019). Furthermore, the technology landscape outlined by the World Bank (2018) identified key areas of consideration that are required for the effective implementation of national digital identity in a country. These key considerations include defining the identity lifecycle, creating a technology assessment framework, forming credential technologies, identifying authentication technologies and protocols, and identifying technologies for risk predictive and business activity analytics (World Bank [WB], 2018).

Defining the Identity Lifecycle

A review of digital identity implementations in different countries shows various approaches to executing the critical considerations outlined by the WB (2018). A starting point of a country's digital identity life cycle is either a new implementation or extending the national population register if it exists. Countries like Argentina (Najles, 2020), Estonia (Sullivan, 2018), and India (Mobasher, 2018) had national population registers. These countries used their population registers as a source of information and starting point for implementing their digital

identity. However, due to social or cultural reasons, such national population registers do not exist or are not used as a starting point for implementing digital identity systems (OECD, 2019).

Regardless of a national public register availability in a country, the major steps in the identity lifecycle remain the same (Grassi et al., 2017; WB, 2018). The major steps in the identity lifecycle are registration or identity proofing; creating and issuing virtual or physical credentials such as digital ID cards; identity claim verification or authentication; verifying access rights or authorization; and identity management of maintenance (WB, 2018). For performing online transactions and accessing public or private services, it is critical to authenticate the user's identity (Boontaetae et al., 2018), where the foundational identity system comes into the picture. However, as noted by the WB (2018), unlike all the critical steps of the identity lifecycle, authorization (verifying access rights) is not the responsibility of the foundational identification system; instead, it is executed by the functional identification systems to verify access rights of the system's users.

Evaluating Digital Identification Technologies

Implementing and executing a digital identity system require different hardware and software technologies that should be well evaluated and assessed using well-defined criteria. Among the major hardware required for digital identity systems are the front desk registration computers and tablets, the backend servers, the biometric scanners, the card readers, and more. In addition, each hardware requires its corresponding operating systems, networking, and operation-specific software to execute the required operation on each hardware or to interact with other hardware components in the digital identity ecosystem.

A clearly defined technology assessment framework facilitates setting selection criteria metrics for hardware and software required for the digital identity system and avoiding a random

selection of products. To this end, the World Bank (2018) outlined an assessment framework with six key assessment parameters such as *maturity*, *performance*, *scalability*, *adoption*, *security*, and *affordability* for evaluating both hardware and software components. Furthermore, to represent responses in each assessment parameter, a three-point scale of the *high*, *medium*, and *low* is suggested for the framework (World Bank [WB], 2018).

Establishing Credential Technologies

The credential technologies that are utilized in the digital identity ecosystem involve hardware and software used for biometric recognition, the physical identification cards and their readers, bar codes and readers, magnetic strips and readers, and mobile devices and identity-related apps (Grassi et al., 2017; Priyasta et al., 2018; Preciozzi et al., 2020; WB, 2018). The WB categorized credential technologies into three sub-technologies: biometrics, cards, and mobile (Husni, 2016). Details for each sub-technologies and associated identity data collected and verified using the technologies are outlined in the WB (2018).

Each sub-technologies used for establishing identity credentials matures at different rates from the others (Smith, 2017; Subha, 2017) and require maturity level assessment metrics, the problems they can and cannot solve, and comparative affordability of the technology. Concerning this, detailed assessment metrics and associated supporting studies (Gelb et al., 2016) that could be used to evaluate each sub-technology are outlined by WB (2018).

Authentication Technologies and Protocols

An authentication and trust framework comprises a set of technical, business, and legal rules that participants of the operation agree to follow and achieve trust (Grassi et al., 2017; WB, 2018). A federated authentication framework comprises governance, standards, and supporting technologies that enable identity providers and the relying parties to trust and authenticate

individuals with known assurance levels across a diverse range of service providers (WB, 2018). Frameworks that support the development of authentication include Open Authorization (OAuth) 2.0, OpenID Connect (OIDC), and SAML (Naik & Jenkins, 2017; Oh & Kim, 2019). Furthermore, the Universal Authentication Framework (UAF) and the Universal Second Factor (U2F) specifications of the Fast Identity Online protocol (FIDO) are used for authentication in government applications in the US, Australia, and Germany (Noor, 2020).

Due to its immutable nature and track records in chronological order of transactions in a decentralized ledger, blockchain is an emerging technology for providing self-sovereign identity use cases in the digital identification domain (Boysen, 2021; Htet et al., 2020). Furthermore, the blockchain platform is architected with privacy by design principles and provides identity system users with a means to control and share only necessary information needed to transact with third parties (Yaga et al., 2019; Zheng et al., 2017). Furthermore, blockchain-based decentralized identity solutions enable individuals to access and control their identification anywhere in the World, potentially supporting cross-border verification of identity (Argento et al., 2021; WB, 2018). The next section outlines the details of the user blockchain-based and distributed ledger technology (DLT) for digital identity security and data protection.

Security and Digital Identity Data Protection

The national digital identity system is a critical data storage system targeted for identity theft and cybercriminal attacks. Hence, securing this critical system and protecting the data it stores are the critical measures that must be put in place to empower citizens and realize significant economic values for a country (Mir et al., 2020). Due to the need to ensure the protection and security of the digital identity systems while providing efficient identity services, digital identity technologies have evolved through stages such as centralized, federated, user-

centric, and self-sovereign identity (Boysen, 2021; Hardjono, 2019; Soltani et al., 2021). Furthermore, blockchain-based identity platforms (Boysen, 2021; Mühle et al., 2018) are also proposed to protect users' data, enabling the users to grant consent to the use of their data.

To improve government digital identity services and enable users to protect their privacy by providing consent and permission to use their data for services, Chalaemwongwan and Kurutach (2018) proposed a blockchain-based digital identity framework for Thailand. Although Chalaemwongwan and Kurutach did not provide enough detail on how the identity data is stored and protected, they showcased how users' experience could be centralized and improved using a blockchain-based single-sign-on service. However, frameworks intended to centralize segmented and distributed identity data are not recommended in the latest digital identity implementation studies (Boysen, 2021; Mühle et al., 2018) that promote a distributed architecture.

The development of blockchain-based platforms is claimed to support distributed identification systems and to provide robust security protocols (Boysen, 2021) for upholding the principle of self-sovereign identity (SSI) (Mühle et al., 2018) by preventing information from being identified, accessed, or misused. Using the right platform and protocol to ensure security lets users build trust, which is among the intangible ingredients of a reliable digital identity system (Abraham, 2020). Furthermore, as Abraham noted, while building users' trust in an information system requires years of flawless execution, a single high-profile security incident could easily result in irreversible damage eroding the users' trust.

For ensuring the security of users' digital identity, identification systems are required (Clark, 2018) to adhere to *principle 6* (securing data by design) and *principle 8* (ensuring security through a comprehensive legal framework) of the *Principles on Identification* (World Bank Group [WBG], 2021). Multiple layers of security measures are employed to secure identity

data by design. These measures are data security at rest and in transit (Hsu & Chao, 2009), using biometric recognition security features to distinguish the identity data owners, and using the identity card's physical security features (WB, 2018). Furthermore, blockchain-based and distributed ledger technology (DTL) solutions are being proposed for digital identity data security that could enable owners to control and consent to their data, the detail of which is outlined in the next sections.

Blockchain and Digital Identity

Blockchain technology and distributed ledger technologies were not built having a digital identity in perspective (WBG & GPFI, 2018). However, recent developments in blockchain-based identity solutions are showing the potential of the technologies used for customer due diligence (CDD) through self-sovereign identity (SSID) (Boysen, 2021; Hardjono, 2019). Among the most recent use cases of the blockchain technology for identity is the plan by the government of Dubai to use blockchain for verifying data such as insurance documents, health records, and passport information which are encrypted and stored on an Emirates digital ID (Kuperberg et al., 2019). In addition, blockchain-based identity and e-government solutions have been in development in Estonia since 2008 (Martinson, 2019). As a result, more advanced use cases of the blockchain technology for digital trust (e-trust) with electronic signature and seals services enable e-Residents to access and use a range of Estonian e-government and private sector services virtually from anywhere in the World (Sullivan, 2018).

Despite the progress in the technology, the feasibility of blockchain for use in SSID and CDD at scale is being disputed around the legal standing of blockchain-based applications, the standards used, and the interoperability of shared ledger systems (WBG & GPFI, 2018). In this regard, the WB (2018) outlined assessments of the blockchain technology using maturity, ease of

adoption, affordability, performance, security, and scalability metrics using a three-point (high, medium, and low) scale. Based on the WB assessment, blockchain technology has a high resilience for security, high accuracy and stable performance, and low integration for adopting the technology. Furthermore, the WBG and GPFI rated the blockchain technology's maturity, scalability, affordability metrics, and associated sub-metrics at a medium level.

Distributed Ledger Technology and Blockchain

Distributed ledger technology (DLT) allows the recording and sharing of data across multiple data stores or ledgers. When applied for financial transactions, DLT enables the network participants to record, share, and synchronize data in a distributed manner (Collomb, 2016). Blockchain-based DLT systems are an append-only chain of data blocks that can contain several transaction records. They are shared across the entire network encrypting the data contained in the blocks to make the transaction details private (Natarajan et al., 2017). Furthermore, the network participants in the blockchain-based DLT can collectively determine the validity of each block using a pre-defined algorithmic validation method (consensus mechanism) (Collomb & Sok, 2016; Natarajan et al., 2017).

There are three types of DLTs: the centralized ledger; the distributed permissionless ledger; and the distributed permissioned ledger (Natarajan et al., 2017). In the centralized ledger, all network participants synchronize their local databases with a centralized electronic ledger maintained and controlled by a trusted central party (Fung & Halaburda, 2016). In a permissioned distributed ledger, each participant of the network owns the complete and up-to-date copy of the entire ledger, and every local addition to the ledger by any node in the network is communicated to all participants of the network (Barrdear & Kumhof, 2021; Fung & Halaburda, 2016). In the case of a permissioned distributed ledger, network participants need

authentication and authorization from a central authority to access the network and make any changes to the ledger (Natarajan et al., 2017).

According to Natarajan et al. (2017), blockchain-based DLT holds the potential to expand financial inclusion by addressing barriers to access to finance, such as affordability of financial services, lack of verifiable ID systems, low payment and credit infrastructure, and lack of secured transaction frameworks. In addition, as Natarajan et al. noted, DLT could also lead to greater financial access and inclusion for underserved communities through facilitating cross-border payments and remittances, digital identity systems, asset registries, and digital currencies. However, despite these potentials of the blockchain-based DLT, the maturity of the technology to support secured financial transactions with reliable performance and at scale are still the subject of intense deliberations (Barrdear & Kumhof, 2021; Fung & Halaburda, 2016).

Summary of Chapter Two

This section outlines the summary of key concepts and topics discussed in this chapter. As a theoretical framework for this study, different theories such as general system theory (Von Bertalanffy, 1972; Von Bertalanffy, 2008), information system theories (Halawi & McCarthy, 2006; Kautz et al., 2020), and design theory (Gregor & Jones, 2007; Walls et al., 1992) are assessed. Systems theory depicts a system as self-regulating and is used to investigate a complex interaction of system elements with one another and the system they are part of (Strauss, 2002; Von Bertalanffy, 1972; Von Bertalanffy, 2008). Nevertheless, since the focus of this study is to explore and establish a standardized data harmonization framework that facilitates users-system and multi-system interactions, the definition and components of design theory (DT) by Gregor and Jones (2007) seem to be more apt. Furthermore, the amalgamation of the DT with big data theories, ISO standards related to national digital identification, and the ten principles of

identification for sustainable development (WBG, 2021) to serve as a unified theoretical framework for this study are discussed.

In line with the focus of this study, which is addressing the lack of a standardized data harmonization framework for integrating a centrally managed digital identity national database (Festus et al., 2017; World Bank Group [WBG], 2017), a review of related literature on data integration approaches is presented. In this regard, data integration approaches use: distributed file systems (Sazontev & Stupnikov, 2019), mediated schema-based queries (Festus et al., 2017), data-mapping based on standard identifiers (Faezov & Dunbrack, 2021; UniProt, 2021), semantic queries based on linked data technique (Mihaylov et al., 2019) are critically assessed. However, the existing body of knowledge related to data integration approaches could not address the data integration needs of the digital identification data but provide valuable lessons for creating the data integration and harmonization framework, which is the target of this study.

Digital identification or electronic identity (e-ID) system is a platform of technologies, processes, and policies that enable a person to prove unambiguously and securely and assert their legal rights in a digital context (Atick, 2016). Robust, inclusive, and responsible national digital identification systems are recognized to be powerful drivers of inclusive and sustainable development (Esquivel-Korsiak & Mittal, 2018). However, to successfully implement a digital identity system in a country, individuals' data protection and privacy rights should be in place to establish regulatory enforcement and restorative frameworks before the implementation and use of digital biometric identity data (Dixon, 2017). In addition, it is recommended to clearly define the role of each stakeholder in the implementation of the digital identity systems and use of the services, using guiding principles and standards (Eaton et al., 2018).

The primary barrier to accessing financial services cited by 26 percent of unbanked residents in low-income countries was the lack of individuals' legal documentation (Demirgüç-Kunt et al., 2018). In response to this, the introduction of digital ID facilitates inclusive financial services, making it easier to fulfill documentation requirements to open accounts for unbanked individuals (WBG & GPFI, 2018). Consequently, to facilitate their public and private services and benefit from the digital economy, countries have started implementing different identification models (OECD, 2019) and identification schemes (Gelb & Clark, 2013). Concerning this, there have been several academic studies (Kamu, 2011; Mekonnen, 2016; Tarekegn, 2017; Worku, 2016), government proclamations (FNG, 2012), government strategic initiatives (MiNT, 2020a), and World Bank-sponsored studies (Gelb, 2016) to enable Ethiopia benefit from the introduction of the national digital registry and digital identification.

The critical considerations for implementing a national digital identification include defining the identity lifecycle, creating a technology assessment framework, forming credential technologies, identifying authentication technologies and protocols, and identifying technologies for risk predictive and business activity analytics (WB, 2018). In addition, a national digital identity system is a critical data storage system targeted for identity theft and cybercriminal attacks. Therefore, securing this critical system and protecting the data it stores are the critical measures that must be put in place to empower citizens and realize significant economic values for a country (Mir et al., 2020). Hence, to ensure the protection and security of the digital identity systems while providing efficient identity services, digital identity technologies have evolved through stages such as centralized, federated, user-centric, and self-sovereign identity (Boysen, 2021; Hardjono, 2019; Soltani et al., 2021). Furthermore, blockchain-based identity platforms (Boysen, 2021; Mühle et al., 2018) and blockchain-based DLT (Natarajan et al., 2017)

are also proposed to protect users' data in national digital identification systems and provide the users the flexibility of providing consent on the use of their data. However, despite the potential of the blockchain and blockchain-based DLT, the maturity of the technology to support secured data transactions with reliable performance and at scale are still the subject of intense discussions (Barrdear & Kumhof, 2021; Fung & Halaburda, 2016).

Chapter 3: Research Method

The problem to be addressed by this study is the lack of a standardized data harmonization framework for integrating a centrally managed digital identity national database with disparate standalone identity data sources for the implementation of digital identity in developing countries (Festus et al., 2017; World Bank Group [WBG], 2017). Integrating data sets from multiple standalone data sources used for civil registration, identification, and functional purposes into a centrally managed national identity database have been a slow and daunting process, as observed in 17 African countries surveyed by the WBG (2017). These data integration and harmonization issues hinder digital identity solution providers from effectively implementing digital identity solutions (Aziz, 2017). As a result, countries fail to realize up to 13 percent of economic benefit from the digital identity solution (White et al., 2019).

In response to the lack of a standardized data integration framework, this qualitative exploratory study aimed to explore, understand, and establish a standardized data harmonization framework used for integrating a national identity database with disparate digital identity data sources. As a result, this chapter's qualitative research methodology and design are used to answer the research questions defined in chapter one under the *Research Questions* section.

Research Methodology and Design

The research questions do not involve dependent or independent variables for quantitative data collection and statistical analysis. Since there is no theory or hypothesis testing with numerical data collection and statistical data analysis, quantitative method research and, by extension, the mixed research methods (Barczak, 2015; Heyvaert et al., 2013) are not appropriate for this study. Therefore, in line with the statement of the problem, purpose, and the associated

research questions defined in this study, the qualitative research method is the most appropriate method for the nature of this study (Creswell & Creswell, 2018).

In line with the qualitative research design chosen for this study, a face-to-face interview of domain experts was conducted to ensure flexibility and entertain new ideas, insights, and inquiries (Creswell & Creswell, 2018). In this regard, to facilitate the invitation and voluntary participation of individuals to collect the required data for this study, an informed consent letter (see Appendix C) was used. In addition, an interview protocol (see Appendix D) was used to execute the face-to-face interview of the participants consistently. Furthermore, audio recordings of participants' responses to the semi-structured interview questions (see Appendix E) and a notepad to capture handwritten notes of their non-verbal responses (see Appendix F) were used during the face-to-face interview process.

To ensure alignment of the purpose of the study to address the research problem, the research design components of this study: the population and sample, the sampling procedure, the instrumentation, the study procedures, data analysis, and ethical assurance details are outlined in subsequent sections.

Population and Sample

A population frames similar items or events of interest for some question and restricts research by defining a group with similar characteristics for examination (Creswell & Creswell, 2018; Frey, 2018). This study's population constitutes domain experts who have had practical experience with the problem area (i.e., national civil registry and national identity). Therefore, the population to conduct this study were solution providers, stakeholders, information technology (IT) managers, and policymakers involved in designing, implementing, and integrating digital identification solutions in Addis Ababa, Ethiopia.

Addis Ababa is the capital of the federal democratic republic of Ethiopia, where the headquarters and ministry offices of the federal government are located. Among the Ethiopian federal offices that are directly or indirectly related to the national identification system and located in Addis Ababa are the immigration nationality and vital events agency (INVEA), the ministry of peace (MoP), the ministry of innovation and technology (MiNT), and the ministry of revenue (MoR). Hence, selecting respondents from these offices in Addis Ababa would provide a diverse and appropriate sampling.

Sampling is the selection of subset elements that could represent the entire population (Boddy, 2016; Dsouza, 2017). The sampling technique enables researchers to draw adequately accurate results, about the population, by studying a selected subset of the population (Boddy, 2016; Dsouza, 2017). Based on experience from prior qualitative studies, which used a sample size of up to nine respondents (Boddy, 2016; Pearsall et al., 2015; Siamagka et al., 2015), a sample size of ten respondents from the government offices (i.e., INVEA, MoP, MiNT, MoR) would be sufficient for this study.

Sampling Procedure

Laying down a well-thought sampling procedure is essential to get a suitable subset of a population for a study (Boddy, 2016). Therefore, this study utilizes a purposive sampling technique that enables the researcher to strategically select participants aligned with the research problem and questions (Gill, 2020). Furthermore, since multiple federal offices (i.e., INVEA, MoP, MiNT, MoR) are responsible for different aspects of Ethiopia's national digital identification system, the purposive sampling approach enables the researcher to gather ideas from stakeholders and domain experts of these federal offices.

The sample-set selection process involved selecting the federal and government offices working on the Ethiopian national digital identification-related systems. After the Northcentral University's (NCU's) institutional review board (IRB) approval, the researcher contacted the offices and individuals through email, phone, or in-person (Lewis, 2021). Next, the study participants were selected based on specific criteria related to their national digital identification system project involvement. In addition, to be selected into the sample set, individuals were required to have experience assuming the roles of software engineers, data analysts, database administrators, IT managers, policymakers, or stakeholders of the national digital identification system.

For applying the proposed sampling approach (Gill, 2020), the specific participants of this study were those (a) who are involved in the design and implementation of the national digital identification-related system, (b) who use the national identification-related data and closely work on the national identification system, and (c) who could contribute to the target of this study. Furthermore, additional participants were recruited using a snowball sampling technique (Gill, 2020) if original participants recommended them based on their knowledge and willingness to participate in the study.

The selection of participants was finalized based on the individuals' willingness to support the study by volunteering their time and completing the informed consent letter (see Appendix C). Upon completing the informed consent letter, schedule options were sent to the individual to participate in a face-to-face interview that gets audio-recorded and transcribed (Babchuk, 2017) for the study's data analysis. Furthermore, to ensure that enough data is collected to reach data saturation (Gill, 2020), more participants satisfying the selection criteria were considered for the interview.

Instrumentation

Exploratory qualitative research, which aims to get new insights about a particular phenomenon, is achieved through collecting experts' opinions using instruments like written artifacts, surveys, or interviews (Creswell, 2014; Milena & Dainora, 2008). For this exploratory qualitative study, open-ended semi-structured questions through face-to-face and virtual interviews enabled the researcher to collect more in-depth information (Milena & Dainora, 2008). Furthermore, the face-to-face interview method would allow capturing verbal answers to the different questions and inferring themes of the study (Cunningham et al., 2017; Fryer, 2001; Gill et al., 2008).

The interview questions captured concepts and themes from the respondents' insights, experiences, and thoughts (Tavory, 2020) to establish strategies to create a standardized data harmonization framework for integrating a national identity database. The interview set contained open-ended and semi-structured questions that encouraged the domain experts to elaborate on their insights, experience, and ideas regarding the study (Fryer, 2001; Tavory, 2020). Furthermore, each main interview question had follow-up probing questions that enabled the researcher to gain additional insight into the participant's responses to the main questions (Schultze & Avital, 2011).

The researcher developed the face-to-face interview questions (see Appendix E) according to the theoretical framework outlined to address the study's research questions. The data collection method followed a consistent interview protocol (see Appendix D), including digital voice recording and handwritten notes (see Appendix F) to capture participants' gestures, facial expressions, nods, body posture, and silence during the interview (Roberts, 2020). In addition, data labeling organized the interview responses and handwritten notes for analysis

during the data analysis process and ensured the confidentiality of personally identifiable information (PII) (Dsouza, 2017; Merriam & Tisdell, 2016).

Validity of Instruments

Validity is the core measure of a qualitative study and is characterized by dependability, credibility, transferability, and confirmability (Hayashi et al., 2019; Leung, 2015; Noble & Smith, 2015). Dependability in qualitative research characterizes whether the research process is authentic, logical, traceable, and documented (Hayashi et al., 2019). This research's dependability was ensured by carefully executing the interview protocol (see Appendix D) during data collection and clearly outlining the data analysis process of the study.

Credibility is attributed to the overall robustness of the data analysis process and depends on the quality of the data collected for a study (Cope, 2014; Noble & Smith, 2015). In addition, the credibility (also known as internal validity) and trustworthiness of a study result depend on an accurate representation of participants' ideas in the study findings (Cope, 2014; Frey, 2018). In this regard, this study's credibility was maintained and verified by involving participants in the study by letting them check the accuracy of their responses.

Transferability, also called external validity, is attributed to the flexibility and applicability of the study findings to be 'fit' for other settings or groups (Cope, 2014; Creswell & Creswell, 2018; Frey, 2018). For example, findings from exploring ideas to create a specific strategy may only be relevant to generalizing the subject or the specific phenomenon (Cope, 2014). However, the findings' capability of being "fit" or transferable can be enhanced by providing sufficient information on the participants and the research context, and the detailed procedure of the research execution (Cope, 2014). Therefore, transferability was ensured by

providing a step-by-step execution plan of the study along with the unique settings of the research process to transfer the findings to other settings.

Reliability of Instruments

Reliability is the repeatability and consistency of instruments when used to gather information in different contexts and settings (Frey, 2018; Leung, 2015). However, the dynamic contexts and nature of qualitative research make it difficult to maintain the reliability of an instrument to be used in different settings without requiring a certain margin of variability (Dsouza, 2017; Leung, 2015). In addition, while validity and reliability are better evidenced in quantitative studies, there are no universally accepted criteria to evaluate them in qualitative studies (Hayashi et al., 2019).

For assessing the reliability of a qualitative study, Golafshani (2003) suggested an approach called processual, which aims to have interconnected temporal explanations, emphasizing the need to link process analysis with temporal outcomes. Furthermore, triangulation, continual data comparison, and inclusiveness of the unusual situation are some of the suggested approaches for improving the reliability of an instrument used for qualitative research (Ospina et al., 2018). However, these approaches require repeated observations and comprehensive data collection over time, which is not realistic for a time-constrained study like this one.

Therefore, to ensure the reliability of this study, the interview instruments were reviewed, using the reputational analysis approach, by subject matter experts such as NCU's dissertation team members' (Johnson et al., 2020; O'Sullivan & Jefferson, 2020). In addition, a triangulation approach was employed using experts who are non-research participants representative of the

broader population in the data collection and analysis process (Ospina et al., 2018) to ensure the reliability of the instruments.

Study Procedures

The procedure followed for this study was guided using the interview protocol for collecting the data needed to address the research questions outlined in Chapter 1. For executing the face-to-face interview, an invitation to participate in this study was communicated through email to potential study participants in offices (i.e., INVEA, MoP, MiNT, MoR) where the target population of this study was found (Meho, 2006). The target population of this study is solution providers, stakeholders, information technology (IT) managers, and policymakers involved in the design, implementation, and integration of digital identification systems in Addis Ababa, Ethiopia.

Upon receiving the response from potential participants, satisfying the criteria, an NCU's IRB-approved pilot study involving at least two domain experts was conducted to determine the completeness of the interview protocol (Yeong et al., 2018). The pilot study was performed following the interview protocol (see Appendix D), using the interview questions (see Appendix E), and without informing the participants that they were part of a pilot study. Performing the pilot study the same way as the actual study enabled to get honest and complete responses from the participants, which were used to review and improve the completeness of the instruments (Creswell & Creswell, 2018; Yeong et al., 2018). To avoid bias in the research findings, the responses and any data collected during the pilot study were not used in the actual study.

After updating the interview instrument following the pilot study, nine participants who met the informed consent criteria were scheduled for face-to-face interviews that could take 45-60 minutes. The face-to-face interviews enabled the researcher to capture detailed responses and

nonverbal gestures and expressions (Roberts, 2020) from each participant. In addition, voice recording and handwritten notes were used to capture participants' responses, gestures, facial expressions, nods, body posture, and silence during the interview (Roberts, 2020). All the face-to-face interviews were conducted within two weeks.

Upon completing the face-to-face interview, data labeling was employed to organize the interview responses and handwritten notes for analysis and ensure the confidentiality of personally identifiable information (Dsouza, 2017; Merriam & Tisdell, 2016). In addition, the transcribed script of participants' responses was communicated to the respective respondents to accurately represent their ideas and improve the study's credibility (Cope, 2014; Frey, 2018). Finally, the data analysis process was performed on the final transcribed script of the respondents.

Data Analysis

For both the data collection and data analysis processes, the research questions defined for this study served as the central controlling units for guiding and governing the processes (Creswell & Creswell, 2018). The qualitative data analysis process involves labeling, coding, segmenting, splitting, and synthesizing the collected data to aggregate, translate and identify trends and themes (Creswell & Creswell, 2018; Nind & Vinha, 2016). Furthermore, thematic units assigned to codes for grouping coherent qualitative data establish semantic associations (Nind & Vinha, 2016). For this study, the participants' responses were segmented, split, and synthesized into six themes to address the study's research questions.

These six national digital ID-related themes were focused on data integration, security considerations, national and international standards, big data management, big data analytics, and the potential national digital ID project risks. The interview questions were categorized into

five related sections (see Appendix E) to facilitate the data analysis process of this study. In each of the five categories identified in the interview instrument, specific thematic units enabled the researcher to capture specific responses from participants. Furthermore, the specific thematic units in each category facilitated the data organization, labeling, and conceptualizing the thematic units to relational themes to detail the findings and conclusions of the study (Dsouza, 2017; Smith, 2015).

Data analysis in this qualitative study followed a series of steps such as organizing, labeling, creating thematic units to establish semantic associations, and deriving the findings and conclusions of the study (Nind & Vinha, 2016; Dsouza, 2017; Smith, 2015). For organizing and labeling the collected data for this study, the audio responses and handwritten notes captured during the interview were transcribed, cataloged, and sorted (Creswell & Creswell, 2018). For facilitating the organization and labeling of the collected data for this study by mapping the audio recording with associated handwritten notes captured during the interview process, the interview order number and the participant's pseudonym were used (see Appendix F).

In the interview instrument of this study, there were thirteen interview questions grouped into five sections (see Appendix E). Each interview question was used to collect a specific thematic unit of the study. To this end, the thirteen thematic units were used for the transcription, cataloging, and sorting of the recorded responses and associated handwritten notes (Frey, 2018; Nind & Vinha, 2016). Furthermore, the direct quotes of participants were also used to describe and reflect on the themes and group them into thematic areas (Creswell & Creswell, 2018).

To protect personally identifiable information (PII) captured during the data collection process, anonymization and replacement of PII data with pseudonyms (Dsouza, 2017; Merriam & Tisdell, 2016). In addition, to avoid representing the same information differently, which

would create redundancy and excessive information (Frey, 2018; Nind & Vinha, 2016), coding, labeling, and substituting related data representations were performed using identical codes, labels, or aliases. Furthermore, to ensure the accuracy of this study's findings through a consistent representation of related information, the data analysis process involved validation by verifying related changes with the corresponding participants (Cope, 2014; Frey, 2018).

Different tools and techniques are available for qualitative data analysis (Ryan, 2009; Salmona & Kaczynski, 2016). A cloud-based qualitative data analysis software called Dedoose (Dedoose, n.d.; Salmona & Kaczynski, 2016) was used for this study. Dedoose is a collaborative, web-based application that facilitates managing, combining, and analyzing mixed-method data (Dedoose, n.d.)

In addition, Dedoose facilitates filtering, coding, excerpting, and identifying patterns through interactive visualization and analytics (Dedoose, n.d; Salmona & Kaczynski, 2016). Furthermore, information such as visuals, sets of text excerpts, and descriptor data code systems can be exported from the Dedoose software to an appropriate format for using the exported information for an intended purpose (Dedoose, n.d.)

For this study, the specific data analysis processes performed using the Dedoose software (Dedoose, n.d.) involved frequency analysis of words and word groups from the transcribed audio responses of the participants (Light & Yasuhara, 2008). In addition, after removing stop words, the high-frequency words and word groups were associated with thematic units to analyze the thematic areas. Furthermore, using the Dedoose tool, patterns of the transcribed, coded, and thematically grouped data were visually analyzed to summarize the findings of this study.

The qualitative data analysis process was executed to explore and extract details for the analysis results about how a national digital ID could be implemented efficiently to benefit

individuals and institutions in terms of cost and time savings or more accessible products and services (White et al., 2019). In addition, the analysis results were explored to address the big data harmonization issue that requires significant time and labor to integrate different data sets into a central digital identity database, which is a gap in the literature (Chalaemwongwan & Kurutach, 2018; Festus et al., 2017). Therefore, upon completing the data analysis process, the results were used to appraise the study's findings, implications, recommendations, and conclusions in Chapters Four and Five.

The findings, implications, recommendations, and conclusions in Chapters Four and Five were outlined based on the analysis result. In this regard, the analysis result was used to justify how the proposed framework fills the knowledge gap in computer science by advancing the field of big data processing and integration (Hilbert, 2016). Furthermore, the implications, recommendations, and conclusions in Chapter Five provide details on the outcome of this study that benefits academics, researchers, and practitioners in identifying approaches needed to create a well-harmonized big data integration solution.

Assumptions

Assumptions are things or points in an argument that the researcher takes for granted without concrete proof or evidence based on their experience, education, belief, and culture (Abrams & Luna, 2015; Ellis & Levy, 2009). Documenting the research assumptions demonstrates that the research proposal has been thoroughly considered and helps to reduce misunderstanding and resistance to proposed research (Ellis & Levy, 2009). Therefore, it is necessary to identify and document potential assumptions in this research.

There are two assumptions that this researcher makes for this research. The first assumption is that the interview participants' responses would provide sufficient inputs to help

outline the strategies needed to establish a standardized data harmonization framework for integrating a national identity database with disparate digital identity data sources. The second assumption is that the participants of the study, involved in different aspects of the national digital identification-related systems implementation, could provide details related to big data analysts need to establish a digital identity related big data analytics system.

Delimitations

Delimitations refer to what the researcher will not do (Leedy & Ormond, 2010). Delimitations are characteristics that define the study's boundaries and constrain the scope of the research problem (Ellis & Levy, 2009; Simon, 2011). The researcher sets delimitations to make the study variable more manageable and utilize available resources in a bounded time and scope towards the study's purpose (Simon, 2011).

There are two delimitations identified for this study. The first delimitation is that domain experts participating in this study were individuals involved in developing and maintaining national digital identification-related systems in Addis Ababa, Ethiopia. The second delimitation is that participants' selection was based on the availability of domain experts in the federal offices associated with developing and maintaining digital identification-related systems.

Limitations

Limitations, which exist in every study (Leedy & Ormrod, 2010), are the potential weaknesses or problems identified and yet cannot be controlled with the study by the researcher (Leedy & Ormond, 2010; Simon, 2011). Limitations are characteristics and influences of a specific topic and design methodology that affect the study process and interpretation of the findings (Ellis & Levy, 2009; Price & Murnan, 2004). Among the limitations of this study is that the level of detail and quality of data collected from domain experts depend on the experiences

and exposure of national digital identification-related systems in the Ethiopian context. In addition, the study data collection was limited by the amount of time available as per the study program requirements. Furthermore, the time required to consider and compare qualitative data analysis methods and their corresponding results were limited by the time available, as per the study requirements.

Ethical Considerations

The fundamental principle of research ethics is to provide the risks and benefits of participating in specific research (Francis-Auton et al., 2020). In addition, among the ethical duties of the researcher is to protect participants from any virtual or physical harm during their participation or after they participated in the study (Bromley et al., 2017; Eduard et al., 2016). Therefore, a NUC's IRB approval was sought before commencing any data collection efforts to avoid ethical repercussions.

After securing the NCU's IRB approval for the commencement of data collection, an IRB-approved informed consent letter (see Appendix C) that outlines the risks and benefits of participating in the study was sent out to potential participants of this study. The consent letter clearly outlines the level of privacy and data protection and the confidentiality of the data collected for this study. Furthermore, to ensure voluntary participation in the study, the consent letter clearly states voluntary participation.

Summary of Chapter Three

This chapter outlined the research methodology proposed for this study. In line with the statement of the problem, purpose, and the associated research questions defined in this study, the qualitative research method is the most appropriate method for the nature of this study

(Creswell & Creswell, 2018). Following the selected research design details, the chapter outlined the research population, sample, and sampling procedure.

Citing experiences from prior qualitative studies (Boddy, 2016; Pearsall et al., 2015; Siamagka et al., 2015), a sample size of ten respondents from the Ethiopian federal offices (i.e., INVEA, MoP, MiNT, MoR) is argued to be sufficient for this study. For identifying the target sample, the sampling procedure outlined the recruitment process. As a result, participants were recruited using the purposive sampling approach (Gill, 2020) followed by the snowball sampling technique (Gill, 2020). The final recruitment and data collection process began after communicating with participants with an NCU's IRB-approved consent letter (see Appendix C).

The chapter also provided details about the research instrumentation, which outlined how the validity and trustworthiness (Dsouza, 2017), the findings of the study, and the reliability (Frey, 2018; Leung, 2015) of the study instrument were ensured. Furthermore, the interview protocol detail outlines the execution steps, from communicating the consent letter to transcribing the face-to-face responses and handwritten notes (Dsouza, 2017; Merriam & Tisdell, 2016; Yeong et al., 2018).

Following the study procedure, which outlined the data collection and preparation process, the chapter outlined the data analysis process. The qualitative data analysis process involves mapping techniques such as labeling and coding to translate and identify trends and themes (Nind & Vinha, 2016). For facilitating the data analysis process, a cloud-based qualitative data analysis software called Dedoose (Dedoose, n.d.; Salmona & Kaczynski, 2016) was used for this study. The Dedoose software facilitates filtering, coding, excerpting, and identifying patterns through interactive visualization and analytics (Dedoose, n.d; Salmona & Kaczynski, 2016).

The qualitative data analysis results were used to explore and extract details about how a national digital ID could be implemented efficiently to benefit individuals and institutions in terms of cost and time savings or more accessible products and services (White et al., 2019). The findings, implications, recommendations, and conclusions in Chapters Four and Five were informed based on the analysis result. Furthermore, the analysis results were used to provide details on the outcome of this study can fill the knowledge gap in the field of big data processing and integration (Hilbert, 2016) and benefit academics, researchers, and practitioners in identifying approaches needed to create a well-harmonized big data integration solution.

Details on the assumptions, delimitations, and limitations (Abrams & Luna, 2015; Ellis & Levy, 2009; Simon, 2011) of this study are also outlined in this chapter. Furthermore, details on the ethical considerations in this study, which provide the risks and benefits of participating in the specific research (Francis-Auton et al., 2020), are outlined. NUC's IRB approval was secured to avoid ethical issues before commencing any data collection efforts.

Chapter 4: Findings

This qualitative exploratory study aimed to explore, understand, and establish a standardized data harmonization framework to integrate a national identity database with disparate digital identification data sources. Using the qualitative research methodology and design outlined in chapter three, the results of this research study are discussed, and a comprehensive assessment of the findings is outlined in this chapter. The two research questions stated in chapter one were used to guide the assessment of the findings of this study.

For collecting the required data regarding the study's guiding research questions, semi-structured face-to-face interview questions were developed (see Appendix E), and face-to-face interviews were conducted using an interview protocol (see Appendix D). The study participants were contacted through email and social media and selected based on relevant eligibility criteria created for the study. The face-to-face and virtual interviews conducted with the participants were voice-recorded and transcribed into text for labeling and themes analysis. The details of the collected data are outlined and analyzed in this chapter.

This chapter is outlined into three main sections: trustworthiness of the data, results, evaluation of the findings, and summary. The *section on the trustworthiness of data* outlines the study processes' dependability, credibility, transferability, and confirmability (Hayashi et al., 2019; Leung, 2015; Noble & Smith, 2015) to ensure the trustworthiness of the final study data. The *results* section describes the findings and analysis of nine semi-structured interviews conducted for this study. Under the results section, the following two guiding research questions are presented with reports and comprehensive discussions on the thematic areas corresponding to each research question.

RQ1. How can national digital identity solution providers and stakeholders be supported in integrating national digital identity data from multiple standalone data sources?

RQ2. What are the national digital identity and related components that big data analysts need to establish for a digital identity big data analytics system?

The thematic areas corresponding to these guiding research questions and reports are presented in subsections. In this regard, six thematic areas for RQ1 and four thematic areas for RQ2 are identified, and a comprehensive description of the participants' experiences is provided for each of the thematic areas. In addition, the *evaluation of the findings* section outlines the assessment of the themes and resulting findings with the current literature. Finally, the *summary* section outlines the key points in the chapter.

Trustworthiness of the Data

The data for this qualitative research is collected using open-ended semi-structured questions through face-to-face and virtual interviews. The interviews captured concepts and themes from the respondents' insights, experiences, and thoughts (Tavory, 2020) to establish strategies to create a standardized data harmonization framework for integrating a national identity database. Furthermore, using face-to-face interviews, more in-depth information and verbal answers to the different questions and themes of the study (Cunningham et al., 2017; Fryer, 2001; Gill et al., 2008) were collected.

The research participants were credible and demonstrated in-depth knowledge of the subject under study. The face-to-face interviews with eight of the nine participants took place at the participants' place of work, for which written permission was secured to recruit the participants and perform the interviews. The ninth participant, with whom a Zoom interview was conducted, was recruited using a snowball technique after receiving independent

recommendations from two of the eight participants. Furthermore, besides the participants' experience confirmed during their informed consent, their publicly available professional network profile and endorsements provide them with more credibility.

All the face-to-face and remote interviews were recorded and transcribed into texts. A transparent description of the processes and protocols used and the steps involved in collecting the data, capturing the interview recordings, the associated transcriptions, and notes for this research were created to ensure the data's trustworthiness. Trustworthiness is a core metric characterized by the research process's dependability, credibility, transferability, and confirmability (Hayashi et al., 2019; Leung, 2015; Noble & Smith, 2015). The subsequent sections outline each of these characteristics in the context of this research.

Dependability

Dependability in qualitative research characterizes whether the research process is authentic, logical, traceable, and documented (Hayashi et al., 2019). This research's dependability is ensured by carefully executing the interview using a traceable interview protocol (see Appendix D). In addition, the interview responses of the face-to-face interviews with each participant are documented, recorded, and transcribed to ensure the authenticity of the data. Furthermore, to achieve dependability and reuse the raw data for other research, the interview recordings and the transcribed documents were saved so that if anyone desired to check or use the raw data, maintaining the confidentiality of the data. Similarly, the data analysis processes of this study are clearly outlined.

Credibility

The credibility (i.e., internal validity) and trustworthiness of a study result depend on an accurate representation of participants' ideas in the study findings (Cope, 2014; Frey, 2018). An

email with the transcription of their responses was sent to each study participant. The researcher received confirmation and feedback emails that ensured each transcription represented the respective participant's response. Furthermore, to ensure the overall robustness of the data analysis process, which is characterized by the credibility and quality of the data collected for a study (Cope, 2014; Noble & Smith, 2015), the data collection process for this study was made as clean and authentic as the researcher could.

Transferability

Transferability, also called external validity, is attributed to the flexibility and applicability of the study findings to be fit for other settings or groups (Cope, 2014; Creswell & Creswell, 2018; Frey, 2018). Due to the small sample size (nine participants) of this study, the applicability of this study's findings to be fit for other settings or groups may be limited. However, the step-by-step execution processes used for the study and the collected data helped offer strong evidence about data integration challenges in other settings.

Confirmability

The interview protocol (see Appendix D) is documented to ensure this study's data confirmability. In addition, the collected data were checked and rechecked throughout the interview process using interview questions (see Appendix E). In this regard, the interview and follow-up questions were documented to ensure the repeatability of the data collection process that ensures the study results are repeatable by other researchers. Furthermore, a concerted effort was put to ensure the repeatability of the results by clearly detailing the categories of themes and coding schemes and identifying the codes and patterns for analyses.

Results

The recruitment email and social media messages were shared with Ethiopia's appropriate National ID program office. In response to the recruitment messages, about fifteen individuals responded. As a result, the final nine individuals who satisfied the eligibility criteria to participate in this study were selected and scheduled for the semi-structured interview of the study.

The interview with each of these selected participants was conducted after getting their respective consent. The participants worked on national digital identity-related projects in Ethiopia and Finland. The participants' demographic details (gender, educational level, years of experience in IT, and national digital id) are shown in Table 1 below.

Table 1

Participants Demographics

Participant ID	Educational Level	*Years of Experience*		*Role in the National ID*
		In IT	*In National ID*	
1	Bachelor	10	5	Senior Engineering Lead
2	Bachelor	8	4	Senior Database Admin
3	Bachelor	13	4	Senior Software Engineer
4	Bachelor	16	5	IT Manager
5	Masters	15	7	IT Manager
6	Bachelor	13	4	R&D Director
7	Bachelor	12	4	IT Director
8	Masters	15	8	Technical Director
9	Masters	13	8	Chief Technology Officer

Each study participant was informed, in their consent letter, that the estimated duration of the interview would be between 45 and 60 minutes. However, there was no attempt to make the

interview durations longer or shorter when interviews last in less than 45 minutes or more than 60 minutes, respectively. As a result, participants were encouraged to provide detailed responses and explanations as they felt relevant to the interview questions. Therefore, with the shortest and the most extended interview sessions taking 28 and 91 minutes, respectively, the average duration of all the interview sessions was 45 minutes.

The use of participants' official working language during the interview was considered essential to enable the participants to provide enough context in their responses and, to collect relevant data for this study without any language barrier. As a result, Amharic, the official Ethiopian language, was chosen to facilitate communication during the recruitment and face-to-face interview with participants in Ethiopia. In line with this, IRB approved the use of certified Amharic translations of the recruitment letter and consent letter (see Appendix C) during the interview process.

While Amharic was used for eight of the nine participants' interviews, the interview with the ninth participant was in English. However, English was used for transcribing all the recorded interview responses. Therefore, the accuracy of the English transcription was affirmed by receiving a confirmation response email from the respective participants.

The data analysis process of this study followed a series of steps to organize, label, code, segment, split, and synthesize the transcribed responses and interview notes to aggregate, translate and identify trends and themes (Creswell & Creswell, 2018; Nind & Vinha, 2016). For organizing and labeling the study's data, the audio responses and handwritten notes captured during the interview were transcribed, cataloged, and sorted (Creswell & Creswell, 2018). Furthermore, thematic units were assigned to codes for categorizing coherent qualitative data to establish semantic associations (Nind & Vinha, 2016).

Ten thematic areas were identified from the transcribed, cataloged, and sorted responses and associated handwritten notes (Frey, 2018; Nind & Vinha, 2016) captured for this study. Furthermore, the direct quotes of participants were also used to describe and reflect on the thematic units and categories (Creswell & Creswell, 2018). Therefore, the following sections and subsections outline the thematic areas and themes pertinent to the guiding research questions.

Research Question 1. **How can national digital identity solution providers and stakeholders be supported in integrating national digital identity data from multiple data sources?**

This research question guides this study in collecting and extracting relevant thematic details from participants' responses to explore, understand, and establish a data harmonization framework for integrating a national identity database from disparate digital identification data sources. Six thematic areas were identified from the participants' responses corresponding to this research's question. Each thematic unit and thematic area are described using relevant participants' experiences and quotes that provide additional context to the comprehensive description of the thematic area or unit.

The analysis and findings of the six thematic areas identified from the data collected, corresponding to *Research Question 1*, are outlined in subsequent sections.

Common Causes of the NDID Data Integration Challenges. This thematic area represents the participants' responses to the first three questions, under *section 1: national digital ID data integration* of the interview questions (see Appendix E), concerning the data sources, data source-related challenges, and data integration challenges from those data sources. The first two thematic units of this thematic area are data format and data quality. All participants who experienced data format also experienced data quality as data integration challenges. Regarding the data format and quality challenges, P3 said, "The data from different sources were different

in format and data quality and needed a reasonable time to address the quality and format issues." Data format issues during NDID data integration, four participants shared their experiences. Furthermore, extending the same experience about the two issues, P5 said

> … for integrating and migrating data from the two sources, [source-1] and [source-2], data format and data quality were not compatible with the NDID system's data format and quality requirements. Apart from the data quality issues, due to the non-standard and vendor-locked software and devices used to collect the biometric data, the biometric images were not compliant with the ISO standards adopted by the NDID system.

All participants' most common data integration challenge relates to the data and system standard. In this regard, P4 describes that unlike the ten-fingers biometric data standard set for the NDID system they were implementing, one of the systems they attempted to integrate with NDID had a two-finger biometric data standard. On the other hand, P5 described how vendor-locked systems and devices cause interoperability issues with standard format biometric data collection devices when used in systems that need to be integrated with the NDID system. In this regard, P1 said, "One aspect that should not be ignored is the potential challenges to integrating data from different data sources and the interoperability of existing hardware devices to work with the new NDID system." Table 2 shows the causes of the NDID data integration challenges.

Table 2

Common causes of the NDID data integration challenges

ID	Thematic Unit	Participants
1	Data format	P1, P2, P3, P5,
2	Data quality	P1, P2, P3, P4, P5, P8
3	Data and system standard	All participants
4	Lack of data protection guidelines	P1, P2, P3, P4, P6, P9

The fourth thematic unit that emerged as an NDID data integration challenge is the lack of data protection policies and guidelines described by six participants. In this regard, P8 described their experience with a data source, [source-3], owner organization. As per P8's description, the owner organization was unwilling to get the data source-3 integrated with NDID due to a lack of data protection and consent guidelines. Furthermore, regarding organizations that collect users' data just for their business purpose without setting a platform for getting users' consent if the users' data is to be shared with a third party, P3 explained

> … Since users give their identity data to an institution or organization, there was a challenge in getting these users' (data owners) consent before migrating their data to the NDID database. The privacy and data protection law, related policies, and guidelines can significantly address such users' consent-related challenges.

In addition, P4 described their experience with the NDID data integration challenges due to the lack of data protection guidelines. Concerning this, P4 provided their recommendation saying, "… legally; there should be a data protection and privacy law that clearly articulates the accountabilities of the data handlers, and penalties of improper data handling or related security incidents." Furthermore, P5 added a similar experience and recommendation, saying, "... data protection is key to maintaining the public trust."

Security Recommendations for NDID Data Integration. This thematic area emerged from participants' responses to the first question, under *section 2*: *security and design considerations for sustainable development* of the interview questions (see Appendix E). The themes under this thematic area included the need for privacy and data protection policy, the usage monitoring and audit trails, the physical and virtual system protection, and the clients' and users' authentication and authorization as security recommendations for NDID data integration.

Regarding the privacy and data protection policy theme, P4 said, "… there should be data protection and privacy law that clearly articulates the accountabilities of the data handles." P5 reinforced the same and said, "Privacy and data protection are key to maintain … trust on NDID". Furthermore, P5 recommended protecting the privacy and security of users' data using a biometric-enabled security system when it is accessed. In addition, P4 suggests having an organization or entity with clearly articulated responsibilities for handling users' complaints about violated data protection rights. Table 3 illustrates the security recommendations for NDID integration.

Table 3

The security recommendations for NDID data integration

ID	Thematic Unit	Participants
1	Privacy and data protection policy	P1, P2, P3, P4, P5, P6, P9
2	Usage monitoring and audit	P1, P7, P8, P9
3	Physical and virtual system protection	P1, P2, P3, P5
4	Clients’ authentication	P1, P3, P8
5	User authentication and authorization	P1, P4, P7, P8, P9

Concerning the usage monitoring and audit theme, P9 described the need for audit logs to ensure that the different actors on the data can be monitored as per defined security policies and procedures. In this regard, P9 said, "Every entry into the personal data storage system should be logged into a secured audit log system. The logged data can later be audited and monitored about who accessed the personal data and the purpose of the access." Similarly, P7 stressed the importance of logging audit trails and said, "… there should be a permanent audit log for logging audit trails of both successful and failed requests. In addition, the audit logs should include the

associated application or system insights to monitor the volumes, sources, and reasons of saucerful and failed requests."

Design and Architectural Recommendations for NDID Data Integration. This thematic area consolidates participants' responses to the second question, under section 2: *security and design considerations for sustainable development* of the interview questions (see Appendix E). Regarding the first theme, Centralized system with distributed data management, P1, P2, P5, and P9 explained their experience and described how essential it is to have a centralized NDID system with distributed and clustered data management system. Regarding a centralized NDID system, P9 described the benefits of a centralized NDID system to end-users. P9 said, "a centralized digital identity system makes the life of individuals easier and enables authorities to integrate with other organizations to exchange information and connect the data related to the same person easily."

Regarding decentralized data management, P5 explains their experience in terms of the federal government structure of Ethiopia. P5 said, "… regional governments and service providers should be enabled to access the system in real-time, which requires a centralized NDID database that is clustered and distributed to support such requests throughout the country." Similarly, P9 reinforced the importance of centralizing and decentralizing NDID data integration from a different perspective. P9 said:

> … when different organizations need to access the NDID data, they can use the centralized digital identification to fetch the decentralized data from the source and connect these data for use without storing a copy of the data unless required. Furthermore, decentralization using dedicated base registries containing basic information about specific domains (drivers' licenses, healthcare, banking services)

avoids having one big, centralized data store that contains everything that could cause a big security risk.

The other design and architectural recommendations for NDID data integration themes focus on the system's graceful failure handling, standard security frameworks, interoperability and integrations, and the use of open sources and open standards. Regarding the need for standard security frameworks for NDID data integration, P3 said, "… ensuring the security of national identity data requires both legal and technical frameworks." P2 also supported this idea, saying, "… the security and privacy of users' data should be secured using a zero-trust security architecture with encryption of data-at-rest and data-in-transit.” Table 4 presents the recommended design and architectural considerations for NDID data integration.

Table 4

The design and architectural recommendations for NDID data integration

ID	Thematic Unit	Participants
1	Centralized system with distributed data management	P1, P2, P5, P9
2	Standard security frameworks	P1, P2, P3
3	Graceful failure handling	P9
4	Interoperability and integrations standards	P1, P2, P3, P4, P5, P6, P8
5	Open source and open standards	P3, P4, P6, P7, P8, P9

Regarding fault-tolerance as a critical design consideration, P9 said, "… a critical system like the national ID should be designed so that it is fault-tolerant, and failures in the different components of the system should be gracefully handled so that the system remains operational.". Most participants recommend profound consideration of interoperability and integration standards in designing and implementing the NDID data integration system. In this regard, P5 described the cost of a design and implementation effort of an NDID data integration system

without interoperability and integration standards. P5 said, "…without clearly defined data collection, integration and interoperability standards, an NDID data integration and system implementation would be a waste of national resources."

Standards and Procedures for NDID Data Integration. This thematic area contains six themes from participants' responses to the two questions under section 3: *national digital ID standards* of the interview questions (see Appendix E). When describing the need for data governance guidelines and policies at a national level, P3 said, "… policies and guidelines can play a great role in addressing users' consent-related challenges.". P9 articulated the same idea, saying, "At a national level, policies, standards, and data governance guidelines are required." Furthermore, P4 recommends a clearly outlined and defined governance for NDID data integration and the collaboration with and among stakeholders in the NDID ecosystem.

> Regarding the need for standards around interoperability and integration, P5 said,
>
> "… these integrations and interoperability details should have standards and guidelines to facilitate the data flow between different systems. Without clearly defined data collection, integration, and interoperability standards, an NDID data integration and system implementation would waste national resources."

Regarding national policies and regulations, P9 stressed the importance of having several NDID-related policies and ensuring their proper execution. P9 said, "… there must be several … policies and controls in place, and it must be ensured that all the authorities that use the digital identity data follow these rules and regulations." On the other hand, regarding national and cross-border standards, P5 described the importance of cross-country NDID system integration and suggested using standard biometric modalities to enable it.

P9 recognized the challenge of having cross-border NDID data integration, saying, "… one big challenge in the digital identification cross-border use is a single individual could have different digital identification in two countries." As a result, P9 acknowledged the need for cross-border standards mainly developed by multiple countries. Table 5 presents the recommended national standards and procedures for NDID data integration.

Table 5

Recommended standards and procedures for NDID data integration

ID	Thematic Unit	Participants
1	Data governance guidelines	P1, P3, P4, P5, P6, P8, P7, P9
2	National policies and regulations	P1, P3, P4, P5, P6, P7, P8, P9
3	National and cross-border standards	P5, P9
4	Institutional and administrative procedures	P1, P6, P7
5	International standards	P1, P2, P3, P4, P5, P6, P8
6	International experiences	P1, P2, P4, P8, P9
7	Good ID Principles	P4, P8

Based on their experience in a multi-year NDID project that resulted in multiple initiatives, P7 describes the issue of not having institutional and administrative procedures, saying "…initiatives will not be in sync with the core objectives of having a centralized NDID system." As a result, P7 recommended a well-defined legal entity to coordinate NDID data integration initiatives. Similarly, besides having well-established national policies and regulations, P4 indicates the need for independent data regulation and governance entity, saying

> … concerning governance, the NDID system should be developed and implemented following laws, policies, standards, and guidelines. Concerning governance, to avoid

users' data handling issues, there should be a separate entity that oversees, governs, and controls the proper execution of security and data protection policies and practices.

In addition to the national standards, procedures, and governance guidelines-related themes, the international standards theme emerged in the responses of almost all the participants. The participants' responses described several specific NDID system's related ISO standards, such as biometric image standards, devices, and systems' interoperability standards. All the participants, P1, P3, P4, P5, P6, and P8, who recommended different international standards, have also recommended different data governance guidelines, national policies, and regulations.

Participants recommended adopting international experiences from countries that have already implemented NDID systems using open standards and open sources. In this regard, describing India's more than 1 million population and its ability to successfully implement an NDID system (Aadhar), and provide good coverage, P1, P2, and P8 have commonly suggested learning experiences from India. In line with this, P8 shared their similar experience, saying, "…the unique NDID number generation best practice was examined using experiences of countries like India, Nigeria, Morocco, Rwanda, Tunisia, and the Philippines."

Regarding the good ID principles developed by the World Bank's identity for development (ID4D) program, P4 and P8 recommended reviewing the good ID principles and ensuring that each of the principles is considered during the NDID data integration implementation. P4 describes how some of the national standards should be adapted from good ID principles, such as inclusiveness, saying, "... anyone living anywhere in the country (remote, mobile, urban) should be included and provided with the NDID card". Extending P4's idea, P8

suggests closely aligning the NDID system implementation with the ten principles outlined in the good ID guidelines of ID4D.

Value Propositions for NDID Data Integration Efforts. The five themes in this thematic area emerged from participants' responses to the fourth question, under *section 3: national digital ID data integration* of the interview questions (see Appendix E). The themes in this thematic area focus on how participants responded to the benefits of an NDID data integration as a value proposition to pursue functional identity data owners for the NDID data integration. All participants unanimously responded on the benefits of NDID data integration on two themes: *eliminating redundant infrastructure costs* and *providing efficient identity authentication services, which benefit service providers and consumers*.

When describing the elimination of redundant information cost as a value proposition of NDID data integration, P4 said, "It avoids the operational cost of collecting the same data while avoiding the unnecessary waste of time and resource needed to re-register residents and provide their NDID number." P1 also supported this idea, saying, "if the data is already collected at different places for different purposes, using the already collected data could save much more human and financial resources." Furthermore, P8 said, "Apart from avoiding duplication of efforts to maintain multiple systems in multiple institutions, a centralized NDID system provides a reliable service to uniquely identify and authenticate individuals from the entire population."

Regarding cost-effective benefits of NDID data integration, P4 said, "NDID system provides central authentication services for uniquely identifying public and private service users." Furthermore, P4 advised against building an identity service focused on an organization's

customers, noting the cost to build the system and afford the operational, software license, and hardware costs. Table 6 presents the value propositions for NDID data integration.

Table 6

Value propositions for NDID data integrating from disparate systems

ID	Thematic Unit	Participants
1	Eliminate redundant infrastructure costs	All participants
2	Efficient and reliable identity authentication	All participants
3	Broader population coverage	P2, P8
4	Solving system scale-up issue	P8, P9
5	Facilitating focus on core business	P4, P5, P8

According to P8's description of the benefits of NDID data integration, the three themes, such as broader population coverage, solving system scale-up issues, and facilitating focus on core business, are interrelated. In this regard, P8 said, "Collaboration of other institutions with the NDID system program enables the program to provide a reliable and scalable service, enabling them to rely on the NDID auth service and focus on their core businesses." Regarding NDID's broader population coverage for reliable identity service, P8 said,

> "… Institutions managing their identity services do not have any business reasons or financial benefits to scale their services to deal with the unknown and uncovered population. As a result, institutions realize that their identity service will be unreliable, with no way to mitigate the risks associated with the unknown population and seek a solution to address these issues."

Risks and Recommended Mitigations for NDID Data Integration. The purpose of the first question, under section 5, risk and mitigation consideration for the national digital ID project of the interview questions (see Appendix E), was to identify potential risks and recommended mitigations for NDID data integration projects. Based on participants' responses,

five themes emerged. First, all participants agreed that institutional and administrative risks should be noted in NDID data integration projects. Concerning this, P7 described the risk and associated mitigation, saying

> … the first and foremost risk that would hinder the NDID system implementation originates from the institutional arrangement created for developing and implementing the NDID system. Therefore, it is essential to create a separate legal entity that is vested and responsible for the overall activities of the NDID system implementation.

Regarding the legal and regulatory-related risks, P3 said, "since the NDID system is an important system, it should be protected by law not to be abused or exploited to put the national security and individuals at risk." Similarly, P4 described related risk and provided mitigation, saying, "… the actual implementation should be maintained by a robust organization, strong organizational leadership, and effective governance structure." P4 further explained the risk mitigation, saying, "robust organizational structure facilitates the creation of NDID ecosystem laws, policies, guidelines and standards, and effectively following up on the …execution of these laws, policies, and guidelines." Table 7 illustrates the potential risks and mitigation recommendations for NDID data integration projects.

Table 7

Potential risks and recommended mitigations for NDID data integration

ID	Thematic Unit	Participants
1	Institutional and administrative	All participants
2	Legal and regulatory	P3, P4, P6, P7, P8, P9
3	Country-specific and cross-border	P1, P6, P7, P9
4	Technical and operational	P1, P7, P8
5	Public awareness	P1, P3, P6, P8

Concerning country-specific and cross-border related risk, P9 said, “… one big challenge in the digital identification cross border use is a single individual could have different digital identification in two countries.” As a mitigation to the risk described, P9 said, “enable countries to connect cross-border digital identifications so that individuals could use their digital identity from one country to access their digital services in another country.”

Regarding country-specific risks, P7 provided a practical example related to similarities between people living in the remote border areas of Ethiopia. P7 said, "Since these border regions are remote, people may not present credible source documents to identify themselves, making it difficult to distinguish these people when providing them with registration IDs." In response to the mentioned risk, P7 suggested cost-effective mitigation. P7 said, "…collect the biographic data, without the biometric data, at the Kebele level with minimal cost, to use the official language and English on the NDID card, and collect the household data."

When describing technical and operational risks, P1 said, “... the operational aspects of the NDID system require considering the number of registration centers, training enough registration officers, and building the technical capabilities of the core team.” Concerning cost-related operational risk and mitigation, P7 added, “the operational cost of the project is for the lifetime of the system and covering this operational cost, and ways of sustaining the system should be clearly defined.

P1, P3, P6, and P8 described their experiences regarding public awareness risks and mitigation. In this regard, P3 said, “The main challenge for NDID system implementation is public awareness. So, users may think of NDID as a government central control system … and a tool for tracking people for devilish purposes.” Similarly, P1 explained public awareness as the first and foremost thing to consider, saying, “The first and foremost thing to be considered in

NDID implementation is creating awareness about the program and the system." Finally, P6 and P8 described public awareness risks and provided mitigations.

As a mitigation to the public awareness issue, P6 said, "The implementation of NDID system requires brainstorming and creating awareness about what it is, why a nation needs it, its practice in other countries, what it needs." Then, extending the public awareness risk mitigation to institutions, P8 described institutional awareness-related risks and provided mitigation. In this regard, P8 said

> … there could be institutions that have already started the implementation of functional ID systems and have already signed license agreements, which could create a parallel effort…. Therefore, to mitigate this and related issues, it is crucial to extensively create awareness about the strategic importance of the unified NDID system and its ability to provide reliable and cost-effective authentication services for all institutions.

Research Question 2. **What are the national digital identity and related components that big data analysts need to establish for a digital identity big data analytics system?**

This research question guides this study in gathering and obtaining relevant thematic details from participants' responses to identify data elements and components for establishing a national digital ID-related big data analytics system. From the semi-structured interview responses of participants, four thematic areas were identified corresponding to this research question. The thematic unit and thematic area, under this research question, are described using relevant participants' experiences and quotes that provide additional context to the comprehensive description of the thematic area or unit.

The analysis and findings of the four thematic areas noted from participants' responses to interview questions corresponding to *Research Question 2* are described in subsequent sections.

System Setup Recommendations for NDID-Related Big Data Analytics. The thematic units in this area are produced from participants' responses to the first question, under *section 4: big data management and analytics* of the interview questions (see Appendix E). For example, three participants in the first thematic unit, P3, P4, and P9, responded to strict security measures. In this regard, configuring firewall access controls and access rights are recommended in setting up an NDID-related big data analytics solution.

When describing their recommendation about security and strict access control, P9 said, "… common protection methods such as firewalls, access controls, and access rights should be carefully considered." P3 supported this idea and said, "there should be a legal framework such as policies and guidelines about how the big data analytics could access the NDID data." On the other hand, P8 described the need for a technical security solution that blocks unauthorized access to the NIDID system and sets up authorization roles at different levels of the NDID system. Extending P8's idea, P7 described the modules and layers of security, saying, "… modules and layers of the system security should be considered at the operating system, at the application software, at the core and peripheral hardware devices, at the data center, and the data transmission channels levels."

When describing the need to set up NDID trace log components for big data analytics, P8 said, "Systems accessing the NDID system are also tracked using audit logs to monitor access to resources and automated user's consent processes." Apart from setting up operational trace logs, P9 described the architecture of the trace logs to be decentralized. P9 stated, "… a decentralized system is a lot more scalable and makes the logs of each system stored in a decentralized storage making the data size a lot smaller and more secure than centralized log data storage." Table 8 illustrates the system setup recommendations for NDID-related big data analytics.

Table 8

System setup recommendation for NDID-related big data analytics

No	Thematic Unit	Participants
1	Configure firewall access controls and security modules	P3, P7, P8, P9
2	Setup operational and system trace logs for each NDID component	P8, P9
3	Maintain the required hardware and software cost	P3, P9
4	Reliable storage, network, and transmission bandwidth	P5
5	Open-source components for cost-effective analytics	P3, P4, P8

Regarding maintaining the required hardware and software, P9 said, "… it is important to identify the required infrastructure resources and the associated cost of setting up the hardware and software infrastructures." Regarding this idea, P3 commented saying, "… maintain systems by procuring different software licenses and hardware infrastructure." For maintaining a reliable network and data transmission bandwidth, P5 said, "For integrating the different identity systems and storing the stream of data from the different sources, there should be a reliable network, enough data transmission bandwidth, and reliable data storage platform." Extending the same recommendation for big data analytics, P5 further said, "For implementing big data analytics solutions, it is important to address the storage and network infrastructures."

P3, P4, and P8 recommend setting up NDID-related big data analytics solutions cost-effectively. P4 said, "NDID system development and implementation teams should use open source and open standard software." Supporting this idea, P3 described the advantage by saying, "open source and open standards to facilitate interoperability and avoid vendor lock-ins." On the other hand, P8 explained the associated flexibility of using open source for big data analytics solutions. P8 said

> … solutions for NDID system-related big data storage management make the big data management solution flexible to integrate different open source-based components that

could facilitate cost-effective analysis and generation of insights. In order to facilitate the portability of solutions, it is recommended to follow standardized analytical practices and use popular and open-source analytics frameworks with cost-effective platforms.

Recommended Components for NDID-Related Big Data Analytics. The thematic units in this area emerged in participants' responses to the second question under *section 4: big data management and analytics* of the interview questions (see Appendix E). Among the recommended components needed for NDID-related big data analytics solutions, P1 and P8 recommended flexible and scalable computing infrastructure. In this regard, P1 said, "For handling and managing big data and multi-dimensional data processing and modeling, software and hardware infrastructures are needed." P9 further elaborated on P1's recommendation commenting on the flexibility and type of infrastructure needed by saying

> … there should be a scalable and flexible computing infrastructure. Let us consider Cloud versus on-premises from this perspective. The Cloud is more scalable and easier to do big data analytics because the cloud platform makes it easier to get scalable storage and computing resources.

Big data analytics tools are among the recommended components of big data analytics components. In this regard, P1, P3, and P4 recommended available big data analytics tools. P3 provided a general comment concerning these tools, saying, "The most important aspect for big data analytics and management is creating a data warehouse and ensuring the required software tools are in place." In a more specific way, P1 promoted P3's recommendation, saying, "… for big data analytics, we need tools such as data integration tools, data visualization tools, business intelligence tools that data analysts and data engineers can use." Furthermore, for handling

specific big data types, P1 said, "… for handling and managing big data and multi-dimensional data processing and modeling, software and hardware infrastructures are needed."

Regarding more specific big data storage and processing platforms, P2, P3, P5, P6, and P9 recommended different platforms that enable big data analysts to perform NDID-related big data analytics. As part of the big data analytics components, among participants' recommended platforms for NDID-related big data analytics are Hadoop, KFKA, Apache Spark, file-based object stores, data warehouse, and Cloud. In this regard, P2 recommended considering and leveraging these technologies, and P2 said, "NDID system design should consider leveraging big data technologies such as KFKA, Apache Spark, and Hadoop ... to facilitate big data analytics and generate insights from the data."

Concerning big data warehouse and its use to generate valuable insights that can inform policymakers, P6 and P8 provided their recommendation. P6 said, "Such a big data warehouse enables the country to generate useful insights that can be used to evaluate policies in action and create new ones to address social and economic issues in the country." P8 supported the idea and further elaborated on P6's recommendation, saying

> … it would be helpful to have a big data warehouse and analytics component that extracts the NDID system usage and each NDID component's log traces and generates insights. The big data analytics generated insights could identify bottlenecks in the system infrastructure, optimize the specific component, and improve the system.

Considering its minimal upfront cost for setup and use, P9 recommended the cloud platform for big data analytics. P9 said, "In the cloud, you only pay for what you use, which means you do not need a big upfront investment, which is required for physical hardware for a

home-grown physical data center or on-premises environments." Table 9 summarizes the recommended tools and platforms for NDID-related big data analytics.

Table 9

Recommended tools and platforms for NDID-related big data analytics

No	Thematic Unit	Participants
1	Flexible and scalable computing infrastructure	P1, P9
2	Big data analytics tools	P1, P3, P4
	Data integration and analysis tools	
	Business intelligence and data visualization tools	
	Multi-dimensional data modeling and processing tools	
3	Big data storage and processing platforms	P2, P3, P5, P6, P9
	Hadoop	
	KFKA	
	Apache Spark	
	File-based object store	
	Data warehouse	
	Cloud platforms	

Recommendations for NDID-Related Big Data Storage and Management. The recommended themes in this thematic area emerged from participants' responses to the second question in *section 4: big data management and analytics* of the interview questions (see Appendix E). Four of the study participants, P1, P2, P8, and P9, recommended the use of distributed storage and management trace logs. P1 said, "…use a distributed cluster architecture to store the NDID data and logs." P2 also supported distributed storage of data and trace logs. Furthermore, P9 shared their general thought saying, "In general, I think that decentralization is a very good approach." P9 further elaborated on their comment, saying, "… a decentralized system is more scalable and stores each system's log in a decentralized storage making the data size a lot smaller and more secure than centralized log data storage."

Regarding the storage of logs and application data, P9 cited their practical experience and recommended setting strict requirements on the location and type of platform (Cloud vs. on-premises). In addition to redacting sensitive data from trace logs, P9 advised against using public cloud data centers outside a country's territory, saying, "… it should be required that any sensitive data, NDID data or otherwise, is stored and processed within the country."

P8 advised a national data center for artificial intelligence (AI) and big data analytics to generate insights, saying

> … operational activities could generate a huge amount of data which is amenable for big data analytics and generation of insights. In this regard, a national data center for artificial intelligence and big data analytics helps generate insights from the data collected about the NDID system usage and the log traces of each NDID component.

Concerning trace log aggregation and generation of insights using a big data warehouse, P5 said, "… big data warehouse created from the integrated NDID systems could provide a holistic view of individuals' experiences at different public and private service providers. Furthermore, such data can be used to create insights that help the government to create accelerated e-government services." Table 11 summarizes the recommendations for NDID-related big data storage and management.

Table 10

Recommendations for NDID-related big data storage and management

No	Thematic Unit	Participants
1	Use distributed data storage architecture for storing trace logs.	P1, P2, P8, P9
2	No sensitive personally identifiable information (PII) in trace logs	P9
3	Maintain a central (national) big data computing and AI data center	P8
4	Trace log aggregation for an end-to-end view	P5

Potential big data features from NDID-related big data analytics. The themes in this thematic area were gathered from participants' responses to all the three questions under *section 4: big data management and analytics* of the interview questions (see Appendix E). Among the features that can be leveraged from NDID-related big data analytics, respondents described scenarios, such as generating insights for real-time fraud detection and alerting (P2) and reverse search of individuals using face-matching artificial intelligence (AI) algorithms (P7).

Regarding the reverse-searching scenario during NDID data integration, P7 said, "One of the potential areas that big data analytics facilitates in the NDID system is, searching individuals using face matching artificial intelligence algorithms. This approach is useful when biometric data are not fully available in source systems." P2 described NDID big data analytics for fraud detection and alerting, saying, "… involve big data analytics for fraud detection and alerting."

Four participants, P5, P6, P7, and P8, described scenarios in which big data analytics-generated insights could be used to accelerate e-government services (P5 and P8), identify and optimize NDID system's operational bottlenecks (P6, P7, P8), and monitor system components' operational efficiency (P5, P8). Concerning accelerated public and private e-services, P8 said, "a centralized NDID system … accelerates providing scalable and reliable services for public and private institutions." Similarly, P5 supported the idea, adding technical details and saying, "… such big data warehouse created from the integrated identity systems could provide a holistic view of individuals' experiences at different public and private service providers. The data could be used to create insights that help the government to create accelerated e-government services." Table 11 summarizes the potential big data features from NDID-related big data analytics.

Table 11

Potential big data features from NDID-related big data analytics

No	Thematic Unit	Participants
1	Real-time insights for fraud detection and alerting	P2
2	Reverse-searching using face-matching AI algorithms.	P7
3	Generating insights that can be used to	P5, P6, P7, P8
	Accelerate public and private e-services	P5, P8
	Monitor system components' operational efficiency	P5, P8
	Identify operational bottlenecks and optimize the system.	P6, P7, P8
	Track end-to-end view of individuals' e-service experiences	P5, P7
	Identify hidden patterns and issues in the NDID system	P7, P8

P7 recommended maintaining a permanent log that will be used to log audit trails of both successful and failed request types and commented on how these data could be used, saying, "... monitor the volumes, sources and reasons of successful and failed requests." P6 also commented on the role of big data analytics-generated insights "to solve many national issues, using insights generated using big data analytics." Concerning NDID-related big data insights' use to identify and optimize NDID system's operational bottlenecks, P8 said,

> …a big data warehouse and analytics component extracts the NDID system usage and the log traces of each NDID component, and generate insights. The big data analytics generated insights could identify bottlenecks in the system infrastructure, optimize the specific component, and improve the system.

P5 and P7 explained how big data analytics-generated insights track end-to-end views of individuals' e-service experiences. While P7 described how insights get tracked end-to-end,

saying "… by linking the source data using the NDID registration numbers", P5 clarified the importance of these insights to improve public and private e-services, saying

> … big data warehouse created from the integrated NDID systems could provide a holistic view of individuals' experiences with different public and private service providers and create insights that help the government create accelerated e-government services.

P7 and P8 expressed the use of NDID-related big data analytics to detect hidden patterns. In this regard, P7 said, "big data analytics generates insights and patterns about individuals' transactions and events from NDID-integrated source systems." Similarly, P8 explained the idea from the NDID system's perspective, saying, "big data analytics can identify hidden patterns and system issues in the NDID system execution and help address those system issues."

Evaluation of the Findings

Various frameworks were proposed to facilitate the big data solution implementation and integration processes (Chalaemwongwan & Kurutach, 2018; Festus et al., 2017). However, approaches proposed in the existing literature do not adequately address the big data harmonization issue that requires significant time and labor to integrate different data sets into a central digital identity database.

This research explored approaches to establish a big data integration and harmonization framework to respond to the problem. The findings of this research were founded on a theoretical framework that was an amalgamation of components of design theory (DT) (Agogué & Kazakçi, 2014; Gregor & Jones, 2007) with concepts from three sources: the theories of big data and data analytics (Daniel, 2018; Guerra et al., 2019), World Bank Group (2021), and standards and frameworks defined by the international organization for standardization (ISO).

To attain this study's aim, two guiding research questions, RQ1, which was focused on NDID data integration, and RQ2, which was focused on NDID-related big data analytics, were defined. The analysis of data collected from the study participants is aligned with the gap in existing research in that the participants recognized the importance of a data integration framework to accelerate an NDID system's timely implementation. Furthermore, in line with the need for a sound data integration framework for timely implementation of the NDID system, the study participants provided valuable insights to recognize the economic benefits that greater adoption of NDID could unlock and enable countries to achieve their digital identity commitment by 2030 (Esquivel-Korsiak & Mittal, 2018; Martin, 2021, Saputro et al., 2020).

In response to RQ1, six focus areas were identified: challenges, security recommendations, design and architectural recommendations, standards and procedures, value propositions, and risks and mitigations for NDID data integration. From the NDID data integration challenges issues, in descending order of prevalence, are data and system standards, lack of data protection guidelines, data quality, and data format. From the NDID data integration security recommendations, the security measures in descending order of prevalence are privacy data protection policy, user authentication and authorization, system protection, usage monitoring, and clients' authentication. In addition, the two most recommended design and architectural considerations for NDID data integration are interoperability and integration standards and the use of open source and open standards.

The fourth focus area is recommendations of standards and procedures for NDID data integration. In this area, the three most recommended areas are developing national policies and regulations, data governance guidelines, and closely following international standards. The fifth focus area in RQ1 was value propositions for NDID data integration. In this area, participants

unanimously indicated two benefits as value propositions of NDID data integration: the elimination of redundant infrastructure costs and efficient and reliable identity authentication. Finally, the six focus areas identified the potential risks and mitigations for NDID data integration projects. Institutional, administrative, legal, and regulatory issues were identified as the most common risks, and participants advised related mitigations.

Regarding RQ2, four focus areas of NDID-related big data analytics were identified and were aligned with the existing literature (Daniel, 2018; Guerra et al., 2019). The first three areas (system setup, analytics components, data storage, and management recommendations) outlined the analysis results of respondents' responses for setting up an NDID-related big data analytics solution. The fourth and final area (features to be leveraged) presented participants' responses on the potential features and scenarios of NDID-related big data analytics.

Summary

This qualitative study aimed to establish strategies to create a data harmonization framework for integrating a national identity database based on experts' insights, experiences, and thoughts captured using a semi-structured interview. The study involved nine participants who had experiences with NDID and related systems' implementations, summarized in Table 1. Valuable data was collected using eight face-to-face and one remote semi-structured and recorded interview. Two research questions guided the analysis results of the study: RQ1 - focused on NDID data integration and RQ2-focused on NDID-related big data analytics.

In response to RQ1, six thematic areas were identified and analyzed from the interviewees' responses. The thematic areas focused on challenges, security recommendations, design, and architectural recommendations, standards and procedures, value propositions, and risks and mitigations for NDID data integration. From the NDID data integration challenges, the

most prevalent thematic units were issues associated with data and system standards, lack of data protection guidelines, and data quality. Among the NDID data integration security recommendations, most participants recommended having a privacy data protection policy.

Regarding RQ2, four thematic areas of NDID-related big data analytics were identified. The first three thematic areas (system setup, analytics components, data storage, and management recommendations) outlined the analysis results of respondents' responses for setting up an NDID-related big data analytics solution. The fourth and final thematic area (features to be leveraged) of the NDID-related big data analytics presented participants' responses on the potential features and scenarios of NDID-related big data analytics.

Finally, the evaluation section presents the significance of the research, summarizing the analysis results outlined in the results section of the chapter.

Chapter 5: Implications, Recommendations, and Conclusions

Integrating data sets from multiple standalone data sources into a centrally managed identity database has been a slow and daunting process for many countries struggling to implement the national identity databases. This problem has existed in 17 African countries surveyed by the WBG (2017), having highly fragmented identity systems maintained in different databases for civil registration, identification, and functional uses. The highly fragmented identity databases observed in most of the 17 African countries (WBG, 2017) and the ad hoc data harmonization process from such fragmented databases into a central digital identity database played a significant role in slowing down the rollout of digital identity (Aziz, 2017; WBG, 2017). Unfortunately, the existing literature does not adequately address this data harmonization issue requiring significant time and labor for digital identity data integrations (Festus et al., 2017), as observed in 17 African countries (WBG, 2017).

This qualitative exploratory study aimed to explore, understand, and establish a data harmonization framework used for integrating a national identity database with disparate digital identity data sources. The study data collected the experiences and perspectives of participants who have had years of experience designing and implementing NDID systems for a country.

A qualitative approach guided by two research questions with a qualitative design method was used to facilitate the research. Consequently, to ensure flexibility and entertaining new ideas, insights, and inquiries (Creswell & Creswell, 2018), a face-to-face interview of domain experts was conducted using a semi-structured set of interview questions. Voluntary participation of individuals was ensured using an informed consent letter (see Appendix C). Furthermore, face-to-face interview questions (see Appendix E) were used to collect the study data in a consistent

approach using an interview protocol (see Appendix D). As a result, participants of the study candidly shared their practical experiences in recorded and one-to-one interview sessions.

The study data collected from the participants was rich and helped obtain a practical and deeper understanding of the challenges of NDID data integration efforts, various recommendations that can be used during NDID realizations, and potential risks and mitigations. In addition, the collected study data is used to establish a framework that might help digital identity solution developers and technology leaders effectively communicate the integration details and accelerate the implementation of NDID databases for countries. Furthermore, the study data was used to define big data analytics characteristics that can create hindsight, insights, data visualizations for data-driven decisions, and machine learning models to detect hidden patterns in the NDID systems.

Regarding the limitations in this study, there was a limitation related to the interview participants' IT industry and NDID implementation experience, predominantly from one country, Ethiopia. Although most participants have diversified IT industry experience in different domains, their practical NDID and related experiences were from the National ID program, vital events registration agency (VERA), and the regional VERA (RVERA), which were all in Ethiopia. There were also limitations related to participants' responses that could be misinterpreted by environmental, anxiety, bias, or stress-related factors (Ellis & Levy, 2009; Price & Murnan, 2004). Furthermore, the limited ability to generalize the results of data collected from the small sample size (nine participants) is another limitation. In addition, bias could easily be introduced in the study, which constitutes limitations since the primary investigator designed and conducted the interviews, interpreted the interview responses from Amharic to English, and performed the analysis.

Chapter 5 is outlined in three main sections: the implications and findings of this study, recommendations for practice, and recommendations for future research. The implications section outlines and reinforces the study's results regarding the two guiding research questions of the study and explains areas where the results are aligned with or deviated from other related studies. The recommendations for practice section outlines how the results from this study can be applied in data integration practices in general and NDID-related data integration. Furthermore, the recommendations for practice section highlights the practical applications of NDID-related big data analytics. The recommendations for future research section provides this study's logical next steps and recommendations on how the study could be extended. Finally, the conclusion section summarizes the chapter's main points and the research study.

Implications

This section outlines and reinforces the study's results based on the two guiding research questions of the study, covering each research question independently, and explains areas where the findings of this study are aligned with or deviated from other related studies.

Research Question 1. How can national digital identity solution providers and stakeholders be supported in integrating national digital identity data from multiple data sources?

The objective of this research question was to explore, understand, and establish a data harmonization framework for integrating a national identity database from disparate digital identity data sources. Concerning this research question, six thematic areas were identified from the research participants' responses, and the analysis and findings of these thematic areas are outlined in Chapter 4. Therefore, the implications of the analysis results, their alignment with, and deviation from related studies are outlined in subsequent sections.

Common Data Integration Challenges. This study suggests that the most common NDID data integration challenges are data format and quality issues. In addition, data format and data quality issues coexisted as data integration challenges. As a result of the data format and data quality issues, NDID data integration projects encountered data standard and compatibility issues that needed considerable time to resolve. This claim is aligned with the existing body of literature that integrating data sets from fragmented data sources into a centrally managed identity database has been a slow and daunting process for many countries struggling to implement the national identity databases (Aziz, 2017).

Although the existing literature does not adequately address the data integration issue that needs significant time and labor for mapping and moving data from heterogeneous sources (Festus et al., 2017), several approaches were proposed. Among the approaches proposed are using distributed file systems (Sazontev & Stupnikov, 2019), using mediated schema-based queries (Festus et al., 2017), data-mapping based on standard identifiers (Faezov & Dunbrack, 2021; UniProt, 2021), using semantic queries based on linked data technique (Mihaylov et al., 2019). In addition, to address the common data interpretation issues identified in this study, defining a unified target data schema and schema alignment from the different data source schemas for integration (Sazontev, 2018; Sazontev & Stupnikov, 2019) could be considered.

This study observed that data formats (schema) vary from source to source in a heterogeneous data integration setting; source schema profiling, scoring, and implementing probabilistic schema suggested by Sazontev and Stupnikov (2019) could be used for defining a unified target schema. Furthermore, given the sheer volume of the NDID data size, the data format and quality issues need an automated solution, unlike the manual data integration and handling of data format issues through manual reviews as observed and proposed in some

literature (Drysdale et al., 2020; Magrane & Consortium, 2010). On the other hand, ISO standards serve as theoretical lenses to assess the standardized data exchange formats, enrolment requirements, and security evaluation needs (ISO, 2009a; ISO, 2014; ISO, 2018b).

Security Recommendations for Data Integration. The security recommendations for NDID data integration included the need for a privacy and data protection policy, the usage monitoring and audit trails, the physical and virtual system protection, and the clients' and users' authentication and authorization. Well-defined data protection and privacy laws that delineate the accountabilities of the data handlers are noted to be the key to maintaining users' trust in the NDID system. In addition to data protection policies and guidelines, the results of this study suggest technical solutions such as the use of a biometric-enabled security system when users' data and the NDID systems are accessed.

Concerning the security recommendations, the findings of this study are aligned with the data privacy ISO standards defined for data integration needs (ISO, 2011; ISO, 2015; ISO, 2018b; ISO, 2019b; ISO, 2020). In addition, the privacy, data protection, and security guideline needs identified in this study are aligned with a proposed authentication and trust framework comprising a set of technical and legal rules to achieve trust between stakeholders (Grassi et al., 2017; WB, 2018). Furthermore, the findings of this study regarding the security of users' digital identity data align with principle 6 (securing data by design) and principle 8 (ensuring security through a comprehensive legal framework) of the Principles on Identification (Clark, 2018; World Bank Group [WBG], 2021).

Design and Architectural Recommendations. Regarding the design and architectural consideration, the result of this study suggests how essential it is to have a centralized NDID system with distributed and clustered data management system. Furthermore, this study claims

that a centralized NDID system makes users' public and private service usage experiences easier, enabling organizations to connect a single user's data from various systems and exchange it with other organizations easily. On the other hand, the decentralized and clustered database architecture enables regionally distributed service providers to access the system in real-time through a centralized NDID system and support requests at the national level.

The design principles published by the World Bank Group (WBG, 2021) are focused on the technical implementation details. As a result, the NDID system should satisfy the principles of trust, security, interoperability, and operational sustainability identity system. Furthermore, regarding centralized versus distributed design approaches observed in this study, centralized, federated, and user-centric identity services are proposed to ensure digital identity systems' efficiency, protection, and security (Boysen, 2021; Hardjono, 2019; Soltani et al., 2021). Using a centralized NDID system for its digitized public services, Estonia reported saving over 1400 years of working time and 2% of GDP annually (Martinson, 2019).

A centralized system with distributed data management design and architectural approach is also suggested to store the basic users' information at a central NDID system keeping the functional and domain-specific details in the distributed domain-specific systems. In this regard, the basic data stored in the centralized NDID system is claimed to be used to fetch and connect detailed data from one or more decentralized domain-specific systems on a need basis, which aligns with the approach prosed by Festus, Sunday, and Jeremiah (2017). Furthermore, such decentralization using dedicated base registries containing basic information about specific domains is claimed to avoid a big security risk associated with one big and centralized data store that contains everything.

Standards and Procedures. Regarding the standards and procedures for NDID data integration, the need for data governance guidelines and policies at a national level is an agreed-upon suggestion. Standards and governance guidelines for NDID data integration and the collaboration with and among stakeholders in the NDID ecosystem are claimed to facilitate integrations and interoperability details. Furthermore, data collection, integration, and interoperability standards are claimed to avoid slow and complex data integration processes and considerably save organizational resources. This claim is aligned with the emphasis given by the international standards organization (ISO), which published more than 22000 ISO standards and related documents (ISO, 2018a; Zhao et al., 2020) to realize the 17 United Nations' sustainable development goals (SDGs).

Of the 22000 ISO standards defined to realize the 17 SDGs, more than 155 of the standards are published for SDG 16 (Zhao et al., 2020), including the SDG 16.9, which focuses on realizing birth registration 2030 (Esquivel-Korsiak & Mittal, 2018). In addition, existing ISO standards are also being reused in different aspects of NDID, such as the ISO standard framework for authentication assurance (ISO, 2013), for assessing the scalability of the data integration at the target identity system.

To successfully implement a digital identity system in a country, individuals' data protection and privacy rights should be in place to establish regulatory enforcement and restorative frameworks (Dixon, 2017). This finding is aligned with the needs of appropriate policies, standards, and procedures identified in this study as vehicles for the successful implementation of the NDID system. In addition, in alignment with the findings of this study, it is recommended to clearly define the role of each stakeholder in the implementation of the NDID

systems and use of the services, using guiding principles and standards (Eaton et al., 2018; WBG, 2018).

Value Propositions for NDID Data Integration. For realizing NDID data integration, the results of this study establish promoting the benefits of NDID as a value proposition to pursue functional identity data owners for the NDID data integration. A centralized NDID data integration effort claimed to eliminate redundant infrastructure costs by saving the operational cost of collecting the same data and avoiding the unnecessary waste of time and resources needed to re-register residents and provide their NDID cards. Furthermore, apart from avoiding duplication of efforts to maintain multiple systems with a subset of the population data in multiple institutions, a centralized NDID system is claimed to provide a reliable service to uniquely identify and authenticate individuals from the entire population.

In alignment with the value proposing findings of this study, several examples worldwide demonstrate robust, inclusive, and responsible national digital identification systems as powerful drivers of inclusive and sustainable development (Esquivel-Korsiak & Mittal, 2018). In addition, a properly designed and implemented digital identity system, with careful considerations of security, privacy, inclusion, and citizen empowerment, is claimed to ensure realizing significant economic values (Mir et al., 2020) from the system. Furthermore, in alignment with the findings of this study, the digital identification initiative is encouraged by the positive impact of digital transformation on development and its financial benefits in developed countries (Billestrup & Stage, 2014) and some developing countries (Gaur & Padiya, 2016; Mićić, 2017).

A digital identification or electronic identity (e-ID) system is a platform of technologies, processes, and policies that enable a person to prove unambiguously and securely and assert their legal rights in a digital context (Atick, 2016). On the other hand, according to Demirgüç-Kunt et

al. (2018), the primary barrier to accessing financial services in low-income countries was the lack of individuals' legal documentation. In this regard, NDID systems are claimed to play a significant role in streamlining the government's electronic cash transfer services (Randall et al., 2017; Reuben & Carbonari, 2017; Sen, 2019). These claims and observations are well-aligned with the individual and organizational value proposition findings of this study.

This study noted that the broader population coverage for identity service, solving system scale-up issues, and enabling businesses to focus on their core business are the added value propositions of centralized NDID data integration. In alignment with this study's observation, Kuperberg et al. (2019) noted that Emirates digital ID cards can store multiple functional and encrypted identity data like insurance documents, health records, and passport information, enabling the ID to serve as an identity and functional ID reliably. In this regard, the results of this study revealed the importance of the NDID data integration to enable the system to provide a reliable and scalable service, enabling businesses to rely on the NDID auth service and focus on their core businesses.

Risks and Recommended Mitigations. This study identified five risks and recommended mitigations for NDID data integration projects. As per the results of this study, some institutional and administrative arrangements created for implementing the NDID system are potential sources of risks that could hinder the NDID system implementation. In this regard, the results of this study suggested a mitigation approach by creating a separate legal entity authorized to oversee and adjust the overall NDID system implementation efforts as needed.

Two of the good ID principles published by the World Bank Group (WBG), under the inclusion pillar, focus on ensuring non-discriminated universal access to legal identity for individuals and removing barriers and gaps associated with cost and technology to access legal

identity (WBG, 2021). However, this study also noted that a lack of well-defined legal and regulatory policies could expose the NDID system to be abused or exploited, thereby putting the national security and individuals at risk, which would cause individual and organizational trust issues. Therefore, despite the WBG (2021) principle for non-discriminated universal access to NDID, a lack of trust could prevent individuals and organizations from using the NDID identity services.

The cross-border NDID interoperability issue was noted as a challenge in the cross-border use of NDID, risking a single individual getting different digital identifications in two or more neighboring countries. This study suggested mitigation by enabling countries to connect cross-border NDID efforts and enable individuals to be uniquely identified and offered digital services in other countries using their country's digital identity. This mitigation is aligned with the cross-border digital identity systems' interoperability which is being developed by countries like Estonia and Finland (Sullivan, 2018). Furthermore, the effort to develop interoperable digital identity systems between countries has been extended to a regional scope like the EU (Sullivan, 2018) and East Africa (Esquivel-Korsiak & Mittal, 2018), involving several countries.

Public awareness is important for the success of NDID implementation. In this regard, while some countries are successfully providing 99% of public services to their citizens as e-services (Martinson, 2019), other countries like Jamaica and Nigeria have faced legal and operational issues when implementing the digital identity system (Dunn, 2020; Uzodike & Onapajo, 2019). For addressing such issues during the NDID system implementation, this study noted that the main challenge is a lack of public awareness. In addition, users may think of NDID as a government central control system or a tool for tracking people for devilish purposes. Furthermore, inadequate regard for public awareness and protection of citizens' rights results in

legal challenges during collecting and storing sensitive identity-related data, as observed in Jamaica, India, and Kenya (Dixon, 2017; Dunn, 2020; Weitzberg, 2020).

Research Question 2. **What are the national digital identity and related components that big data analysts need to establish for a digital identity big data analytics system?**

The objective of this research question was to identify data elements and components for establishing a national digital ID-related big data analytics system. Concerning this research question, four thematic areas were identified from the research participants' responses, and the analysis and findings of these thematic areas are outlined in Chapter 4. Therefore, the subsequent sections outline the implications of this study's analysis results and their alignment with and deviation from related studies.

Infrastructure Setup for Big Data Analytics. This study suggested setup recommendations for NDID-related big data analytics infrastructure. Among these recommendations, configuring the big data analytics platform with a strict security control was focused on the details and importance of careful configurations of common infrastructure protection methods such as firewalls, access controls, and access rights. In this regard, a legal policies and guidelines framework about how big data analytics pipelines could access the NDID data and the need for associated technical security solutions to set different authorizations to access the NIDID system was also suggested.

This study's infrastructure setup findings align with the World Bank (2018), which outlined an evaluation and assessment framework for NDID-related hardware and software components with key parameters such as maturity, performance, scalability, adoption, security, and affordability. In addition, NDID implementation and execution are facilitated by tools (OECD, 2019) that involve identification-related hardware and software technologies (Grassi et

al., 2017; Priyasta et al., 2018; Preciozzi et al., 2020). Furthermore, big data computing infrastructure needs should be carefully evaluated (Kharat & Singhal, 2017).

The study suggested identifying the required infrastructure resources and the associated cost of setting up the hardware and software infrastructures regarding the required hardware and software. Furthermore, the recommended core infrastructure components for implementing big data analytics solutions are maintaining hardware and software, a reliable network with the required data transmission bandwidth, and a reliable data storage platform. In this regard, the use of open source and open standard big data management solution components is suggested to facilitate cost-effective analysis and generation of insights, improve interoperability, and avoid vendor lock-ins.

Big Data Analytics Platform Components. This study revealed the need for a flexible and scalable computing infrastructure for handling and managing NDID-related big data and multi-dimensional data processing and modeling. Compared to the on-premises-based big data analytics platforms, the Cloud platform is suggested for its flexibility in getting scalable storage and computing resources. Furthermore, Cloud service providers' "pay for what you use" cost model and the Cloud platform’s minimal upfront cost for setup are noted as advantages over the on-premises platform for avoiding the need for a big upfront investment set up equivalent on-premises platform.

There is no current literature regarding NDID-related big data analytics. However, the intelligent data integration strategies for integrating massively large-scale and heterogeneous proposed by Mihaylov et al. (2019) and studies on the analysis, modeling, and interpretation of insights (Akhtar et al., 2019; Gandomi & Haider, 2015) have alignments with the findings of this study. This study suggested big data analytics tools for data integration & analysis, business

intelligence and data visualization, and multi-dimensional data modeling and processing. In this regard, the specific platforms suggested in this study to facilitate big data analytics and generate insights are Hadoop, KFKA, Apache Spark, file-based object stores, data warehouse, and Cloud.

Big Data Storage and Management. This study covers big data storage and management for NDID-related big data analytics. The study results indicate the technical importance of decentralized and distributed storage and management of trace logs. As per the result of this study, a decentralized system is claimed to be a lot more scalable and facilitates the NDID components to be stored in decentralized storage, making the data size a lot smaller and more secure than centralized log data storage. Furthermore, the results of this study promote setting strict security requirements on the location and type of platform (Cloud vs. on-premises), employing redaction of sensitive data from trace logs, and performing sensitive data storage and analytics within the territory of the NDID-system's country.

The findings of this study are aligned with the role of NDID, and the big data storage architecture needed for NDID-related data analytics. In this regard, NDID is claimed to facilitate online transactions and accessing public and private services (Boontaetae et al., 2018), and the distributed data storage architecture of the NDID system is promoted and recommended in the latest digital identity implementation studies (Boysen, 2021; Mühle et al., 2018). Furthermore, biometric recognition security is recommended for securing identity data at rest and in transit (Hsu & Chao, 2009).

For physically securing sensitive NDID data storage and analytics within a country, this study recommends establishing a national data center with artificial intelligence (AI) and big data analytics platforms to generate NDID system-related insights. Furthermore, the national data center with AI and big data analytics platforms is argued to facilitate the generation of

insights from the data collected about the NDID system usage and the log traces of each NDID component. Furthermore, this study claims that an integrated NDID systems data warehouse provides a holistic view of individuals' experiences with different public and private service providers and creates insights that help the government create accelerated e-government services.

Big Data Analytics for NDID System. Among the outcomes that could be leveraged from NDID-related big data analytics that this study identified are scenarios such as generating insights for real-time fraud detection and alerting. Furthermore, this study noted that a reverse search of individuals using face-matching AI algorithms are advanced big data analytics outcomes to be leveraged when biometric data are not fully available in source systems. Furthermore, big data analytics-generated insights are claimed to identify and optimize the NDID system's operational bottlenecks and monitor NDID system components' operational efficiency. In addition, big data analytics generated insights from a centralized NDID system data warehouse is claimed to provide a holistic view of individuals' experiences across connected e-services and accelerate providing scalable and reliable e-services by public and private institutions.

In alignment with existing studies, the findings of this study similarly argue that big data analytics is the automated processing of big data in a cost-effective, efficient and innovative way to generate insights that enable enhanced decision making (Akhtar et al., 2019). In addition, big data analytics facilitates the generation of useful insights based on varying data types from heterogeneous sources (Guerra et al., 2019; Sazontev & Stupnikov, 2019; Yu & Wu, 2020). Among such insights, big data analytics generated insights from audit trails data are claimed to detect hidden patterns, identify system-specific performance issues and operational bottlenecks in the NDID system infrastructure, optimize specific NDID components and improve the system.

Recommendations for Practice

This section is outlined in two sub-sections in line with the two research questions and corresponding study results. The first sub-section outlines recommendations to be considered by NDID solution providers and stakeholders for integrating NDID data from multiple data sources. The second sub-section outlines recommendations regarding the NDID-related components that big data analysts need to establish for a digital identity big data analytics system.

Recommendations from Research Question 1. The current study's practical applications concerning the first research question were used to explore, understand and establish a data harmonization framework for integrating a national identity database from standalone digital identity data sources. Concerning the first research question, the practical applications are summarized into three main areas: the NDID data integration challenges, the value propositions of the NDID system, and the set of recommendations developed because of this study.

Figure 2

NDID Data Integration Challenges, Value Propositions, and Recommendations

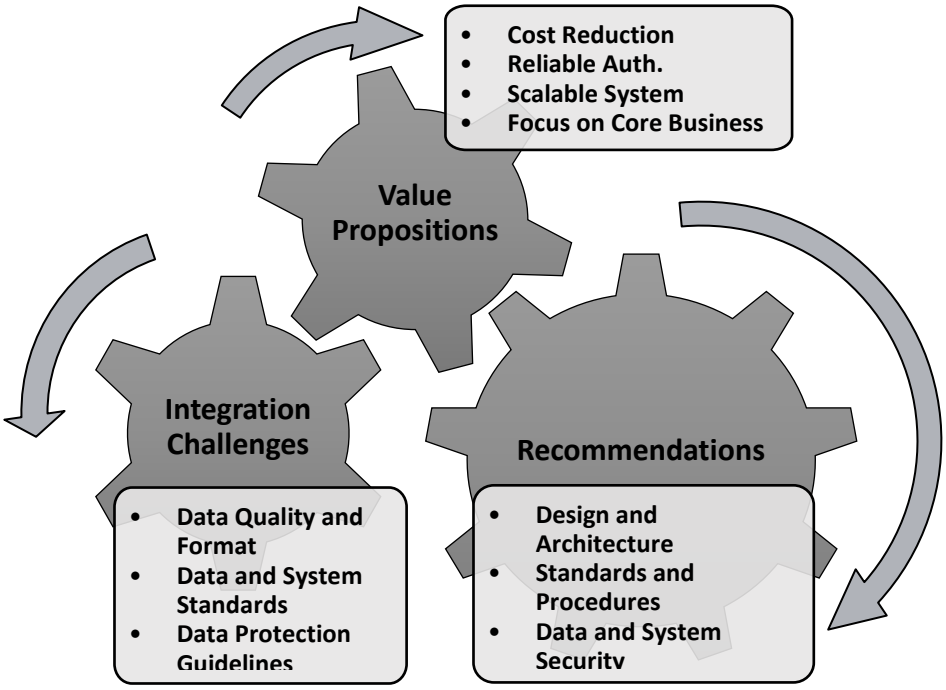


Figure 1 depicts the NDID data integration challenges that need to be overcome, the value propositions that encourage the NDID data integration implementation, and the corresponding recommendations to be considered during NDID data integration implementation. As illustrated in Figure 1, it is recommended to synchronously consider these three aspects to understand the challenges, recognize the values, and utilize the recommendations of NDID data integration implementation. Each aspect is adequately outlined in the subsequent sections.

Address the Common NDID Integration Challenges. When implementing the NDID data integration project, it is important to understand the most common NDID data integration challenges. These data integration challenges involve data format and quality issues, causing compatibility with the NDID system's data format and quality requirements. Since non-standard and vendor-locked software and devices cause data compliance issues with the ISO standards set for the NDID system, it is important to avoid using non-standard and vendor-locked systems. Furthermore, since the lack of data protection policies and guidelines was identified as an issue that caused difficulty in getting access to data sources for integration, it is recommended to define data protection policies and guidelines clearly.

Clearly defined data protection policies and guidelines encourage organizations owning source systems to integrate their system with the NDID system. Furthermore, it is recommended that organizations that collect users' data have a platform for getting the users' consent if the data collected just for the organizations' business is to be shared with a third party.

Recommendations for NDID Data Integration. The practical applications and recommendations for NDID data integration involve the design and architecture, standards and procedures, data and system security, and risk mitigation recommendations. Regarding the NDID system design, an NDID system having centralized services with distributed data management

architecture is recommended to facilitate accelerated public and private service user experiences. For example, for a federal and regional administrative set up in countries like Ethiopia, the decentralized and clustered database architecture enables regionally distributed service providers to access the system in real-time through a centralized NDID system and support requests at the national level. Furthermore, the design and architecture of a critical system like the NDID should support fault tolerance and gracefully handle failures in the different components of the system so that the system remains operational.

For facilitating the NDID system's interoperability with other systems, the use of open standards and open-source frameworks and profound consideration of interoperability and integration standards in designing and implementing the NDID data integration system is recommended. In addition, the standards and governance guidelines for NDID data integration and the collaboration with and among stakeholders in the NDID ecosystem should be clearly defined to facilitate integrations and interoperability details. Furthermore, clearly defining data collection, integration, and interoperability standards avoid slow and complex data integration processes and considerably save organizational resources.

Regarding privacy and data protection policies, it is recommended to have well-defined data protection and privacy laws that delineate the accountability of the data handlers for maintaining users' trust in the NDID system. In addition, technical solutions such as using a biometric-enabled authentication system, monitoring and audit trails, physical and virtual system protection, and NDID clients' and users' authorization are among the recommended data protection and security measures. Furthermore, in case of users complain about violated data protection rights, it is recommended to set up a support team or dedicated entity with clearly defined responsibilities for handling users' complaints.

For national and cross-border integrations, standard biometric modalities are recommended so that individuals can be recognized and provided with public or private services in a cross-border setting. Furthermore, Institutional and administrative procedures are recommended for successfully implementing NDID data integration initiatives. In this regard, establishing a legal entity to oversee NDID-specific standards is recommended to foster the proper execution of security and data protection policies and practices. Furthermore, such a legal entity is recommended to properly harmonize ISO standards with the country's national context and adopt international NDID experiences using open standards and open sources. Use of ISO standards such as biometric image standards, devices, and systems' interoperability standards.

It is recommended to create a separate legal entity authorized to oversee and adjust the overall NDID system implementation efforts to address risks associated with institutional and administrative arrangements created for implementing the NDID system. It is recommended that such a legal entity establish a robust organization, strong organizational leadership, and effective governance structure to maintain legal and regulatory issues and policies covering the NDID system and the data it stores.

It is recommended to collect the individuals' biographic data with their household details for facilitating inclusive NDID coverage in remote regions of a country cost-effectively without the complex biometric data collection process. Furthermore, for implementing cross-border NDID access, it is recommended for countries to connect cross-border NDID integration efforts.

There are operational recommendations for successfully executing the NDID system. These are organizing the number and schedules of NDID registration centers, preparing enough registration officers with sufficient data protection and handling training, building the technical capabilities of the core team, and having a support team for handling users' complaints.

Furthermore, it is recommended to define a cost-sharing model to minimize the long-term financial burden of the NDID system and share its operational cost with the NDID service clients and sustain the system for a long time.

To effectively implement and use the NDID services, it is recommended to create public and organizational awareness of the NDID system. The suggested approaches are brainstorming and creating awareness about the NDID system and its practical advantages. For creating organizational awareness, the suggested approaches outline the strategic importance of the centralized NDID system and its power to provide reliable and cost-effective authentication services for any institution. Furthermore, using the recommended value propositions outlined below facilitates persuading NDID clients.

Value Propositions for NDID Data Integration. The following are the advantages of a centralized NDID system that can be used as value propositions for NDID data integration.

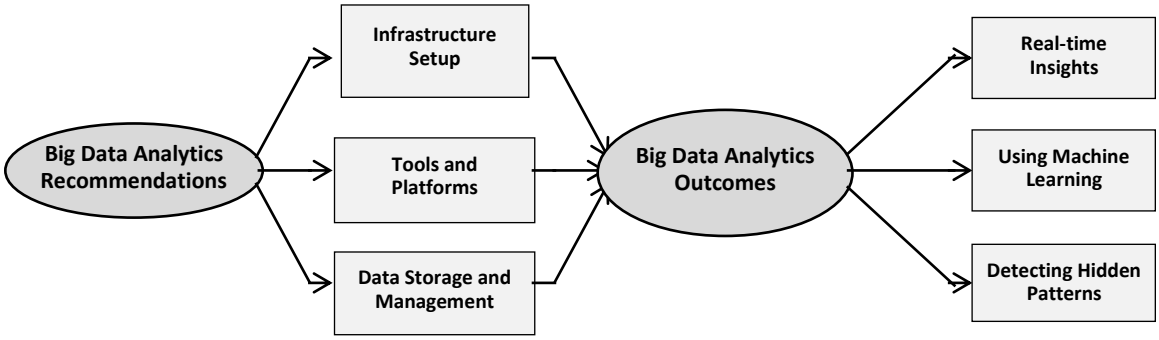
- Eliminating redundant infrastructure and operational costs of collecting the same data in multiple institutions.
- Avoiding maintenance of multiple systems with a subset of the population data in multiple institutions
- Providing services to identify and authenticate individuals from the entire population uniquely and avoiding risks with organization-specific services
- Enabling businesses to focus on their core businesses by providing a reliable and scalable authentication service.

Recommendations from Research Question 2. The practical applications of the current study concerning the second research question, which was used to identify data elements and components for establishing a national digital ID-related big data analytics system, are outlined

here. Three key recommendations related to big data storage and management, tools and platforms, and infrastructure setup are described. Furthermore, based on the recommendations, the big data analytics outcomes that can serve as the NDID-related big data analytics value propositions are also described. These NDID-related big data analytics recommendations and outcomes are shown in Figure 3.

Figure 3

NDID-Related Big Data Analytics Recommendations and Effects



Big Data Analytics Recommendations. As a result of this study, the three recommendations required for the practical application of NDID-related big data analytics are focused on the infrastructure setup, the tools and platforms, and the data storage and management needed for big data analytics.

When configuring a big data analytics platform, it is recommended to follow strict security control using standard infrastructure protection methods such as firewalls, access controls, and access rights. In this regard, the legal policies and guidelines frameworks and associated technical security solutions used to protect the NDID system data should be carefully reviewed when setting up big data analytics pipelines that have access to the NDID data. In addition, for a more secure, scalable, and managed component-specific log data size, it is recommended to set up decentralized log data storage.

It is recommended to identify the infrastructure and the associated cost of setting up the computing resources needed for big data analytics and employ open source and open standard big data analytics solutions to facilitate cost-effective analysis and generation of insights. In this regard, careful use of the Cloud platform is recommended for its flexibility in getting scalable storage and computing infrastructure for handling NDID-related non-sensitive big data analytics. In addition, more specific data storage and processing platforms that enable big data analysts to perform NDID-related big data analytics are also recommended. The specific platforms include Hadoop, KFKA, Apache Spark, file-based object stores, and data warehouses.

Regarding security, it is recommended to set strict security requirements on the location and type of platform (Cloud vs. on-premises), employing redaction of sensitive data from trace logs, and performing sensitive data storage and analytics within the territory of the NDID-system's country. This study recommends establishing a national data center with AI and big data computing platforms to generate NDID system-related insights that could help the government create accelerated e-government services.

Big Data Analytics Outcomes. The NDID-related big data analytics recommended outcomes that this study identified are generating insights for real-time fraud detection, NDID systems' issue alerting, and an advanced reverse search of individuals using face-matching algorithms. In addition, the NDID-related big data analytics is recommended for generating insights used to identify and optimize the NDID system's operational bottlenecks and monitor NDID system components' operational efficiency. Furthermore, insights with a holistic view of individuals' experiences from different e-service points are also recommended to accelerate providing scalable and reliable e-services by public and private institutions.

Big data analytics generated insights from the NDID system's audit trails data are recommended to identify system-specific issues and operational bottlenecks in the NDID system infrastructure. Such insights are recommended to help optimize NDID components and improve the NDID system's efficiency. Furthermore, NDID-related insights are recommended to detect hidden patterns and system issues and help address those issues in the NDID system.

Recommendations for Future Research

The findings of this study regarding the NDID data integration challenges are based on the participants with diversified IT industry experience in different domains, which are predominantly from one country, Ethiopia. In addition, most of the participants have their practical NDID and related experiences from the National ID program, vital events registration agency (VERA), and the regional VERA (RVERA), which were all in Ethiopia. So, to broaden the applicability of the findings of this study, and maximize the contributions to the body of knowledge, using or adopting the approaches of this study and researching a developed country's NDID system is a great next step.

The small sample size (nine participants) used for this study could limit the ability to generalize the results of this study in different settings. Furthermore, given the limited number of participants, responses from these participants could easily be misinterpreted by environmental, anxiety, bias, or stress-related factors (Ellis & Levy, 2009; Price & Murnan, 2004). In addition, bias could easily be introduced in the study, which constitutes limitations since the primary investigator is the same person who interviewed the participants of the study and interpreted their responses from Amharic to English.

This study attempted to contribute approaches to accelerated NDID data integration and setting up NDID-related big data analytics for improved data-driven decisions and NDID-based

public and private services. In this regard, accelerated adoption of e-ID (Bughin et al., 2019) relies on the advancement of personal and professional digital skills (Kane, 2019) and the expansion of personal digital devices and big data computational platforms (Sharma et al., 2017). Furthermore, the digital transformation across businesses (Schwertner, 2017) and increased positive attitude of users towards technology such as social networks (Jahanmir & Cavadas, 2018) have a positive impact on accelerating the adoption of digital technology in both government and private organizations. Therefore, researching the impact of countries' skilled ICT workforce and technological adoption rates on the rate of NDID system implementation could provide an additional dimension to maximize the body of knowledge needed for successful NDID system implementation and execution.

Big data analytics projects are complex and require skilled information technology individuals and computing infrastructure (Horita et al., 2017; Pan et al., 2016). The findings of this research regarding the NDID-related big data analytics are mainly based on the theoretical responses of the study participants, having limited hands-on experience with big data analytics projects involving sophisticated computing infrastructure for big data. Therefore, expanding and verifying the findings of this study by involving actual NDID systems that have currently implemented big data analytics would reduce the shortcomings of this study and maximize the contribution of the study.

Conclusions

This qualitative exploratory study aimed to explore, understand, and establish a data harmonization framework used for integrating a national identity database with disparate digital identity data sources. Integrating data sets from multiple standalone data sources into a centrally managed identity database has been a slow and daunting process, as observed in many countries

struggling to integrate highly fragmented functional identity systems (WBG, 2017). Developing a standardized big data integration and harmonization framework addresses the time-consuming and tedious integration of diverse data sets into a central identity database. This study has proposed an organized data integration approach and contributed toward developing a standardized data integration and harmonization framework for an NDID system.

This study's contribution to the data integration framework approach facilitates communication between the digital national identity solution stakeholders and accelerates the digital identity system's implementation for enabling countries to achieve their digital identity commitment by 2030 (Esquivel-Korsiak & Mittal, 2018; Martin, 2021, Saputro et al., 2020). In addition to the NDID data integration framework, this study was set out to identify data elements and components for establishing a national digital ID-related big data analytics system.

Although various frameworks were proposed to facilitate the big data solution implementation and integration processes (Chalaemwongwan & Kurutach, 2018; Festus et al., 2017), approaches proposed in the existing literature do not adequately address aspects of the NDID-related big data analytics. Therefore, the contributions of this study regarding the NDID-related big data analytics can facilitate the practical implementation of NDID big data analytics and fill the knowledge gap in the field of computer science by advancing the field of big data processing and integration (Hilbert, 2016). Furthermore, the contributions of this study can benefit academics, researchers, and practitioners in identifying approaches needed to create a well-harmonized big data integration solution.

References

Abraham, S. (2020). Building Trust: Lessons from Canada's Approach to Digital Identity. *ORF Issue Brief No. 367, Observer Research Foundation*. https://worldreach.com/wp-content/uploads/2020/06/ORF_IssueBrief_367_Building-Trust-Canada.pdf

Abrams, C., & Luna, A. G. (2015). The reality of the researcher: Addressing assumptions and biases. *UCLA: Center for the Study of Women*. Retrieved from https://escholarship.org/uc/item/82h2t0fs

Aier, S., & Fischer, C. (2011). Criteria of progress for information systems design theories. *Information Systems and E-Business Management, 9*(1), 133-172. https://doi.org/10.1007/s10257-010-0130-8

Agogué, M., & Kazakçi, A. (2014). 10 Years of C–K Theory: A Survey on the Academic and Industrial Impacts of a Design Theory. *Anthology of Theories & Models of Design*, 219–235. https://doi.org/10.1007/978-1-4471-6338-1_11

Akhtar, P., Frynas, J. G., Mellahi, K., & Ullah, S. (2019). Big Data-Savvy Teams' Skills, Big Data-Driven Actions and Business Performance. *British Journal of Management, 30*(2), 252–271. https://doi.org/10.1111/1467-8551.12333

Akhtar, Z., Hadid, A., Nixon, M. S., Tistarelli, M., Dugelay, J. L., & Marcel, S. (2018). Biometrics: In search of identity and security (Q & A). *IEEE MultiMedia, 25*(3), 22-35. https://doi.org/10.1109/MMUL.2018.2873494

Anusha, K., Rajesh, N., Kavitha, M., & Ravinder, N. (2021). Comparative Study of MongoDB vs Cassandra in big data analytics. *5th International Conference on Computing Methodologies and Communication (ICCMC)*, 1831–1835. https://doi.org/10.1109/ICCMC51019.2021.9418441

Ardagna, C. A., Bellandi, V., Bezzi, M., Ceravolo, P., Damiani, E., & Hebert, C. (2021). Model-based big data analytics-as-a-Service: Take big data to the next level. *IEEE Transactions on Services Computing, Services Computing, 14*(2), 516–529. https://doi.org/10.1109/TSC.2018.2816941

Argento, L., Buccafurri, F., Furfaro, A., Graziano, S., Guzzo, A., Lax, G., Pasqua, F., & Saccà, D. (2021). ID-Service: A Blockchain-Based Platform to Support Digital-Identity-Aware Service Accountability. *Applied Sciences (2076-3417), 11*(1), 165. https://doi.org/10.3390/app11010165

Atick, J. (2016). Digital identity: the essential guide. In *ID4Africa Identity Forum*.

Aziz, A. A. (2017). Overcoming data integration challenges. National Identity Management Commission (NIMC). Retrieved from https://nimc.gov.ng/docs/overcoming_identity_challenges_24April2017.pdf

Babchuk, W. A. (2017). Qualitative Research: A Guide to Design and Implementation. *Adult Education Quarterly, 67*(1), 71–73. https://doi.org/10.1177/0741713616671930

Barczak, G. (2015). Publishing qualitative versus quantitative research. *Journal of Product Innovation Management*, *32*(5), 658. https://doi.org/10.1111/jpim.12277

Barrdear, J., & Kumhof, M. (2021). The macroeconomics of central bank digital currencies. *Journal of Economic Dynamics and Control*, 104148. https://doi.org/10.1016/j.jedc.2021.104148

Bernal, M. V. (2017). Retail payments innovations in Peru: Modelo Peru and financial inclusion. *Journal of Payments Strategy & Systems*, *10*(4), 343–351.

Billestrup, J., & Stage, J. (2014, June). E-government and the Digital Agenda for Europe. In *International Conference of Design, User Experience, and Usability*, 71-80. Springer, Cham.

Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting & Social Change,* 130, 99–113. https://doi.org/10.1016/j.techfore.2017.07.027

Boddy, C. R. (2016). Sample size for qualitative research. *An International Journal of Qualitative Market Research, 19*(4), 426-432. https://doi.org/10.1108/QMR-06-2016-0053

Boontaetae, P., Sangpetch, A., & Sangpetch, O. (2018). RDI: Real Digital Identity Based on Decentralized PKI. *2018 22nd International Computer Science and Engineering Conference (ICSEC)*, 1–6. https://doi.org/10.1109/ICSEC.2018.8712663

Boysen, A. (2021). Decentralized, Self-Sovereign, Consortium: The Future of Digital Identity in Canada. *Frontiers in Blockchain, 4*. https://doi.org/10.3389/fbloc.2021.624258

Bromley, E., Mikesell, L., & Khodyakov, D. (2017). Ethics and science in the participatory era: A Vignette-Based Delphi Study. *Journal of Empirical Research on Human Research Ethics, 12*(5), 295-309. https://doi.org/10.1177/1556264617717828

Brown, W., 3rd, Weng, C., Vawdrey, D. K., Carballo-Diéguez, A., & Bakken, S. (2017). SMASH: A Data-driven Informatics Method to Assist Experts in Characterizing Semantic Heterogeneity among Data Elements. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2016, 1717–1726.

Buchanan, R. (2019). Systems thinking and design thinking: The search for principles in the world we are making. *She Ji: The Journal of Design, Economics, and Innovation, 5*(2), 85–104. https://doi.org/10.1016/j.sheji.2019.04.001

Bughin, J., Deakin, J., & O'Beirne, B. (2019). Digital transformation: Improving the odds of success. *McKinsey Quarterly*.

Chalaemwongwan, N., & Kurutach, W. (2018). A Practical National Digital ID Framework on Blockchain (NIDBC). *15th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 497–500. https://doi.org/10.1109/ECTICon.2018.8620003

Champion, C., Kuziemsky, C., Affleck, E., & Alvarez, G. G. (2019). A systems approach for modeling health information complexity. *International Journal of Information Management, 49*, 343–354. https://doi.org/10.1016/j.ijinfomgt.2019.07.002

Choi, D. D., Laughlin, B., & Schultz, A. E. (2021). Mobile Communication Technology and National Identity in Sub-Saharan Africa. *Journal of Engineering*, 1477.

Clacy, A., Goode, N., Sharman, R., Lovell, G. P., & Salmon, P. (2019). A systems approach to understanding the identification and treatment of sport-related concussion in community rugby union. *Applied Ergonomics, 80*, 256–264. https://doi.org/10.1016/j.apergo.2017.06.010

Clark, J. M. (2018). *Public Sector Savings and Revenue from Identification Systems: Opportunities and Constraints (English)*. Washington, D.C. World Bank Group. http://documents1.worldbank.org/curated/en/745871522848339938/pdf/Public-Sector-Savings-and-Revenue-from-Identification-Systems-Opportunities-and-Constraints.pdf

Cleland, V., & Hartsink, G. (2019). The value of the Legal Entity Identifier for the payments industry. *Journal of Payments Strategy & Systems, 13*(4), 322–336.

Collomb, A., & Sok, K. (2016). Blockchain/distributed ledger technology (DLT): What impact on the financial sector?. Digiworld Economic Journal, (103), 93-111.

Cope, D. G. (2014). Methods and Meanings: Credibility and Trustworthiness of Qualitative Research. Oncology Nursing Forum, 41(1), 89–91. https://doi.org/10.1188/14.ONF.89-91

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed method approaches.* Thousand Oaks, CA. SAGE Publications.

Creswell, J. W., & Miller, D. L. (2000). Determining Validity in Qualitative Inquiry. *Theory Into Practice, 39*(3), 124-130. https://doi.org/10.1207/s15430421tip3903_2

Cronin, C. (2014). Using case study research as a rigorous form of inquiry. *Nurse Researcher, 21*(5), 19.

Cunningham, J. A., Menter, M., & Young, C. (2017). A review of qualitative case methods trends and themes used in technology transfer research. *The Journal of Technology Transfer, 42*(4), 923. https://doi.org/10.1007/s10961-016-9491-6

Daniel, B. K. (2018). Reimaging Research Methodology as Data Science. *Big Data and Cognitive Computing, 2*(1), 4. https://doi.org/10.3390/bdcc2010004

Dedoose. (n.d.). *User guide: Combining and analyzing mixed method data*. Retrieved from dedoose: https://www.dedoose.com/userguide

Demirgüç-Kunt, A., Klapper, L., Singer, D., Ansar, S., and Hess, J. (2018). *The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution.* Overview booklet. Washington, DC: World Bank. https://globalfindex.worldbank.org/

Dixon, P. (2017). A Failure to "Do No Harm"--India's Aadhaar biometric ID program and its inability to protect privacy in relation to measures in Europe and the US. *Health and technology, 7*(4), 539-567.

Drysdale, R., Cook, C. E., Petryszak, R., Baillie-Gerritsen, V., Barlow, M., Gasteiger, E., Gruhl, F., Haas, J., Lanfear, J., Lopez, R., Redaschi, N., Stockinger, H., Teixeira, D., Venkatesan, A., Blomberg, N., Durinx, C., & McEntyre, J. (2020). The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences. *Bioinformatics (Oxford, England), 36*(8), 2636–2642. https://doi.org/10.1093/bioinformatics/btz959

Dsouza, M. J. (2017). The practice of qualitative research. *An International Journal of Qualitative Research in Organizations and Management, 12*(3), 247-248. https://doi.org/10.1108/QROM-09-2016-1416

Dunn, H. S. (2020). Risking identity: a case study of Jamaica's short-lived national ID system. *Journal of Information, Communication & Ethics in Society, 18*(3), 329–338. https://doi.org/10.1108/JICES-04-2020-0040

Eaton, B., Hedman, J., & Medaglia, R. (2017). Three different ways to skin a cat: financialization in the emergence of national e-ID solutions. *Journal of Information Technology*, 1-17. https://doi: 10.1057/s41265-017-0036-8

Eduard J. Beck, Wayne Gill, & Paul R. De Lay. (2016). Protecting the confidentiality and security of personal health information in low- and middle-income countries in the era of SDGs and Big Data. *Global Health Action, 9*(0), 1–7. https://doi.org/10.3402/gha.v9.32089

Ellis, T. J., & Levy, Y. (2009). Towards a guide for novice researchers on research methodology: review and proposed methods. *Issues in Informing Science & Information Technology, 6*, 323-337.

Espinosa-González, A. B., Delaney, B. C., Marti, J., & Darzi, A. (2019). The impact of governance in primary health care delivery: a systems thinking approach with a European panel. *Health Research Policy and Systems, 17*(1). https://doi.org/10.1186/s12961-019-0456-8

Esposti, S. D. (2014). When big data meets dataveillance: The hidden side of analytics. *Surveillance & Society, 12*(2), 209. https://doi.org/10.24908/ss.v12i2.5113

Esquivel-Korsiak, V., Mittal, A. (2018). Study of Options for Mutual Recognition of National IDs in the East African Community (English). Identification for Development Washington, D.C.: World Bank Group. Retrieved from http://documents.worldbank.org/curated/en/337501535031584335/Study-of-Options-for-Mutual-Recognition-of-National-IDs-in-the-East-African-Community

Faezov, B., & Dunbrack, R. L. (2021). PDBrenum: A webserver and program providing Protein Data Bank files renumbered according to their UniProt sequences. *PLoS ONE*, *16*(7), 1–15. https://doi.org/10.1371/journal.pone.0253411

Federal Negarit Gazeta. (2012). *Proclamation on the registration of vital events and national identity card (Proc. No. 760/2012)* [Ethiopia]. Retrieved from https://www.refworld.org/docid/5ec7f94d4.html

Festus, O. A., Sunday, A. O., & Jeremiah, A. O. (2017). Schematic structure of national data harmonization system for identity management. *European Scientific Journal, 13*(3), 318. https://doi.org/10.19044/esj.2016.v13n3p318

FIDO. (n.d.). *FIDO Alliance: Simpler, Stronger Authentication- Solving the World's Password Problem*. Retrieved from https://fidoalliance.org/

Francis-Auton, E., Warren, C., Braithwaite, J., & Rapport, F. (2020). Exploring the recruitment, ethical considerations, conduct and information dissemination of an audiology trial: a pretrial qualitative study (q-COACH). *Trials, 21*(1), 28, 1-15. https://doi.org/10.1186/s13063-019-3968-1

Frey, B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation* (Vols. 1-4). Thousand Oaks, CA: SAGE Publications, Inc. https://doi:10.4135/9781506326139

Fryer, D. (2001). *Doing Qualitative Research Differently: Free Association, Narrative and the Interview Method* Holloway, W. and Jefferson, T. (2000) Sage, London, 166pp, ?15.99 ISBN 0-7619-6426-6 (softback). Journal of Community & Applied Social Psychology, 11(4), 324–327. https://doi.org/10.1002/casp.603

Fung, B. S., & Halaburda, H. (2016). Central bank digital currencies: a framework for assessing why and how. *Available at SSRN 2994052*.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Garcia, L., Bolleman, J., Gehant, S., Redaschi, N., Martin, M., & UniProt Consortium. (2019). FAIR adoption, assessment and challenges at UniProt. *Scientific Data, 6*(1), 1–4. https://doi.org/10.1038/s41597-019-0180-9

Gaur, A. D., & Padiya, J. (2016). A Study Impact of ‘Digital India ‘in ‘Make in India’ Program in IT & BPM Sector. In *Fourteenth AIMS International Conference on Management*, 325-331.

Gelb, A., & Clark, J. (2013). Identification for development: The biometrics revolution. *Center for Global Development Working Paper*, (315).

Gelb, A., Mukherjee, A., & Diofasi, A. (2016). *Iris Recognition: Better than Fingerprints and Falling in Price.* Center for Global Development. Retrieved from: https://www.cgdev.org/blog/iris-recognition-better-fingerprints-and-falling-price

Gelb, A. H., & Palacios, R. J. (2016). *ID4D Country Diagnostic: Ethiopia* (No. 142092, pp. 1-40). The World Bank, Washington, DC. Retrieved from http://pubdocs.worldbank.org/en/822621524689442102/Ethiopia-ID4D-Diagnostic-Web040418.pdf

Gerber, A., le Roux, P., & van der Merwe, A. (2020). Enterprise Architecture as Explanatory Information Systems Theory for Understanding Small- and Medium-Sized Enterprise Growth. *Sustainability, 12*(8517), 8517. https://doi.org/10.3390/su12208517

Gill, P., Stewart, K., Treasure, E., & Chadwick, B. (2008). Methods of data collection in qualitative research: interviews and focus groups. *British Dental Journal, 204*(6), 291–295. https://doi.org/10.1038/bdj.2008.192

Gill, S. L. (2020). Qualitative Sampling Methods. *Journal of Human Lactation, 36*(4), 579–581. https://doi.org/10.1177/0890334420949218

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The qualitative report, 8*(4), 597-607.

Goldkuhl, G. (2004). Design theories in information systems – A need for multigrounding. *Journal of Information Technology Theory and Application (JITTA), 6*(2), 59–72. Retrieved from https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1127&context=jitta

Grassi, P., Fenton, J., Lefkovitz, N., Danker, J., Choong, Y. Y., Greene, K., & Theofanos, M. (2017). Digital identity guidelines: enrollment and identity proofing. *National Institute of Standards and Technology. Special Publication (SP) 800-63A.* https://doi.org/10.6028/NIST.SP.800-63a

Gregor, S. (2006). The nature of theory in information systems. *MIS quarterly*, 611-642.

Gregor, S., & Jones, D. (2007). The anatomy of a design theory. *Journal of the Association for Information Systems, 8*(5), 312–335. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.232.743&rep=rep1&type=pdf

Guerra, F., Sottovia, P., Paganelli, M., & Vincini, M. (2019). Big Data Integration of Heterogeneous Data Sources: The Re-Search Alps Case Study. *2019 IEEE International Congress on Big Data (BigDataCongress)*, 106-110. https://doi.org/10.1109/BigDataCongress.2019.00027

Halawi, L. A., & McCarthy, R. (2006). Which Theory Applies: An Analysis of Information Systems Research. Issues in Information Systems, 7(2), 252–256.

Harbitz, M. (2016). Digital identity. *World development report*, 194-197.

Hardjono, T. (2019). Federated Authorization over Access to Personal Data for Decentralized *Identity Management. IEEE Communications Standards Magazine, 3*(4), 32–38. https://doi.org/10.1109/MCOMSTD.001.1900019

Hatchuel, A., Le Masson, P., Reich, Y., & Subrahmanian, E. (2017). Design theory: a foundation of a new paradigm for design science and engineering. *Research in Engineering Design, 1*. https://doi.org/10.1007/s00163-017-0275-2

Hayashi, P., Abib, G., & Hoppen, N. (2019). Validity in Qualitative Research: A Processual Approach. *The Qualitative Report, 24*(1), 98.

Heripracoyo, S., & Kurniawan, R. (2016). Big Data Analysis with MongoDB for Decision Support System. *Telkomnika, 14*(3), 1083–1089. https://doi.org/10.12928/telkomnika.v14i2.3115

Herwig, R., Hardt, C., Lienhard, M., & Kamburov, A. (2016). Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nature Protocols, 11*(10), 1889. https://doi.org/10.1038/nprot.2016.117

Heyvaert, M., Hannes, K., Maes, B., & Onghena, P. (2013). Critical appraisal of mixed methods studies. *Journal of Mixed Methods Research, 7*, 302-327. https://doi.org/10.1177/1558689813479449

Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development Policy Review, 34*(1), 135-174. https://doi.org/10.1111/dpr.12142

Horita, F. E. A., de Albuquerque, J. P., Marchezini, V., & Mendiondo, E. M. (2017). Bridging the gap between decision-making and emerging big data sources: An application of a model-based framework to disaster management in Brazil. *Decision Support Systems,* 97, 12–22. https://doi.org/10.1016/j.dss.2017.03.001

Hsu, Chin-Ming, & Chao, Hui-Mei. (2009). Securing Computerized Personal Data during Transit and at Rest Using Programmable System on Chip (PSoC) Technology. *2009 WRI*

World Congress on Computer Science and Information Engineering, Computer Science and Information Engineering, 4, 415–420. https://doi.org/10.1109/CSIE.2009.490

Htet, M., Yee, P. T., & Rajasekera, J. R. (2020). Blockchain based Digital Identity Management System: A Case Study of Myanmar. *2020 International Conference on Advanced Information Technologies (ICAIT)*, 42–47. https://doi.org/10.1109/ICAIT51105.2020.9261785

Hue, T. T. (2019). The determinants of innovation in Vietnamese manufacturing firms: an empirical analysis using a technology–organization–environment framework. *Eurasian Business Review, 9*(3), 247–267.

Husni, E. (2016). Digital signing using national identity as a mobile ID. *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 261–264. https://doi.org/10.1109/ISITIA.2016.7828668

International Organization for Standardization. (2009). *Information technology — Security techniques — Security evaluation of biometrics.* Retrieved from https://www.iso.org/standard/51521.html

International Organization for Standardization. (2011). *Information technology — Security techniques — Privacy framework* (ISO/IEC Standard No. 29100:2011) https://www.iso.org/standard/45123.html

International Organization for Standardization. (2013). *Information technology — Security techniques — Entity authentication assurance framework* (ISO/IEC Standard No. 29115:2013) https://www.iso.org/standard/45138.html

International Organization for Standardization. (2014). *Information technology — Biometric data interchange formats — Part 7: Signature/sign time series data* (ISO/IEC Standard No. 19794-7:2014) https://www.iso.org/standard/55938.html

International Organization for Standardization. (2015). *Information technology — Security techniques — Privacy capability assessment model* (ISO/IEC Standard No. 29190:2015) https://www.iso.org/standard/45269.html

International Organization for Standardization. (2018a). Contributing to the UN SDGs with ISO standards. ISO Central Secretariat, Geneva - Switzerland. From https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100429.pdf

International Organization for Standardization. (2018b). *Information technology — Guidance for biometric enrolment* (ISO/IEC Standard No. TR 29196:2018) https://www.iso.org/standard/70951.html

International Organization for Standardization. (2019a). *Information technology — Security techniques — Code of practice for protection of personally identifiable information (PII) in public clouds acting as PII processors* (ISO/IEC Standard No. 27018:2019) https://www.iso.org/standard/76559.html

International Organization for Standardization. (2019b). *IT Security and Privacy — A framework for identity management — Part 1: Terminology and concepts.* (ISO/IEC Standard No. 24760-1:2019) https://www.iso.org/standard/76559.html

International Organization for Standardization. (2020). *Information technology — Online privacy notices and consent* (ISO/IEC Standard No. 29184:2020) https://www.iso.org/standard/70331.html

Jahanmir, S. F., & Cavadas, J. (2018). Factors affecting late adoption of digital innovations. *Journal of business research*, *88*, 337-343.

Johnson, J. L., Adkins, D., & Chauvin, S. (2020). A Review of the Quality Indicators of Rigor in Qualitative Research. *American Journal of Pharmaceutical Education*, *84*(1), 138–146. https://doi.org/10.5688/ajpe7120

Jung, H., & Chung, K. (2021). Social mining-based clustering process for big-data integration. *Journal of Ambient Intelligence and Humanized Computing*, *12*(1), 589 – 600. https://doi.org/10.1007/s12652-020-02042-7

Kadadi, A., Agrawal, R., Nyamful, C., & Atiq, R. (2014). Challenges of data integration and interoperability in big data. *2014 IEEE International Conference on Big Data (Big Data)*, 38–40. https://doi.org/10.1109/BigData.2014.7004486

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., & Herwig, R. (2011). ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Research*, *39(Database issue)*, D712–D717. https://doi.org/10.1093/nar/gkq1156

Kane, G. (2019). The technology fallacy: people are the real key to digital transformation. *Research-Technology Management*, *62*(6), 44-49.

Kautz, K., Bjerknes, G., Fisher, J., & Jensen, T. (2020). Applying Complex Adaptive Systems Theory to Understand Distributed Participatory Design in Crowdsourced Information Systems Development. *Australasian Journal of Information Systems*, 24. https://doi.org/10.3127/ajis.v24i0.2225

Kharat, A. T., & Singhal, S. (2017). A peek into the future of radiology using big data applications. *The Indian journal of radiology & imaging*, *27*(2), 241. https://doi.org/10.4103/ijri.IJRI49316

Khoo, M., Rozaklis, L., & Hall, C. (2012). A survey of the use of ethnographic methods in the study of libraries and library users. *Library and Information Science Research*, *34*(2), 82–91. https://doi.org/10.1016/j.lisr.2011.07.010

Kuperberg, M., Kemper, S., & Durak, C. (2019). Blockchain usage for government-issued electronic IDs: A survey. *In International Conference on Advanced Information Systems Engineering*, 155-167. Springer, Cham 2019. https://doi.org/10.1007/978-3-030-20948-3_14

Leedy, P. D., & Ormrod, J. E. (2010). *Practical research: Planning and design* (9th ed.) New York: Merril.

Leung, D., Lo, A., Nang Fong, L. H., & Law, R. (2015). Applying the Technology-Organization-Environment framework to explore ICT initial and continued adoption: An exploratory study of an independent hotel in Hong Kong. *Tourism Recreation Research*, *40*(3), 391.

Leung, L. (2015). Validity, reliability, and generalizability in qualitative research. *Journal of Family Medicine & Primary Care*, *4*(3), 324–327. https://doi.org/10.4103/2249-4863.161306

Levina, N. (2021). All Information Systems Theory Is Grounded Theory. *MIS Quarterly*, *45*(1), 489–494. https://doi.org/10.25300/MISQ/2021/15434.1.7

Lewis, K. B., Graham, I. D., Boland, L., & Stacey, D. (2021). Writing a compelling integrated discussion: a guide for integrated discussions in article-based theses and dissertations. *International Journal of Nursing Education Scholarship*, *19*(1), 1–9. https://doi.org/10.1515/ijnes-2020-0057

Liedtka, J., King, A., & Bennett, K. (2013). *Solving Problems with Design Thinking: Ten Stories of What Works*. Columbia University Press.

Light, J., & Yasuhara, K. (2008). Analyzing large free-response qualitative data sets — a novel quantitative-qualitative hybrid approach. *2008 38th Annual Frontiers in Education Conference*. https://doi.org/10.1109/FIE.2008.4720426

Lim, S. S., Cho, H., & Sanchez, M. R. (2009). Online Privacy, Government Surveillance and National ID Cards. *Communications of the ACM, 52*(12), 116–120. https://doi.org/10.1145/1610252.1610283

Magrane, M., & Consortium, U. (2010). UniProt Knowledgebase: a hub of integrated data. Nature Precedings, 1. https://doi.org/10.1038/npre.2010.5092.1

Mansoor, Z., & Williams, M. J. (2018). Systems approaches to public service delivery: lessons from health, education, and infrastructure. *Systems of public service delivery in developing countries*, 2018-06.

Markus, M. L., Majchrzak, A., & Gasser, L. (2002). A design theory for systems that support emergent knowledge processes. *MIS Quarterly, 26*(3), 179–212. Retrieved from https://www.jstor.org/stable/pdf/4132330.pdf

Martin, A. (2021). Aadhaar in a Box? Legitimizing Digital Identity in Times of Crisis. *Surveillance & Society, 19*(1), 104-108. https://doi.org/10.24908/ss.v19i1.14547

Martinson, P. (2019). Estonia–the Digital Republic Secured by Blockchain. *PricewaterhouseCoopers: London, UK*, 1-12. Retrieved from https://www.pwc.com/gx/en/services/legal/tech/assets/estonia-the-digital-republic-secured-by-blockchain.pdf

Meho, L. I. (2006). E-mail interviewing in qualitative research: A methodological discussion. *Journal of the American Society for Information Science & Technology, 57*(10), 1284–1295. https://doi.org/10.1002/asi.20416

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative Research: A Guide to Design and Implementation* (4th ed.). San Francisco, CA: Jossey-Bass.

Milena, Z. R., & Dainora, G. (2008). Qualitative Research Methods: A Comparison between Focus-Group and In-Depth Interview. *Annals of the University of Oradea: Economic Science, 4*(1), 1279–1283.

Mhamane, S., & Shriram, P. (2018). Railway Ticket Verification and Dynamic seat Allocation using Aadhar Card. *International Conference on Inventive Computation Technologies (ICICT)*, 293–296. https://doi.org/10.1109/ICICT43934.2018.9034437

Mićić, L. (2017). Digital transformation and its influence on GDP. *ECONOMICS-Innovative and Economic Research*, *5*(2), 135-147.

Mihaylov, I., Kańduła, M., Krachunov, M., & Vassilev, D. (2019). A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models. *Biology Direct, 14*(1), N.PAG. https://doi.org/10.1186/s13062-019-0249-6

Millová, K., & Blatný, M. (2015). *Personality Development: Systems Theories*. In International Encyclopedia of the Social & Behavioral Sciences (Second Ed.), 879–883. https://doi.org/10.1016/B978-0-08-097086-8.23035-3

Mir, U. B., Kar, A. K., Dwivedi, Y. K., Gupta, M. P., Sharma, R. S. (2020). Realizing digital identity in government: Prioritizing design and implementation objectives for Aadhaar in India. *Government Information Quarterly*, *37*(2), Article 101442. https://doi.org/10.1016/j.giq.2019.101442

Mittal, A. (2018). *Catalog of Technical Standards for Digital Identification Systems (English)*. Washington, D.C. World Bank Group. Retrieved from

https://documents.worldbank.org/en/publication/documents-reports/documentdetail/707151536126464867/catalog-of-technical-standards-for-digital-identification-systems

Mobasher, M. B. (2018). Identity cards and identity conflicts: a cross-national analysis of national ID cards and the lessons for Afghanistan. *Indian Law Review (24730580), 2*(2), 159–177. https://doi.org/10.1080/24730580.2018.1547078

Moon, M. R. (2017, September 22). Can Central IRBs Replace Local Review? *Journal of Law, Medicine & Ethics, 45*(3), 348. https://doi.org/10.1177/1073110517737533

Mountasser, I., Ouhbi, B., Hdioud, F., & Frikh, B. (2021). Semantic-based big data integration framework using scalable distributed ontology matching strategy. *Distributed and parallel databases: An International Journal of Data Science, Engineering, and Management*, 1-47. https://doi.org/10.1007/s10619-021-07321-6

Mudadu, M. de A., & Zerlotini, A. (2020). Machado: Open source genomics data integration framework. *GigaScience, 9*(9). https://doi.org/10.1093/gigascience/giaa097

Mühle, A., Grüner, A., Gayvoronskaya, T., & Meinel, C. (2018). A survey on essential components of a self-sovereign identity. *Computer Science Review, 30*, 80–86. https://doi.org/10.1016/j.cosrev.2018.10.002

Naik, N., & Jenkins, P. (2017). Securing digital identities in the cloud by selecting an apposite Federated Identity Management from SAML, OAuth and OpenID Connect. *2017 11th International Conference on Research Challenges in Information Science (RCIS)*, 163–174. https://doi.org/10.1109/RCIS.2017.7956534

Najles, Julian. (2020). *Argentina ID Case Study: The Evolution of Identification (English). Identification for Development.* Washington, D.C.: World Bank Group.

http://documents.worldbank.org/curated/en/318351582559995027/Argentina-ID-Case-Study-The-Evolution-of-Identification

Natarajan, H., Krause, S., & Gradstein, H. (2017). *Distributed ledger technology and blockchain.* International Bank for Reconstruction and Development / the World Bank, Washington, DC. https://doi.org/10.1596/29053

Nashipudimath, M. M., Shinde S. K., & Jain, J. (2020). An efficient integration and indexing method based on feature patterns and semantic analysis for big data. *Array*, *7*(100033-). https://doi.org/10.1016/j.array.2020.100033

Nind, M., & Vinha, H. (2016). Creative interactions with data: using visual and metaphorical devices in repeated focus groups. Qualitative Research, 16(1), 9-26. https://doi.org/10.1177/1468794114557993

Noble, H., & Smith, J. (2015). Issues of validity and reliability in qualitative research. *Evidence Based Nursing 18*(2), 34–35. https://doi.org/10.1136/eb-2015-102054

Noor, A. (2020). FIDO: Fast IDentity Online. *ISSA Journal*, *18*(12), 22–26.

Ospina, S. M., Esteve, M., & Lee, S. (2018). Assessing Qualitative Studies in Public Administration Research. *Public Administration Review*, *78*(4), 593–605. https://doi.org/10.1111/puar.12837

O'Sullivan, T. A., & Jefferson, C. G. (2020). A Review of Strategies for Enhancing Clarity and Reader Accessibility of Qualitative Research Results. *American Journal of Pharmaceutical Education*, *84*(1), 147–155. https://doi.org/10.5688/ajpe7124

Pan, Y., Tian, Y., Liu, X, Gu, D, & Hua, G. (2016). Urban big data and the development of city intelligence. *Engineering. 2*(2), 171-178. https://doi.org/10.1016/J.ENG.2016.02.003

Papakitsou, V. (2020). Qualitative Research: Narrative approach in sciences. *Dialogues in Clinical Neuroscience & Mental Health, 3*(1), 63–70. https://doi.org/10.26386/obrela.v3i1.177

Pearsall, E. A., Meghji, Z., Pitzul, K. B., Aarts, M.-A., McKenzie, M., McLeod, R. S., & Okrainec, A. (2015). A qualitative study to understand the barriers and enablers in implementing an enhanced recovery after surgery program. *Annals of Surgery, 261*(1), 92-96. https://doi.org/10.1097/sla.0000000000000604

Power, D. J. (2014). Using 'big data' for analytics and decision support. *Journal of Decision Systems, 23*(2), 222-228. https://doi.org/10.1080/12460125.2014.888848

Priyasta, D., Cesar, W., Susanti, Y., & Junde, J. (2018). Java Card Approach to Emulate The Indonesian National Electronic ID Smart Cards. *Scientific Journal of Informatics, 5*(2), 224–234. https://doi.org/10.15294/sji.v5i2.16347

Preciozzi, J., Garella, G., Camacho, V., Franzoni, F., Di Martino, L., Carbajal, G., & Fernandez, A. (2020). Fingerprint Biometrics From Newborn to Adult: A Study From a National Identity Database System. *IEEE Transactions on Biometrics, Behavior, and Identity Science, 2*(1), 68–79. https://doi.org/10.1109/TBIOM.2019.2962188

Price, J., & Murnan, J. (2004). “Research Limitations and the Necessity of Reporting Them.” *American Journal of Health Education, 35* (2004): 66-67.

Randall, D., Ardic Alper, O. P., Varghese, M. M., & Traversa, M. (2017). *Global financial inclusion and consumer protection survey: 2017 report* (No. 122058, pp. 1-89). The World Bank. https://doi.org/10.1596/28998

Rashid, M., Caine, V., & Goez, H. (2015). The encounters and challenges of ethnography as a methodology in health research. *International Journal of Qualitative Methods,14*(5). https://doi:10.1177/1609406915621421

Reuben, W., & Carbonari, F. (2017). Identification as a national priority: the unique case of Peru. *Center for global development working paper*, (454).

Roberts, R. E. (2020). Qualitative Interview Questions: Guidance for Novice Researchers. *The Qualitative Report, 25*(9), COV1.

Rogers, Everett (16 August 2003). *Diffusion of Innovations*, 5th Edition. Simon and Schuster

Ruggia, R., Delgado, A., Abin, J., González, L., & Garbusi, P. (2016). Managing consistency in e-government transactions: the case of Uruguay. *In Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance*, 313-322.

Ryan, M. (2009). Making Visible the Coding Process: Using Qualitative Data Software in a Post-Structural Study. *Issues in Educational Research, 19*(2), 142–161.

Salmona, M., & Kaczynski, D. (2016). Don't blame the software: Using qualitative data analysis software successfully in doctoral research. *Forum: Qualitative Social Research 17*(3), 42-64. https://doi.org/10.17169/fqs-17.3.2505.

Saputro, R., Pappel, I., Vainsalu, H., Lips, S., & Draheim, D. (2020). Prerequisites for the Adoption of the X - Road Interoperability and Data Exchange Framework: A Comparative Study. *Seventh International Conference on eDemocracy & eGovernment (ICEDEG)*, 216–222. https://doi.org/10.1109/ICEDEG48599.2020.9096704

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z., Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., … Sherry, S.

T. (2021). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 49*(D1), D10–D17. https://doi.org/10.1093/nar/gkaa892

Sazontev, V. (2018, October). Methods for Big Data Integration in Distributed Computation Environments. *In DAMDID/RCDL,* 238-244.

Sazontev, V., & Stupnikov, S. (2019). An Extensible Approach for Materialized Big Data Integration in Distributed Computation Environments. *2019 Ivannikov Memorial Workshop (IVMEM)*, 33–38. https://doi.org/10.1109/IVMEM.2019.00011

Schou, J., & Hjelholt, M. (2018). Digital citizenship and neoliberalization: governing digital citizens in Denmark. *Citizenship Studies*, *22*(5), 507-522. https://doi.org/10.1080/13621025.2018.1477920

Schultze, U., & Avital, M. (2011). Designing interviews to generate rich data for information systems research. *Information and Organization*, *21*(1), 1-16. https://doi.org/10.1016/j.infoandorg.2010.11.001

Schwertner, K. (2017). Digital transformation of business. *Trakia Journal of Sciences*, *15*(1), 388-393.

Sen, S. (2019). A Decade of Aadhaar: Lessons in implementing a foundational ID system. *ORF Issue Brief No, 292*.

Shatnawi, M. Q., Yassein, M. B., Abuein, Q., & Nsuir, L. (2019). Big data analytics tools and applications: survey. *Data Science, E-Learning and Information Systems*, 1–4. https://doi.org/10.1145/3368691.3368741

Sharma, P. K., Chen, M. Y., & Park, J. H. (2017). A software defined fog node based distributed blockchain cloud architecture for IoT. *Ieee Access*, *6*, 115-124.

Siamagka, N.T., Christodoulides, G., Michaelidou, N., & Valvi, A. (2015). Determinants of social media adoption by b2b organizations. *Industrial Marketing Management*, 51, 89-99. https://doi.org/10.1016/j.indmarman.2015.05.005

Simon, H. A. (1996). The sciences of the artificial (3rd ed.). London: MIT Press. Retrieved from https://monoskop.org/images/9/9c/Simon_Herbert_A_The_Sciences_of_the_Artificial_3rd_ed.pdf

Simon, M. K. (2011). *Dissertation and scholarly research: Recipes for success*. (2011 Ed.). Seattle, WA, Dissertation Success, LLC.

Smith, J. A. (2015). *Qualitative psychology: A practical guide to research methods* (3rd ed.). Thousand Oaks, CA: SAGE.

Smith, S. (2016). Voice and Facial Recognition to Be Used in Over 600 million Mobile Devices by 2021. *Juniper Research, London.* Retrieved from https://www. juniperresearch. com/press/press-releases/voice-and-facial-recognition-to-be-used-in-over-60

Snyder, C. (2012). A case study of a case study: analysis of a robust qualitative research methodology. *The Qualitative Report*.

Soltani, R., Nguyen, U. T., & An, A. (2021). A Survey of Self-Sovereign Identity Ecosystem. *Security & Communication Networks,* 1–26. https://doi.org/10.1155/2021/8873429

Strauss, Daniël F. M. (2002). The scope and limitations of Von Bertalanffy's systems theory. *South African Journal of Philosophy, 21*(3), 163–179

Subha, R. (2017). Biometrics in Internet of Things (IoT) Security. *International Journal of Engineering Research and General Science, 5*(5), 37-42.

Sullivan, C. (2018). Digital identity–From emergent legal concept to new reality. *Computer Law & Security Review, 34*(4), 723-731.

Suryani, A. (2013). Comparing Case Study and Ethnography as Qualitative Research Approaches. *Jurnal Ilmu Komunikasi, 5*(1). https://doi.org/10.24002/jik.v5i1.221

Taherdoost, H. (2018). A review of technology acceptance and adoption models and theories. *Procedia Manufacturing, 22*, 960–967.

Tam, S.-M., & Van-Halderen, G. (2020). The five V's, seven virtues and ten rules of big data engagement for official statistics. *Statistical Journal of the IAOS, 36*(2), 423–433. https://doi.org/10.3233/SJI-190595

Tammpuu, P., & Masso, A. (2019). Transnational Digital Identity as an Instrument for Global Digital Citizenship: The Case of Estonia's E-Residency. Information Systems Frontiers: *A Journal of Research and Innovation, 21*(3), 621. https://doi.org/10.1007/s10796-019-09908-y

Tavory, I. (2020). Interviews and Inference: Making Sense of Interview Data in Qualitative Research. *Qualitative Sociology, 43*(4), 449. https://doi.org/10.1007/s11133-020-09464-x

Timoshenko, K., Kuruppu, C., Badshah, I., & Ambalangodage, D. (2020). A Systems Approach to Comprehend Public Sector (Government) Accounting (pp. 50–59). https://doi.org/10.1007/978-3-030-22493-6_6

Tornatzky, L.G., & Fleischer, M. (1990). *The processes of technological innovation*. Lexington, Massachusetts, Lexington Books.

Tsang, E. W. (2014). Case studies and generalization in information systems research: A critical realist perspective. *The Journal of Strategic Information Systems, 23*(2), 174-186. https://doi.org/10.1016/j.jsis.2013.09.002

Tungpantong, C., Nilsook, P., & Wannapiroon, P. (2021). A Conceptual Framework of Factors for Information Systems Success to Digital Transformation in Higher Education

Institutions. *9th International Conference on Information and Education Technology (ICIET)*, 57–62. https://doi.org/10.1109/ICIET51873.2021.9419596

Tyagi, A. K., Rekha, G., & Sreenath, N. (2018). Is your Privacy Safe with Aadhaar?: An Open Discussion. *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, 318–323. https://doi.org/10.1109/PDGC.2018.8745836

UniProt: the universal protein knowledgebase in 2021. (2021). *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Uzodike, U. O., & Onapajo, H. (2019). Beyond the card reader: anti-election rigging technology and national security in Nigeria. *Insights on Africa (Sage Publications Inc.)*, *11*(2), 145–161. https://doi.org/10.1177/0975087819845194

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46* (2): 186–204.

Von Bertalanffy, L. (1972). The History and Status of General Systems Theory. *Academy of Management Journal*, *15*(4), 407–426. https://doi.org/10.2307/255139

Von Bertalanffy, L. (2008). An Outline of General System Theory. *Emergence: Complexity and Organization*, *10*(2), 103-123.

Wakimoto, D. K. (2013). Ethnographic methods are becoming more popular in LIS research. *Evidence Based Library and Information Practice, 8*(1), 96–98.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *Information Systems Research*, *3*(1), 36–59. https://doi.org/10.1287/isre.3.1.36

Wang, W., & Wang, Y. (2016). Research on currency substitution in developing country based on multiple motivations and big data analytics. *Revista Iberica de Sistemas e Tecnologias de Informacao (RISTI),* (E12), 411.

Waterman, K. K., & Bruening, P. J. (2014). Big data analytics: *Risks and responsibilities. International Data Privacy Law, 4(*2), 89-95. https://doi.org/10.1093/idpl/ipu002

Weitzberg, K. (2020). Biometrics, race making, and white exceptionalism: The controversy over universal fingerprinting in Kenya. *The Journal of African History, 61*(1), 23-43. https://doi.org/10.1017/S002185372000002X

Wenz, K. M., Palacios, R. J., Mills, S. (2018). *Incentives for Improving Birth Registration Coverage: A Review of the Literature (English)*. Identification for Development. Washington, D.C.: World Bank Group.

White, O., Madgavkar, A., Manyika, J., Mahajan, D., Bughin, J., McCarthy, M., Sperling, O. (2019). *Digital identification: A key to inclusive growth. McKinsey Global Institute.* Retrieved from https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/digital%20identification%20a%20key%20to%20inclusive%20growth/mgi-digital-identification-report.pdf

Whitley, Edgar A. (2018). *Trusted digital identity provision: GOV.UK Verify's federated approach.* CGD policy paper, 131. Center for Global Development, Washington, USA.

World Bank. (2018). *Technology Landscape for Digital Identification*. Washington, DC: World Bank License: Creative Commons Attribution 3.0 IGO (CC BY 3.0 IGO). https://elibrary.worldbank.org/doi/abs/10.1596/31825#:~:text=Technology%20Landscap

e%20for%20Digital%20Identification.%20Robust%2C%20inclusive%2C%20and,nets%20and%20facilitating%20the%20development%20of%20digital%20economies.

World Bank Group (2017). *The state of identification systems in Africa – a synthesis of country assessments (English)*. Washington, D.C.: World Bank Group. Retrieved from http://documents.worldbank.org/curated/en/156111493234231522/The-State-of-identification-systems-in-Africa-a-synthesis-of-country-assessments

World Bank Group. (2018). *Identification for Development (ID4D) global dataset*. Retrieved from https://datacatalog.worldbank.org/dataset/identification-development-global-dataset

World Bank Group. (2021). *Principles on Identification for Sustainable Development: Toward the Digital Age - Second Edition (English)*. Washington, D.C.: World Bank Group. https://documents1.worldbank.org/curated/en/213581486378184357/pdf/Principles-on-Identification-for-Sustainable-Development-Toward-the-Digital-Age-Second-Edition.pdf

wwPDB consortium. (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research, 47*(D1), D520–D528. https://doi.org/10.1093/nar/gky949

Yaga, D., Mell, P., Roby, N., & Scarfone, K. (2019). Blockchain technology overview. *arXiv preprint arXiv:1906.11078*. https://doi.org/10.6028/NIST.IR.8202

Yakubiv, V., & Yakubiv, R. (2019). System of Organizational and Economic Support of Human Resources Management at Enterprises. *Journal of Vasyl Stefanyk Precarpathian National University, 6*(3–4), 88–95. https://doi.org/10.15330/jpnu.6.3-4.88-95

Yeong, M. L., Ismail, R., Ismail, N. H., & Hamzah, M. I. (2018). Interview Protocol Refinement: Fine-Tuning Qualitative Research Interview Questions for Multi-Racial Populations in Malaysia. *The Qualitative Report, 23*(11), 2700.

Yilmaz, K. (2013). Comparison of quantitative and qualitative research traditions: Epistemological, theoretical, and methodological differences. *European Journal of Education, 48*(2), 311 – 325. https://doi.org/10.1111/ejed.12014

Yu, X., & Wu, Q. (2020). Multi-source Heterogeneous Data Association Technology to Build Public Safety Big Data Integration Research. *2020 International Conference on Big Data Economy and Information Management (BDEIM)*, 17–20. https://doi.org/10.1109/BDEIM52318.2020.00012

Zhang, J., Yang, X., & Appelbaum, D. (2015). Toward effective big data analysis in continuous auditing. Accounting Horizons *American Accounting Association, 29* (2), 469-476. https://doi.org/10.2308/acch-51070

Zhao, X., Castka, P., & Searcy, C. (2020). ISO Standards: A Platform for Achieving Sustainable Development Goal 2. *Sustainability, 12*(22), 9332.

Zheng, Z., Xie, S., Dai, H., Chen, X., & Wang, H. (2017, June). An overview of blockchain technology: Architecture, consensus, and future trends. *2017 IEEE international congress on big data (BigData congress)*, 557-564. https://doi.org/10.1109/BigDataCongress.2017.85

Appendix A

Search Strategy: Parameters and Databases

Search Parameters	Search Target	Source or Database
"Digital Identification", "National Digital Identity", "National digital identity security", "Citizen's Registration", "National digital identification card", "Vital events' registration", "Data Integration", "Distributed Databases", "Blockchain implementation" "Distributed Ledger Technology", + "Ethiopia"	Searching for scholarly reviewed articles related to the National Identification System design and implementation, Security, Data integration,	Online Library: Northcentral University (NCU) NCU Library: https://library.ncu.edu/ • Academic Search Complete • Business Source Complete • Directory of Open Access Journals • IEEE Xplore Digital Library • SAGE Knowledge • ScienceDirect • Springer Nature Journals
	Searching for journal articles, master's theses or doctoral dissertations related to the National Identification System design and implementation	Online Library: Northcentral University (NCU) Dissertations and theses - A-Z Databases: • Computing & Information Technology: Journal Articles: ProQuest • Computing & Information Technology: Conference Proceedings: ProQuest • Dissertation Resources: Dissertations & Theses: ProQuest
	Searching for journal articles related to the National Identification System design and implementation	Google scholar Library - Articles
	Searching for master's thesis or doctoral dissertations related to the National Identification System design and implementation	Online Library: Addis Ababa University (AAU) • Institutional Repository: http://etd.aau.edu.et/
	ISO standards, Whitepapers, country reports, PPT presentations, guidelines related to the National Identification System design and implementation	Search Engines: • Google • Bing
"Vial Events Registration", "Citizen's Registration", "National Digital Identification Card", "Digital Identification system", "National ID"	Searching for government whitepapers, laws or proclamations, strategic documents related to Ethiopia's National Identification System design and implementation	Ethiopia's ministries and parliament websites: • Ministry of Innovation and Technology (MiNT): https://mint.gov.et/?lang=en • Ministry of Peace (MoP): http://www.peace.gov.et/ • House of peoples representatives 'of the FDRE: http://www.hopr.gov.et/ • Federal Democratic Republic of Ethiopia: Office of the prime minister: https://www.pmo.gov.et/government/

Appendix B

The Theoretical Framework Detail

The details of the theoretical framework components are depicted in the figure below. The components are the identification standards (Mittal, 2018), the big data concepts (Akhtar et al., 2019; Gandomi & Haider, 2015), the information systems design theory (Gregor & Jones, 2007), and the design principles on identification for sustainable development (World Bank Group [WBG], 2021).

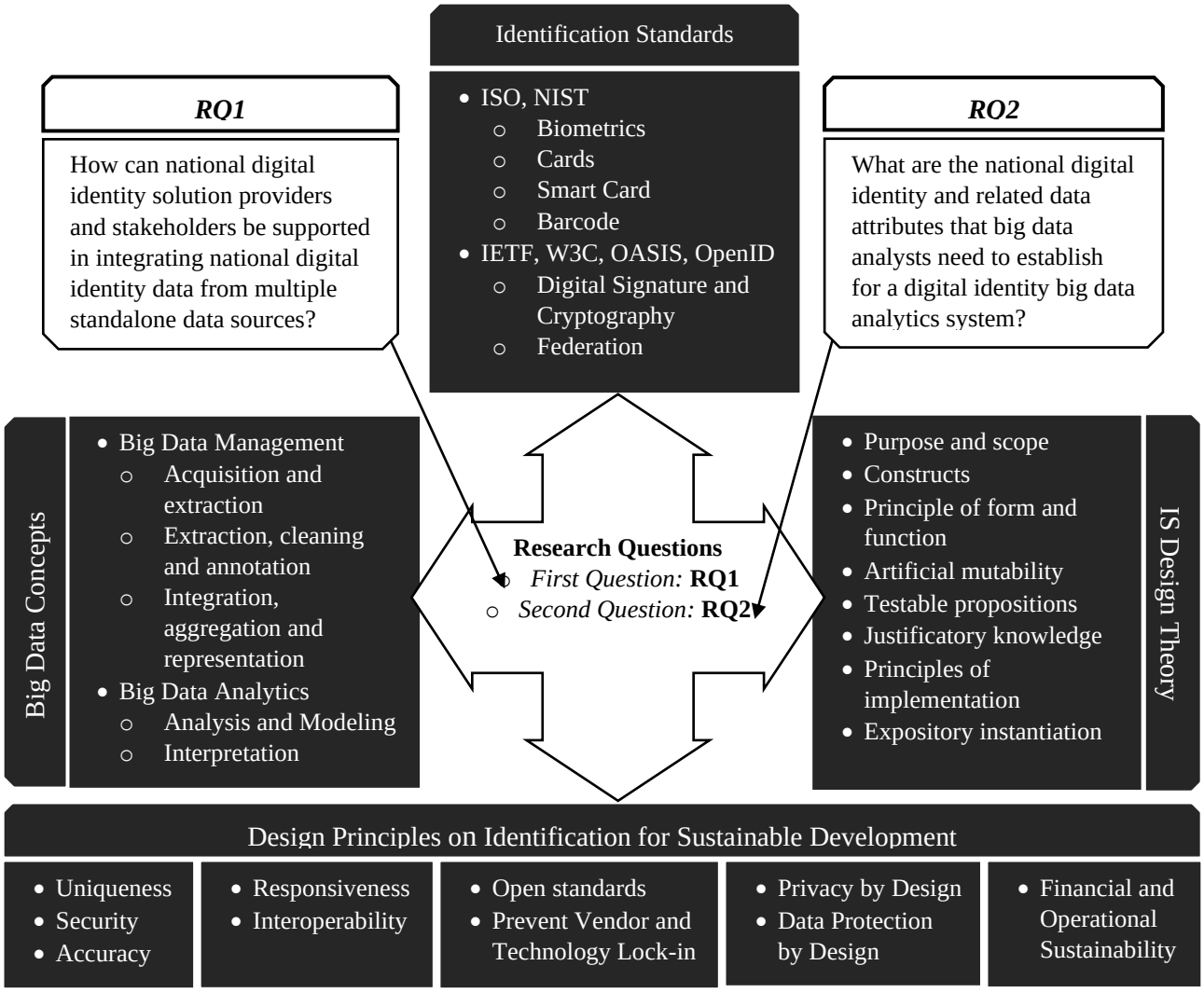


Figure. Theoretical Framework (Detailed)

Appendix C

Consent Letter

Introduction

My name is Denekew A. Jembere, and I am a doctoral student at Northcentral University (NCU).

I am conducting a research study about integrating a national identity database with different digital identity data sources. The name of this research study is "Exploring Strategies to Establish a Standardized Big Data Integration and Harmonization Framework for a National Identity Database." I am seeking your consent to participate in this study.

Please read this document to determine if you would like to participate. Your participation is completely voluntary, and I will address your questions or concerns at any point before or during the study.

Eligibility

You are eligible to participate in this study if:

1. You are age 18 or older.
2. You have a bachelor's degree.
3. You have experience with a national digital ID system implementation. Examples: Indian Aadhar, Estonian e-Estonia, Ethiopian National digital Id systems.
4. You have worked as a software engineer, data analyst, database administrator, or information technology manger in a national digital ID system implementation.

I hope to include ten people in this research.

Activities

If you decide to participate in this study, you will be asked to do the following activities:

1. Participate in an interview (in-person or via Zoom) for 45 to 60 minutes
2. Review your interview transcript via email for 10 - 15 minutes

During these activities, you will be asked questions about:

- National digital Id data integration
- National digital Id standards
- The security and design considerations of a National digital Id
- Big data management and analytics for a National digital Id system
- Risk and mitigation consideration for a National digital Id project

All activities and questions are optional: you may skip any part of this study that you do not wish to complete and may stop at any time.

If you need to complete the activities above in a different way than I have described, please let me know, and I will attempt to make other arrangements.

Risks

There are no foreseeable risks or discomforts associated with this study. You can still skip any question you do not wish to answer, skip any activity, or stop participation at any time.

Benefits

If you participate, there are no direct benefits to you. This research may increase the body of knowledge in the subject area of this study.

Privacy and Data Protection

I will take reasonable measures to protect the security of all your personal information, but I cannot guarantee the confidentiality of your research data. In addition to me, the following people and offices will have access to your data:

- My NCU dissertation committee and any appropriate NCU support or leadership staff
- The NCU Institutional Review Board

This data could be used for future research studies or distributed to other investigators for future research studies without additional informed consent from you or your legally authorized representative.

I will securely store your data for 3 years. Then, I will delete electronic data and destroy paper data.

How the Results Will Be Used

I will publish the results in my dissertation. I may also share the results in a presentation or publication. Participants will not be identified in the results.

Recording

I would like to record your responses with a voice recorder or Zoom during the interview.

Compensation

After you complete the review of your interview response script, you will receive a $30 (or 30 USD) equivalent compensation via email.

Contact Information

If you have questions, you can contact me at: d.jembere5435@o365.ncu.edu.

My dissertation chair's name is Dr. Will Tribbey. They work at Northcentral University and are supervising me on the research. You can contact them at: wtribbey@ncu.edu.

If you have questions about your rights in the research or if a problem or injury has occurred during your participation, please contact the NCU Institutional Review Board at irb@ncu.edu or 1-888-327-2877 ext. 8014.

Voluntary Participation

If you decide not to participate, or if you stop participation after you start, there will be no penalty to you: you will not lose any benefit to which you are otherwise entitled.

Appendix D

Interview Protocol

For consistent execution of the interview procedure, the following interview protocol will be followed throughout the interview process for each participant.

1. Confirm that the participant has received and reviewed the informed consent letter, and then obtain verbal consent before the interview.
2. Create a relationship with each participant by providing precise and consistent information about the purpose of the study. Express the researcher's appreciation of the individual's participation, the significance, and the relevance of their contribution to the study.

   My name is Denekew A. Jembere, and I am a doctoral student at Northcentral University (NCU). I am conducting a research study to explore, understand and establish a standardized data harmonization framework for integrating a national identity database with disparate digital identity data sources. Thank you so much for voluntarily deciding to participate in this study. I appreciate your participation and would like to inform you that your contribution to this study is relevant to the significance of the study.

3. Check and ensure that the participant's questions and concerns about the interview process are addressed before starting the interview.
4. Use a distinct notepad designated by the participant's pseudonym to capture written notes about the participant's reactions during the interview process (see Appendix F). Audio record the pseudonym to map the written note and the participant's audio response.
5. For each participant, use the participant's pseudonym to identify and associate the audio record and the written notes notepad of each interview.
6. Encourage each participant to speak freely and share their experiences.
7. Perform the interview, asking questions one after the other and the associated follow-on questions for any additional clarification.

   For interview questions that need clarifications, use the following follow-on questions:

   a. Would you please tell me more about [… response topic]?
   b. When you said [rephrase or repeat the response], did I understand you correctly?

8. Audio-record each participant's responses and take written notes about their body language and any non-verbal responses.
9. At the end of each interview, thank the participant and express appreciation for their time, participation, and significant contribution to the study. In addition, provide a timetable to share the transcription of their responses for their verification.

Appendix E

Interview Questions

Section 1: National Digital ID Data Integration

1. Data sources: Do you work with integrating National digital ID data from various data sources? If so, how many disparate data sources do you work with?
2. Data source challenges: Did you have any challenges that you experienced in locating national digital ID-related data sources for integration? If so, would you please provide some examples of the challenges in obtaining the data sources for integration and how you overcame them?
3. Data integration challenges: Did you have any challenges that you experienced in integrating the national digital ID-related data sources once the data sources have been identified? If so, would you please provide some examples of the integration challenges and how you overcome them?
4. Data integration benefits: What are your thoughts on the benefits of integrating national digital ID-related data from disparate systems?

Section 2: Security and Design Considerations for Sustainable Development

1. Security requirements: How should national digital ID system users or employees with access to residents' national digital ID data on the service or client-side be held accountable for keeping the data safe?
2. Design considerations: Would you describe the architectural considerations you recommend for developing a national digital ID system?

Section 3: National Digital ID Standards

1. National standards: Would you describe the required national policies, standards, and governance guidelines needed to establish a national digital ID system?
2. Intonational standards: Would you describe the required international standards to develop a national digital ID system?

Section 4: Big Data Management and Analytics

1. Big data analytics solutions: What are the setup considerations you recommend for National digital ID system-related big data analytics solutions?

2. Big data analytics components: What are the critical system components concerning hardware and software needed to establish a national digital ID system-related Big Data analytics system?
3. Big Data storage and management: What are your recommended considerations for the National digital ID system-related Big Data storage and management?

Section 5: Risk and Mitigation Consideration for National digital ID Project

1. Risks and mitigations: Would you please describe the potential risks associated with mitigations for national digital ID data integration projects?
2. Additional considerations: Any additional considerations that you would recommend for stakeholders or individuals considering national digital ID data integration projects?

Appendix F

Notepad for Handwritten Notes

1. Interview order number: _____
2. Interview Date: _____
3. Participant's pseudonym: _____
4. Participant's Organization: _____
5. Participant's Org Role: _____

Q-1: _____
_____
_____
_____
_____
_____
Q-2: _____
_____
_____
_____
_____
_____
Q-3: _____
_____
_____
_____
_____
_____
Q-4: _____
_____
_____
_____
_____
_____
Q-5: _____
_____
_____
_____
_____
_____