Universiteit Leiden
Leiden Institute of Advanced Computer Science (LIACS)

Master's Thesis

A FAIR-Compliant Semantic Interoperability Framework for Maternal Healthcare Data Integration Between Community Health Workers and Referral Hospitals in Kenya

Arya Deshmukh
s4182987

MSc Computer Science
Leiden University

1st supervisor:	Prof. Dr. Mirjam van Reisen
2nd supervisor:	Dr. L. Cao

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Mirjam van Reisen, for her invaluable guidance, encouragement, and continuous support throughout this research. I am especially grateful for the opportunity she provided to participate in the field visit to Kenya. The experience of engaging directly with healthcare professionals and stakeholders significantly shaped this research and provided invaluable insights into the real-world challenges of maternal healthcare interoperability.

I would also like to thank Dr. L. Cao for serving as my second supervisor and for his support throughout this research.

My sincere appreciation goes to Kamukunji Community Empowerment Initiative (KCEI) and Pumwani Maternity Hospital for warmly welcoming us during our visit and generously sharing their time, expertise, and knowledge. Their willingness to discuss their workflows, challenges, and experiences provided the practical foundation upon which this research was built.

I am particularly grateful to Whitney from KCEI and William Nandwa from Pumwani Maternity Hospital for their openness, patience, and valuable contributions during the stakeholder discussions. Their insights into community and hospital maternal healthcare workflows greatly influenced the development and evaluation of the proposed framework.

I would also like to acknowledge all the Community Health Workers, hospital staff, and other stakeholders who participated in discussions and shared their experiences. Their perspectives ensured that this research remained grounded in practical healthcare needs and real-world interoperability challenges.

Finally, I would like to thank my family and friends for their unwavering encouragement, patience, and support throughout my Master's studies. Their belief in me has been a constant source of motivation throughout this journey.

Abstract

Maternal healthcare information is generated across multiple healthcare settings, including Community Health Workers (CHWs) and referral hospitals. Although these systems often collect complementary information about the same patient, differences in data structures, identifiers, workflows, and terminology limit interoperability and continuity of care. Existing interoperability approaches primarily focus on syntactic data exchange through standards such as Fast Healthcare Interoperability Resources (FHIR), while the semantic integration of heterogeneous maternal healthcare data remains a significant challenge, particularly in low-resource settings.

This research presents a FAIR-compliant semantic interoperability framework for integrating maternal healthcare data between Community Health Workers and referral hospitals in Kenya. Using a Design Science Research methodology, the study combines stakeholder engagement, ontology engineering, Ontology-Based Data Access (OBDA), semantic mappings, Resource Description Framework (RDF) materialisation, and knowledge graph technologies to create a shared semantic representation of maternal healthcare information.

A maternal health ontology was developed by extending an existing hospital ontology with concepts representing community-level healthcare data. Relational datasets representing Community Health Unit and hospital information were mapped to the ontology using Ontop without modifying the underlying databases. The resulting RDF was materialised into two independently deployed AllegroGraph repositories, one held by each organisation, with cross-system correspondence links present in both so that information originating from either healthcare setting can be integrated at query time through federated SPARQL without either dataset being merged into the other. The framework was designed to support continuity of care, referral traceability, pregnancy record reconciliation, and cross-system analysis of maternal syphilis screening while adhering to the FAIR principles of Findability, Accessibility, Interoperability, and Reusability.

The proposed framework was evaluated through technical validation, SPARQL-based competency queries, and stakeholder feedback obtained from Community Health Workers and hospital staff. The results demonstrate that ontology-driven semantic integration can effectively connect heterogeneous maternal healthcare datasets while preserving their original structures and supporting meaningful cross-system

querying. The framework further illustrates how semantic technologies can complement existing healthcare interoperability standards by addressing semantic inconsistencies that remain unresolved through structural interoperability alone.

This research contributes a practical semantic interoperability architecture for maternal healthcare in Kenya and demonstrates the potential of FAIR data principles, ontology engineering, and knowledge graph technologies to improve healthcare data integration and continuity of care in fragmented healthcare environments.

Contents

Acknowledgements i

Abstract ii

List of Tables x

1 Introduction 1
- 1.1 Problem Statement . . . 1
- 1.2 Synthesis of Problem Statement . . . 2
- 1.3 Identified Research Gaps . . . 4
- 1.4 Research Objective . . . 4
  - 1.4.1 Sub-Objectives . . . 4
- 1.5 Research Questions . . . 5
  - 1.5.1 Sub-Research Questions . . . 6
- 1.6 Hypothesis . . . 6
- 1.7 Thesis Structure . . . 7
- 1.8 Timeline . . . 8

2 Theoretical Foundation 9
- 2.1 FAIR Data . . . 9
  - 2.1.1 Findable . . . 9
  - 2.1.2 Accessible . . . 9
  - 2.1.3 Interoperable . . . 10
  - 2.1.4 Reusable . . . 10
  - 2.1.5 FAIR in healthcare . . . 10
  - 2.1.6 FAIR-OLR: Ownership, Localisation, and Regulatory Compliance . . . 10
- 2.2 Healthcare Interoperability . . . 11
  - 2.2.1 Foundational Interoperability . . . 12
  - 2.2.2 Structural (syntactic) interoperability . . . 12
  - 2.2.3 Semantic Interoperability . . . 13
  - 2.2.4 Organisational interoperability . . . 13

2.3 Semantic Web Technologies and RDF . . . 14
2.4 Healthcare Ontologies and Vocabularies . . . 14
2.5 Ontology based Data Access and Semantic Mapping . . . 15
2.6 Knowledge Graphs and Triple Stores . . . 16
2.7 Common Data Models and OMOP . . . 16

3 Research Methodology **18**
3.1 Introduction . . . 18
3.2 Research Philosophy . . . 18
3.3 Background: The COMPASS Project . . . 19
3.4 Research Design . . . 19
3.4.1 Design Science Research . . . 19
3.4.2 Justification for Design Science Research . . . 20
3.5 Research Process . . . 20
3.5.1 Phase 1: Problem Investigation and Context Analysis . . . 20
3.5.2 Phase 2: Requirement Analysis and designing a problem statement . . . 21
3.5.3 Phase 3: Semantic Modelling and Architecture Development . 21
3.5.4 Phase 4: Evaluation . . . 21
3.6 Data Sources . . . 22
3.6.1 Community Health Unit Data . . . 22
3.6.2 Hospital Data . . . 22
3.6.3 Secondary Data Sources . . . 22
3.7 Stakeholder Engagement . . . 22
3.8 Architecture Development . . . 23
3.8.1 Mock Data Generation . . . 23
3.8.2 Maternal Health Ontology . . . 23
3.8.3 Ontology Mapping . . . 24
3.8.4 Knowledge Graph Construction . . . 24
3.8.5 Pipeline Orchestration . . . 24
3.8.6 Deployment . . . 24
3.9 Evaluation . . . 25
3.9.1 Technical Evaluation . . . 25
3.9.2 Use case Evaluation . . . 25
3.9.3 Stakeholder Feedback . . . 26
3.9.4 Evaluation Against the Humanitarian Data Space Framework 26
3.10 Ethical Considerations . . . 27
3.11 Use of AI in this research . . . 27

4 Literature Review **28**
4.1 FAIR Healthcare Data and FAIRification . . . 28
4.2 FHIR and healthcare interoperability . . . 29
4.3 Semantic Interoperability and RDF Based Healthcare Systems . . . 30
4.4 Healthcare Ontologies and Controlled Vocabularies . . . 30
4.5 Maternal Healthcare Interoperability in Kenya . . . 31

5 Context Description **33**
5.1 Kenyan Healthcare Environments . . . 33
5.2 Community Health Workers in Kenya . . . 33
5.3 Community Health Centre . . . 34
5.4 Pumwani Maternity Hospital . . . 35
5.5 Maternal Syphilis as a Continuity-of-Care Use Case . . . 35
5.6 Existing Digital Health Infrastructure . . . 36
5.7 Interoperability Challenges . . . 37

6 Current Maternal Healthcare Data Workflows and Interoperability Challenges **38**
6.1 Introduction . . . 38
6.2 Current Data Collection Workflow . . . 38
6.2.1 Community Health Unit (KCEI) . . . 38
6.2.2 Pumwani maternity Hospitals . . . 39
6.3 Structural Comparison of The Two Data Sources . . . 39
6.3.1 No shared identifier . . . 39
6.3.2 Asymmetric contact information . . . 39
6.3.3 A referral field that doesn't link . . . 39
6.3.4 Pregnancy Tracked Twice . . . 40
6.3.5 Syphilis Data . . . 40
6.4 Data Access Constraints . . . 41
6.5 Informal data . . . 41
6.6 Synthesis . . . 41

7 Ontology Design **43**
7.1 Introduction . . . 43
7.2 Requirements Derived from Stakeholders and Analysis . . . 43
7.2.1 Functional Requirements . . . 43
7.2.2 Semantic Interoperability Requirements . . . 44
7.2.3 Non functional requirements . . . 44
7.3 Ontology . . . 45
7.3.1 Origin and Extension Approach . . . 45
7.3.2 Ontology Composition . . . 46

7.4 Functional Requirement Specifications . . . 47
7.4.1 FR1 – Cross System Client Matching . . . 48
7.4.2 FR2 – Syphilis Screening and Diagnosis Traceability . . . 48
7.4.3 FR3 – Referral Traceability . . . 49
7.4.4 FR4 – Pregnancy Record Reconciliation . . . 49
7.4.5 FR5 – Continuity of Care . . . 49
7.5 Conclusion . . . 50

8 Knowledge Graph Construction and Pipeline Implementation 51
8.1 Introduction . . . 51
8.2 Constructing the Synthetic CHU and Hospital Databases . . . 51
8.2.1 Population Design . . . 52
8.3 Ontology Based Data Access (OBDA) . . . 53
8.4 Mappings . . . 55
8.4.1 The Structural Mappings . . . 55
8.4.2 Entity Resolution Mapping . . . 56
8.4.3 Mapping Scope . . . 59
8.5 RDF Materialisations . . . 60
8.6 Pipeline Overview (ETL process) . . . 60
8.6.1 Extraction Process . . . 62
8.6.2 Transformation Process . . . 62
8.6.3 Post-Processing . . . 62
8.6.4 Loading Process . . . 63
8.7 The Knowledge Graph (AllegroGraph) . . . 63
8.8 Resolution of the Non Functional Requirements . . . 64
8.8.1 NFR1 – No Modification to Source Database . . . 64
8.8.2 NFR2 – Rerunnable without Manual Rework . . . 64
8.8.3 NFR3 – Confidentiality through Authenticated Access . . . 64
8.8.4 NFR4 – Extensibility . . . 65
8.8.5 NFR5 – Machine Readability . . . 66
8.8.6 NFR6 – Scalability . . . 67
8.8.7 NFR7 – Localisation . . . 67
8.9 SPARQL based Validation . . . 68
8.9.1 Client Matching Query (FR1) . . . 68
8.9.2 Syphilis Continuity of Care Query (FR2) . . . 68
8.9.3 Pregnancy Linkage Result (FR4) . . . 70
8.9.4 Referral Linkage Results (FR3) . . . 70
8.9.5 Combined Continuity of Care Chain (FR5) . . . 70
8.9.6 Referral Traceability Without a Cross System Match . . . 71
8.10 Operationalising FAIR principles . . . 72
8.10.1 Findable . . . 72

8.10.2 Accessible . . . 72
8.10.3 Interoperable . . . 73
8.10.4 Reusable . . . 73
8.10.5 Ownership . . . 73
8.10.6 Localisation . . . 73
8.10.7 Regulatory Compliance . . . 74

9 Evaluation and Findings **75**
9.1 Introduction . . . 75
9.2 Technical Evaluation . . . 75
9.3 Use Case Evaluation . . . 77
9.3.1 Cross System Client Matching . . . 77
9.3.2 Syphilis Detection . . . 78
9.3.3 Referral Traceability . . . 78
9.3.4 Pregnancy Record Reconciliation . . . 79
9.3.5 End to End Continuity of Care . . . 79
9.3.6 Traceability Without Cross System Correspondence . . . 80
9.4 Stakeholder Feedback . . . 80
9.4.1 Response to the Deployed System . . . 84
9.5 Summary . . . 84

10 Discussion and Conclusion **85**
10.1 Discussion . . . 85
10.1.1 Revisiting The Hypothesis . . . 85
10.1.2 Position Against Literature Gaps . . . 86
10.1.3 Positioning Against the Humanitarian Data Space Evaluation 87
10.1.4 Limitations . . . 91
10.2 Future Work . . . 91
10.3 Conclusion . . . 92

List of Abbreviations **95**

Bibliography **97**

A Field Visit to Kenya **104**
A.1 Sensitive Data Sovereignty Conference . . . 104
A.2 Visit to Pumwani Maternity Hospital and KCEI . . . 105

B Stakeholder Information **107**

C Ontology Development and Knowledge Graph **108**

D Implementation Repository **111**

E Research Ethics and Consent **112**

F AllegroGraph Repository Access **115**

F.1 FAIR Data Point Registration 115

F.2 Africa Health Data Space 116

F.3 Demonstration Video 116

List of Tables

6.1 Structural gaps and the requirements that resolve them 42

8.1 H2 Tables Constructed for Each Data Source 52
8.2 Mapping File Summary 54
8.3 Fields and Tables outside current mapping scope 59

9.1 Technical Evaluation Summary 76

10.1 Position Against the Eight Step Data Transformation Process 89

B.1 Stakeholders consulted during the field visit. 107

Chapter 1

Introduction

1.1 Problem Statement

Healthcare systems are increasingly transitioning toward digitisation via the adoption of Electronic Health Records (EHRs), Health Information Systems (HIS), and interoperable healthcare technologies. These systems generate large volumes of sensitive healthcare information including laboratory observations, diagnoses, medications, referrals, antenatal records, and patient histories. Because healthcare information contains highly sensitive personal and clinical data, healthcare systems require secure, privacy preserving, and semantically interoperable approaches capable of supporting continuity of care, healthcare decision making, and long term healthcare data reuse. [1–3]

The increasing digitisation of healthcare systems has intensified the need for healthcare data infrastructures that support interoperability while maintaining accessibility, machine readability, semantic consistency, and privacy preserving healthcare data management. FAIR principles introduced by Wilkinson et al. (2016) emphasize that healthcare data should be Findable, Accessible, Interoperable, and Reusable.[1] Healthcare systems using FAIR have better interoperability, machine readability, healthcare analytics, and prove to have long term reuse of healthcare information across institutions. Recent FAIRification research within healthcare environments has additionally demonstrated the increasing use of semantic technologies such as RDF, linked data approaches, ontologies, and standardized vocabularies for enabling machine readable healthcare ecosystems and interoperable clinical data infrastructures. [2, 3]

Despite these developments, healthcare interoperability remains a major challenge because healthcare information frequently exists across fragmented and disconnected systems that use different workflows, identifiers, structures, reporting formats, and healthcare terminology. Although standards such as HL7 and Fast Healthcare Interoperability Resources (FHIR) support structured healthcare data

exchange, many healthcare systems continue to experience semantic interoperability limitations where healthcare information may technically be exchanged but fails to preserve consistent conceptual meaning between systems. Research on healthcare interoperability therefore increasingly emphasizes the importance of semantic technologies, RDF based healthcare representation, linked data, standardized vocabularies, and FAIR compliant semantic infrastructures for enabling meaningful healthcare data integration and machine readability. [4]

These interoperability challenges are particularly significant within maternal healthcare environments where continuity of care, referral coordination, antenatal monitoring, laboratory observations, and follow up care are essential for early detection and management of maternal health conditions such as syphilis and other pregnancy related complications. Maternal healthcare information is frequently generated across multiple disconnected healthcare environments including community healthcare systems, Community Health Workers (CHWs), referral systems, and hospital clinical infrastructures, specially in low resource countries like Kenya. Community Health Workers play a central role within Kenya's maternal healthcare ecosystem by conducting household visits, antenatal monitoring, maternal health education, referrals, and followup care within local communities. CHWs frequently serve as the first point of contact for pregnant women within community settings and collect important maternal healthcare information including pregnancy status, maternal danger signs, antenatal attendance, and referral information. [5, 6]

1.2 Synthesis of Problem Statement

Although FHIR improves syntactic interoperability, it doesn't completely address semantic interoperability challenges between community generated maternal healthcare records and referral hospital generated clinical records. Community Health Workers and hospitals frequently collect related maternal healthcare information using different structures, terminologies, workflows, and reporting approaches. This therefore, limits effective integration and reuse of healthcare information across the maternal care pathway.

Furthermore, FAIR principles and semantic technologies such as RDF, ontologies, linked data, and knowledge graph databases remain underutilized within maternal healthcare interoperability research, particularly in low resource healthcare environments.

This problem belongs to a broader, actively studied class. How data generated outside a clinical encounter should be trusted and integrated once it reaches a health facility's own record. Kawu et al. (2024) studied Patient Generated Health Data (PGHD) within the VODAN Africa research programme. It was found that clinicians only trust externally generated data once its quality and provenance can be checked.

And they also explicitly recommend that such data be kept separated from facility collected clinical data rather than merged into it, with the connection between the two preserved and traceable[7]. Their own case illustrates why this still matters. The VODAN Africa system itself supported only data collected within the facility, from outpatient and antenatal registers, with data generated outside the facility left unaddressed[7].

A CHW's household observation is not patient generated in the sense Kawu et al. study. But, it hares the same structural problem. Information produced outside the hospital's own clinical workflow. The value of this information depends on it being verifiably, traceably connected to what the hospital later records, rather than blended into it without record of its origin. This research addresses that same class of problem for a specific case. Community Health Worker (CHW) recorded maternal health observations in Kenya and their traceable reconciliation with hospital generated clinical records. A reconciliation this thesis's architecture performs by keeping CHU and hospital data in genuinely separate repositories, connected only through explicit, individually traceable correspondence links(Section 8.4.2).

This fragmentation problem sits inside a larger and more urgent concern. Patient data is sensitive data, and who is permitted to access it, and from where is not a settled question. In December 2025, Kenya and the United States signed a Health Cooperation Framework and a Data Sharing Agreement. An earlier draft reportedly gave the United States broad access to Kenya's national health data under US federal law rather than Kenyan law.[8] Public objection forced the signed version to commit instead to sharing only deidentified, aggregated data under Kenya's own Data Protection Act (2019) and Digital Health Act (2023). The agreement remains under active legal challenge in the Kenyan courts at the time of writing.[8–11] This dispute makes concrete what this thesis treats as a design requirement rather than an afterthought. Sensitive patient data should remain in the jurisdiction and under the control of the facility that produced it, with information and analytics generated through visiting that data rather than transferring it. This is the principle FAIR-OLR adds to FAIR as Localisation, and the basis of VODAN Africa's data visiting architecture that is already piloted at Pumwani Maternity Hospital itself (Section 2.1.6,Section 5.4) [12, 13].

Among these interoperability issues, maternal syphilis is used throughout this thesis as an use case through which the proposed architecture is designed, built, and evaluated. Syphillis was used because it sits precisely at the fault line this thesis addresses. A condition that requires both a community level trigger, and a hospital level test. Both are required to be recognised as the same clinical event across two systems that currently share neither an identifier nor a common vocabulary. The clinical and epidemiological basis for this choice is set out in Section 5.5, and the structural evidence from the field visit that confirmed its fit is presented in

Section 6.3.5.

Therefore, this research proposes the design and evaluation of a FAIR compliant semantic interoperability architecture that integrates maternal healthcare data between Community Health Workers and referral hospitals in Kenya through a shared maternal health semantic model. The proposed architecture aims to support continuity of care, referral tracking, maternal risk identification, and syphilis related cross system analysis while preserving semantic consistency, machine readability, and healthcare data usability.

1.3 Identified Research Gaps

The literature reveals several important research gaps:

First, the majority of healthcare interoperability research focuses on hospital to hospital integration, and not on interoperability between community healthcare systems and referral hospitals.

Second, interoperability approaches are heavily dependent on FHIR and syntactic interoperability and not focusing enough on semantic interoperability.

Third, limited research combines FAIR principles, RDF semantic modelling, standardised vocabularies, and knowledge graphs in maternal healthcare systems.

Lastly, low focus on semantic interoperability specifically within maternal syphilis healthcare continuity between different levels of healthcare systems.

This thesis address these gaps and proposes a FAIR compliant semantic interoperable framework for maternal healthcare integration between community health centres and hospitals in Kenya, with a focus on syphilis detection.

1.4 Research Objective

The objective of this research is to design and evaluate a FAIR compliant semantic interoperability framework for integrating maternal healthcare data between Community Health Workers and referral hospitals in Kenya.

1.4.1 Sub-Objectives

A. Understand the current maternal healthcare interoperability challenges within the Kenyan healthcare context

- Analyse the current maternal healthcare workflows used by Community Health Workers and referral hospitals.
- Identify how maternal healthcare data is collected, managed, shared, and reused across community and hospital healthcare environments.

- Identify interoperability challenges, semantic inconsistencies, and continuity of care limitations affecting maternal healthcare data integration.

B. Define the system specifications and interoperability requirements for maternal healthcare data integration

- Identify stakeholder requirements through engagement with Community Health Workers and hospital healthcare staff.
- Analyse the suitability of existing healthcare standards, vocabularies, and semantic technologies for maternal healthcare interoperability.
- Define the functional and semantic interoperability requirements necessary for integrating maternal healthcare data between community and hospital systems.

C. Develop a FAIR-compliant semantic interoperability methodology for maternal healthcare data integration

- Explore semantic technologies such as RDF, linked data, ontologies, and standardized healthcare vocabularies.
- Develop a semantic representation approach for maternal healthcare data integration.
- Design a conceptual interoperability framework and semantic data integration methodology capable of supporting machine-readable and FAIR-compliant maternal healthcare information exchange.

D. Validate and evaluate the proposed interoperability framework through stakeholder engagement

- Assess the usability and relevance of the proposed framework within real healthcare environments.
- Gather stakeholder feedback regarding interoperability, continuity of care, and practical usability.
- Evaluate whether the framework supports semantic interoperability, FAIR compliance, and integration of maternal healthcare information between Community Health Workers and referral hospitals.

1.5 Research Questions

This research aims to answer the following primary research question:

> *How can a FAIR compliant semantic interoperability framework support the integration and usability of maternal healthcare data between Community Health Workers and referral hospitals in Kenya?*

1.5.1 Sub-Research Questions

A. What are the current interoperability challenges affecting maternal healthcare data integration between Community Health Workers and referral hospitals in Kenya? (Chapter 6)

- How is maternal healthcare data currently collected and managed across community and hospital healthcare environments?
- What semantic, technical, and workflow limitations affect continuity of maternal healthcare information?

B. What specifications and interoperability requirements are necessary for integrating maternal healthcare data between community and hospital healthcare systems?(Chapter 7)

- What stakeholder requirements exist regarding maternal healthcare data integration and reuse?
- Which healthcare standards, vocabularies, and semantic technologies are most suitable for this context?

C. How can semantic technologies and FAIR principles be used to support interoperable maternal healthcare data integration?(Chapter 8)

- How can RDF, linked data, and semantic representation approaches support machine-readable maternal healthcare data?
- How can a FAIR compliant interoperability methodology support integration between Community Health Workers and hospital systems?

D. How do stakeholders evaluate the usability and practical value of the proposed interoperability framework?(Chapter 9)

- To what extent does the framework support semantic interoperability and continuity of care?
- How relevant and usable is the proposed framework within real maternal healthcare environments?

1.6 Hypothesis

This research hypothesizes that a FAIR compliant semantic interoperability framework can improve the integration and usability of maternal healthcare data between Community Health Workers and referral hospitals in Kenya.

The use of semantic technologies such as RDF, linked data, and standardized healthcare vocabularies is expected to improve interoperability and reduce fragmentation between community and hospital healthcare systems.

The framework is also expected to support better continuity of care by improving how maternal healthcare information such as antenatal records, referrals, and follow-up data is shared and reused.

The practical value of the framework will be evaluated through interviews and feedback from Community Health Workers and hospital staff involved in maternal healthcare workflows.

1.7 Thesis Structure

Chapter 1 – Introduction
Introduces the research problem, objectives, research questions, interoperability challenges, and study context.

Chapter 2 – Theoretical Foundations
Discusses FAIR principles, healthcare interoperability, semantic web technologies, RDF, healthcare vocabularies, and semantic knowledge graphs.

Chapter 3 – Research Methodology
Describes the Design Science Research approach, stakeholder engagement methods, data collection, semantic modeling approaches, and framework evaluation methods.

Chapter 4 – Literature Review
Reviews existing literature on FAIR healthcare data, FHIR, semantic interoperability, RDF-based healthcare systems, healthcare vocabularies, and maternal healthcare interoperability.

Chapter 5 – Context Description
Explains the Kenyan healthcare environment, Community Health Worker systems, referral hospital context, and interoperability challenges.

Chapter 6 – Current Maternal Healthcare Data Workflows and Interoperability Challenges
Presents the findings from stakeholder interviews and field observations, analyses the current maternal healthcare workflows at the Community Health Unit and referral hospital, and identifies the interoperability challenges that motivate the proposed framework.

Chapter 7 – Ontology Design
Presents the design and development of the maternal healthcare ontology, including the extension of the existing hospital ontology, the semantic modelling decisions, ontology mappings, and the shared conceptual model used to achieve semantic interoperability.

Chapter 8 – Knowledge Graph Construction and FAIRification Pipeline
Describes the implementation of the proposed semantic interoperability framework, including the Ontop mappings, RDF materialisation, AllegroGraph knowledge graph, and the end-to-end FAIRification pipeline used to integrate the Community Health Unit and hospital datasets.

Chapter 9 – Evaluation and Findings
Presents the technical validation, SPARQL competency query results, stakeholder

feedback, and evaluates the proposed framework against the research objectives and interoperability requirements.

Chapter 10 – Discussion and Conclusion
Discusses the contributions of the research, reflects on the findings and limitations, and outlines recommendations for future work on FAIR-compliant semantic interoperability in maternal healthcare.

1.8 Timeline

- **September 2025 – December 2025** – Participation in the COMPASS project and DSIP course. Initial exploration of FAIR data, healthcare interoperability, semantic technologies, and maternal healthcare systems.
- **January 2026** – Research visit to Kenya, including visits to Kamukunji Health Centre and Pumwani Maternity Hospital. Meetings and discussions with Community Health Workers, hospital staff, and healthcare stakeholders. Attendance at the Sensitive Data Sovereignty Conference at Tangaza University.
- **February 2026** – Literature review on FAIR principles, healthcare interoperability, RDF, linked data, semantic web technologies, and maternal healthcare interoperability.
- **March 2026** – Analysis of maternal healthcare workflows, ANC data structures, interoperability challenges, and stakeholder requirements.
- **April 2026** – Semantic modelling, ontology design, semantic mapping, RDF representation development, and selection of healthcare vocabularies and standards.
- **May 2026** – Development of the FAIRification pipeline, RDF transformation process, interoperability framework, and setup of the AllegroGraph triple store
- **June 2026** – System testing, interoperability assessment, FAIR compliance evaluation, stakeholder interviews, usability evaluation, and framework assessment within healthcare workflows.
- **July 2026 – August 2026** – Thesis writing, revision, final evaluation, discussion of findings, and thesis submission preparation.

Chapter 2

Theoretical Foundation

The foundation of this thesis is based on FAIR data principles, healthcare interoperability, semantic web technologies, RDF, healthcare vocabulary, and semantic knowledge graph.

2.1 FAIR Data

FAIR data principles were introduced in 2016 as guidelines for improving data management.[1] It stands for Findable, Accessible, Interoperable, and Reusable. What that basically means is that data should be easily found, accessed, combined with other data, and reused by both humans and computers. This improves transparency, collaboration, and the long term impact of work. The principles were developed to enhance both human and machine interaction with data resources and have since become widely adopted within healthcare and biomedical research.

2.1.1 Findable

Findable means a dataset or an individual record within it must carry a globally unique, persistent identifier, and be described by rich enough metadata that a person or a machine can locate it without already knowing it exists.[1] In a maternal healthcare context this means a specific client's record, or a specific referral, should be addressable on its own, rather than only discoverable by manually searching a database export. This is the function a URI performs in RDF (Section 2.3), and it is why every entity produced by this thesis's architecture is identified by one rather than by a row number local to a single table.

2.1.2 Accessible

Accessible means that both humans and computers can retrieve data using standardized, open protocols. It ensures that while data must be "as open as possible, as

closed as necessary", the access conditions and authorization protocols are always clear. [1] Accessible does not mean unrestricted. For sensitive healthcare information, the function this pillar performs is making the access control itself part of the data infrastructure, not an afterthought bolted onto a public dataset.

2.1.3 Interoperable

Interoperable means that research data and metadata must use formal, standardized, and machine actionable languages so that they can be easily integrated, exchanged, and combined across different systems, platforms, and applications. [1] This is a stronger requirement than simply being able to open a file. Two systems can exchange a file successfully (Section 2.2.3) while still disagreeing on what a given field means. The Interoperable pillar is what a shared ontology and a common RDF representation are built to satisfy, which is the specific function Chapter 7's maternal health ontology performs for this thesis.

2.1.4 Reusable

Reusable means data and metadata are described with enough provenance, licensing, and context that they can be correctly understood and reused by someone who was not involved in producing them.[1] This is the pillar most dependent on the other three actually being met. Data that is Findable, Accessible, and Interoperable but carries no record of where it came from or under what terms it may be used is still not genuinely reusable.

2.1.5 FAIR in healthcare

In healthcare, FAIR data principles improve data sharing, interoperability, analytics, and secondary healthcare research. This is particularly important in maternal healthcare data, because data is generated across multiple levels. In such a system, records may exist independent of each other with different syntactics. FAIR hence helps by providing a conceptual foundation for integration of these healthcare data management.

2.1.6 FAIR-OLR: Ownership, Localisation, and Regulatory Compliance

The four FAIR pillars describe how data should be found, accessed, combined, and reused. They say comparatively little about where data is permitted to physically reside, who retains authority over it once it has been shared, and whether the infrastructure handling it actually complies with the legal framework governing the

jurisdiction it was collected in. For sensitive personal health data, this is not a minor omission. FAIR-OLR extends FAIR with three further principles, developed within the VODAN Africa research programme[14] to close exactly this gap.[15]

Ownership under FAIR-OLR is deliberately not framed as belonging to the patient alone. Van Reisen et al. describe it as federated. The health facility that produces the data and the data subject the record describes both hold the possibility of controlling which data or metadata is shared. Localisation requires that data be held in data handling sources under the local control of the facility that produced it and remains responsible for it, rather than in infrastructure external to that facility. Regulatory Compliance requires that the jurisdiction's regulatory requirements, where the data is produced are respected.[14]

Jati et al. (2025) apply FAIR-OLR to Indonesia's SATUSEHAT platform, a national health data exchange system that combines millions of medical records. Their assessment found the clinic's existing SATUSEHAT integration satisfied almost none of them. No persistent identifiers, no metadata of any kind, no licensing information, and critically no way for the clinic or the patient to control, update, or delete data once submitted to the national platform. Their proposed prototype addressed this by giving the clinic its own FAIR Data Point and its own local AllegroGraph repository. After the patient data remains under the clinic's own control and is only submitted to SATUSEHAT as a secondary, regulatorily required step rather than being the clinic's sole copy of its own records.[15]

The Ownership and Localisation principles specifically are operationalised within VODAN Africa through a data visiting architecture. A data editor is deployed at each location where data is produced. Data is entered once, and it remains stored locally at that site's own repository rather than being aggregated into a single, externally controlled store.[14] This is the precise architectural pattern this thesis's own two repository design (Section 8.7) follows.

2.2 Healthcare Interoperability

Healthcare interoperability is the ability of different healthcare systems, applications, and organisations to exchange and use healthcare information. It allows sharing of data across different platforms and healthcare entities while preserving its integrity, quality, and usefulness. This affects continuity of care, research, and facilitates healthcare across multiple healthcare settings.

Healthcare interoperability has become increasingly important with the widespread adoption of Electronic Health Records (EHRs), Health Information Systems (HIS), laboratory information systems, and digital health platforms. In today's healthcare environments, patient information is rarely stored with a single system. Rather, data is generated by different healthcare providers, like hospitals and community

health workers. Interoperability helps to integrate these heterogeneous data sources into a coherent and usable healthcare record.

The Healthcare Information and Management Systems Society (HIMSS) describes interoperability as the ability of different information systems to exchange data accurately, effectively, and consistently while preserving the meaning of the exchanged information.[16] Achieving this requires not only technical connectivity but also standardised data structures, common terminologies, and shared semantic understanding.

Healthcare interoperability is generally classified into four levels: foundational interoperability, structural (syntactic) interoperability, semantic interoperability, and organisational interoperability. Each level has additional capabilities that build upon the previous one.

2.2.1 Foundational Interoperability

This is the most basic level of interoperability. It allows a healthcare information system to send data to another through communication networks and standard communication protocols. At this level, the receiving system is capable of receiving the data but doesn't require to understand it.

2.2.2 Structural (syntactic) interoperability

Structural interoperability defines the format, syntax and organization of data exchange (e.g., message format standards) to the level of data field-by-field results. Syntactic interoperability is mainly achieved by healthcare standard like Health Level Seven (HL7), Clinical Document Architecture (CDA), and Fast Healthcare Interoperability Resources (FHIR). They specify common data structures, message formats, and protocols that are required for systems from different organisations to exchange information.

Syntactic interoperability addresses how information is exchanged rather than what it means. Different systems may exchange data, but may not be able to interpret it correctly. For example, CHWs may record maternal health observations using local terms like high fever, or swollen body parts. Hospitals will use clinical terms and diagnostic terms, like RPR/VDRL test recorded as "positive" or "reactive". Although data systems might successfully exchange these data, this would affect meaningful integration.

Therefore, structural interoperability can't solve the issue of data integration between heterogeneous healthcare systems.

2.2.3 Semantic Interoperability

HIMSS defines semantic interoperability as a "provider for common underlying models and codification of the data including the use of data elements with standardized definitions from publicly available value sets and coding vocabularies, providing shared understanding and meaning to the user."[16]

Unlike syntactic interoperability, semantic interoperability ensures that when data is sent through a system, its meaning is preserved and consistently interpreted by both the sending and the receiving system. This requires the underlying data elements to be machine processable concepts rather than free text or locally defined codes. To achieve this we use standardised clinical terminologies and coding systems such as SNOMED CT, ICD-10/11, and LOINC, combined with semantic web technologies such as RDF and OWL that allow those concepts to be linked and reasoned over. Therefore semantic interoperability is considered a higher level of interoperability.

FHIR has become one of the most widespread used interoperability standard because it organises the healthcare information as Patient, Observation, Encounter, Condition, Procedure, and Medication. These can then be exchanged using standard web technologies and APIs, making healthcare integration easier than previous standards. FHIR resources also include coded data elements, such as the CodeableConcept type used in fields like Observation.code or Condition.code. These can be bound to a standardised terminology such as SNOMED CT, LOINC, or ICD. This terminology binding gives FHIR the technical capacity to carry semantic meaning not just structure. However this binding is optional and depends entirely on implementation. A system only achieves semantic interoperability through FHIR if the party populating a resource actually maps the recorded observation to a coded term, rather than leaving it as free text. When this mapping is not done consistently, the exchanged data remains structurally standardised but semantically ambiguous.[17]

This difference between semantic and syntactic interoperability is crucial in this thesis. Both CHU and hospital system are technically capable of exchanging information, the systems can't automatically interpret it, unless both are mapped into a shared semantic model. This research addresses that gap and develops a maternal health ontology that maps both CHU and hospital datasets.

2.2.4 Organisational interoperability

Organisational interoperability as defined by HIMSS, refers to "governance, policy, social, legal and organizational considerations to facilitate the secure, seamless and timely communication and use of data both within and between organizations, entities and individuals. These components enable shared consent, trust and integrated end user processes and workflows."[16]

In Kenya, national frameworks such as the Digital Health act and the Data Protection Regulations establish requirements for secure and lawful handling of healthcare information across organisations.[18, 19]

The four levels of interoperability discussed above demonstrate that meaningful health data integration needs more than basic connectivity. As discussed, standards such as FHIR are not enough, as they don't address semantic and organisational differences. This thesis primarily focuses on semantic interoperability between CHU and hospital maternal healthcare data.

2.3 Semantic Web Technologies and RDF

Semantic Web Technologies was first introduced by Berners-Lee, Hendler, and Lassila. The goal was to move from a web of linked pages to a web of linked data, where data points, not the whole document, can be connected and queried.[20]

The core idea was Resource Description Framework, RDF. RDF represents information as a 3 part statement (a subject, predicate, and an object) called a triple. [21] For example, "Client 01 hasDiagnosis Syphilis" is a single RDF triple that describes a diagnosis. Each subject and predicate is usually identified using an URI, which gives it a unique and unambiguous identity. This makes it different from a relational tables with a column, which is only relevant in that database.

RDF are useful for healthcare integration because they don't require every dataset to share the same fixed schema before they are combined. Two datasets in a RDF, even if they came from different data schema like CHU and hospital data schema can be merged by simply loading their triples into the same graph. Share entities (clients and patients) become common connection points between the two datasets. This differs from joining relational database where join only works if both have identical keys and structure.[22]

Ontology is RDF schema and the Web Ontology Language (OWL), built on RDF by allowing classes, properties, and logical relationships between them defined.[23] RDF can then be queried by using SPARQL, a query language built for graph structured data. RDF, OWL, and SPARQL are the backbone of semantic web based systems, including this thesis, where the maternal health ontology is represented in OWL, and the resulting data is represented and queried as RDF.[24]

2.4 Healthcare Ontologies and Vocabularies

An ontology is a formal, explicit specifications of the concepts within a domain and the relationships between them [25]. In healthcare, ontologies and vocabularies ensure that when two systems are talking about a concept, like a diagnosis, they are referring to the same thing.

A large number of standardised clinical vocabularies already exist. SNOMED CT is the most comprehensive, covering diagnoses, symptoms, procedures, and clinical finding in detail, ICD-10 is used worldwide for disease classification, and LOINC is used for laboratory and clinical observations. [26–28] Vocabularies like these allow diagnosis in one hospital system to be interpreted and used in another.

However, in practice, adapting to these vocabularies is not always realistic. They are huge in size, complex, and designed for clinical coded terms. That would fit the hospital data well, but for CHU it won't reasonably fit well given how they record their information. Mapping data recorded by CHWs onto a vocabulary like SNOMED CT would mean forcing informal community observations into a coding system that was never designed for informal terms.

Hence, this thesis builds on an existing hospital ontology developed by another student. This thesis adds on classes and properties specific to the Community Health Unit(CHU) setting. This means that the resulting shared model already has a clinically grounded structure for hospital side concepts (patient encounters, visits, and diagnostic records) added onto CHU specific concepts (household registration, community health surveillance, and community level referrals) as a distinct layer. The CHU vocabulary isn't forced on the clinically coded vocabulary, and the hospital data isn't forced into the informal community terms. Rather both sides retain their own natural representation, connected through shared classes and properties.

2.5 Ontology based Data Access and Semantic Mapping

Most healthcare data is stored in relational database, and converting them directly into RDF is quite impractical. Ontology-based Data Access (OBDA) defines a set of mappings describing how existing relational database tables and columns correspond to classes and properties in an ontology, without changing the original data. Once these mappings are defined, the relational data can be queried or converted into RDF through materialisation.

Ontop is one of the widely used implementations for this process.[29] It takes a relational database, an ontology, and a set of mappings and uses them to materialise the mapped data into RDF file. Where several producers each hold their own data, a separate mapping file can be materialised per producer against the same shared ontology, producing one RDF graph per producer rather than a single combined one. For example, a mapping file may describe that each row in PATIENT table corresponds to the PATIENT class in the ontology and other columns such as name and date of birth are its data properties. Because the mapping is external to the original database, the system can keep operating, and the semantic mapping file can

work alongside it.

This mapping approach is useful when integrating data from multiple, structurally different databases, since each one can be mapped independently onto the same shared ontology without changes to the original system.

2.6 Knowledge Graphs and Triple Stores

Once RDFs are generated, they can be organise into a knowledge graph. A knowledge graph is a network of entities connected by relationships, built from RDF triples, where any entity can be linked to another entity, given a relationship exists. This network doesn't require a pre-defined, rigid, schema. This flexibility makes a knowledge graph well suited for data drawn from multiple sources. New entities, relationships, or new data systems can be accommodated without redesigning the underlying structure.

Storing and querying RDFs require a dedicated triple store, a database built for graph structured data and typically queried via SPARQL. A number of triple store implementations exist, AllegroGraph being one. AllegroGraph alongside storing and querying RDF triples, also has visual graph exploration tools, making it possible to trace relationships across a graph.[30]

Knowledge graphs are used because hospital and community health information tend to be relational by nature and representing them explicitly rather than reconstructing them makes it considerably easier to query and visualise.

A knowledge graph does not have to reside in a single store. SPARQL supports federated querying through the `SERVICE` keyword. This allows a query executed against one endpoint to delegate part of its graph pattern to a second, remote endpoint and join the results. This matters for healthcare data specifically because it means two organisations can each retain and control their own repository while a question spanning both can still be answered. No copy of either dataset has to be transferred to a central store for the query to run. This is the query time counterpart to the Localisation principle described in Section 2.1.6. It is also the mechanism this thesis uses to keep the Community Health Unit and hospital graphs separate while still answering questions that cross them.

2.7 Common Data Models and OMOP

Common data models like OMOP are an important feature of healthcare interoperability literature.

OHDSI community research show that OMOP improves structural harmonisation and helps large scale healthcare analytics. OMOP help transforms heterogeneous healthcare datasets into standardized structures for research and analytics.[31]

However, literature shows that structural standardisation by itself doesn't fully resolve semantic interoperability problems. OMOP harmonises healthcare schemas, but it doesn't explicitly represent semantic relationships between fragmented healthcare entities.

Therefore, there is an increasing need for structural harmonisation combined with semantic technologies, controlled vocabularies, and linked data approaches.

Chapter 3

Research Methodology

3.1 Introduction

This thesis looks at FAIR data architecture and semantic interoperability between ANC data in Community Health Units (CHUs) and referral hospitals in Kenya. It tries to solve the challenge of a fragmented healthcare system, followed continuity of care, referral tracking, and identifications of diseases such as syphilis.

To solve these challenges, I use a Design science research (DSR) method. It is suitable here because the goal is not only to understand the problem, but also develop and evaluate the architecture to solve the problem. I develop a FAIR-compliant semantic interoperability architecture consisting of a maternal health ontology, ontology based mappings, RDF representations, and a knowledge graph infrastructure.

This chapter describes the research philosophy, research design, data collection methods, stakeholder engagement process, architecture development process, and evaluation strategy.

3.2 Research Philosophy

Research philosophy are a set of assumptions and beliefs underlying the way we approach our study.

In this research I adopted a pragmatist philosophy. It solves practical problems through application of appropriate methods and technologies, like technical methods, stakeholder engagement, and evaluation for real world problems, rather than sticking to a single philosophical perspective. The purpose of this research is not purely theoretical, but I plan on aiming to develop a practical solution for improving the maternal healthcare system within Kenyan healthcare by applying proper interoperability between CHUs and hospitals. Further, it will be evaluated by the usefulness and applicability of the architecture rather than comparison to some theoretical baseline.

3.3 Background: The COMPASS Project

The work presented in this thesis was preceded by a related project carried out before the research began. Between September and December 2025, as part of the Data Science in Practice course at Leiden University, I along with Lina Dadic (the author from whom the hospital ontology is inherited) worked on COMPASS (Contextual Operations Multi-Point Access to Secure SPARQL).

COMPASS addressed a problem adjacent to the one this thesis takes up. Health data in Kenya is generated across more than nine thousand facilities running one of several electronic medical record systems, and reaches the Ministry of Health only as aggregated monthly reports transmitted manually. There is no central access point through which record level data can be queried. There is also no means of controlling or tracing access to it once a report has been sent. Centralizing that data is not an option, as doing so would remove it from the facilities that own it. The project therefore designed a federated access point through which data could be queried while remaining physically at its source, governed by explicit access control. Withing that project my contribution was the FAIRification pipeline connecting a relational schema to a triple store through Ontology Based Data Access.

Two things carried forward into this thesis. First is the technical direction. COMPASS compared several relational to RDF mapping tools and selected an OBDA approach using Ontop. It established local storage with federated querying as the way to reconcile access with ownership. Both decisions are inherited here, and both are argued for independently in Chapter 8. The second is the problem framing. COMPASS treated data sovereignty as a design constraint rather than a compliance matter, which is the position this thesis takes in Section 2.1.6.

What did not carry forward is any of the implementation. The pipeline described in Chapter 8 was written from scratch for this research. No code, mapping, schema, or dataset was reused. COMPASS is therefore prior context that shaped the approach taken here, not a component of the architecture this thesis presents.

3.4 Research Design

3.4.1 Design Science Research

This study uses Design Science Research (DSR) as the primary research methodology. It focuses on the creation and evaluation of the architecture designed to solve organizational problems. Unlike theoretical research which focuses on understanding phenomena, DSR aims on solving existing situations.

3.4.2 Justification for Design Science Research

There were three primary reasons for choosing this research design:

First, it identifies a clear defined practical problem. The lack of semantic interoperability between CHU and hospital data causes genuine issues with continuity of care and delays the proper care and detection of diseases. Second, the research creates a working architecture rather than just theoretical knowledge and ideas. And finally, the success and usefulness can be evaluated through real word users and stakeholders.

The methodology hence enhances the real world application and usefulness of both the objectives and expected outcomes of the study.

3.5 Research Process

This research was done in 4 phases

3.5.1 Phase 1: Problem Investigation and Context Analysis

The first phase started with visiting Nairobi, Kenya to attend a conference and visit the Kamukunji Community Empowerment Initiative (KCEI), and Pumwani hospital and conduct interviews to understand the maternal healthcare environment and identify existing challenges.

The research was also informed by participation in the Sensitive Data Sovereignty Conference held at Tangaza University in January 2026. The conference focused on FAIR data, healthcare interoperability, data sovereignty, and privacy preserving healthcare systems within African healthcare environments. Discussions and presentations during the conference provided additional insight into real world interoperability challenges, FAIRification approaches, and the importance of semantic technologies within healthcare systems in low resource settings.

Further literature review was done to understand the general problem with data handling and sharing. With the help of the interviews and literature review I could understand and analyse the maternal healthcare workflow, and analyse the existing standards.

The goal of this phase was to establish a clear understanding of the interoperability challenges affecting maternal healthcare data exchange between CHUs and hospitals, and review FAIR principles and semantic technologies.

3.5.2 Phase 2: Requirement Analysis and designing a problem statement

In the second phase, the focus was on identifying functional and semantic requirements for the interoperability architecture. These came from interviews with CHWs, healthcare professionals, analysing the maternal healthcare workflows, and available healthcare datasets and literature.

The main focus points were continuity of care, referral tracking, maternal risk identifiers, maternal risk identifiers, and data interoperability. Within this, maternal syphilis was selected during this phase as the specific clinical use case for the architecture. This selection was not made in advance of the fieldwork, but emerged from it. It was the point at which the interview findings from CHWs and hospital staff, the structural comparison of the two organisations' data schemas, and the literature gap identified converged on the same condition (Section 1.3). This converged on the same condition, providing both a real world reason for the two systems to need to communicate and a concrete, testable requirement against which the proposed architecture could be evaluated.

This led to the generation of the functional, semantic, and non-functional requirments and the foundation of the design of the semantic model and architecture.

3.5.3 Phase 3: Semantic Modelling and Architecture Development

In the third phase, I focused on development of the architecture for semantic interoperability.

I studied the data schema for both CHU and hospital, and created the databases for both. Then, I developed a maternal health ontology in Protégé, defined the semantic relationships. The ontology acted as the shared semantic model between the healthcare datasets. Then I did some ontology based mapping on Ontop. Finally I created RDFs and developed knowledge based graphs in AllegroGraphs.

3.5.4 Phase 4: Evaluation

The final phase included the demonstration and evaluation of the architecture.

It included RDF materialization validation, knowledge graph validation, interoperability demonstration, collection feedback from stakeholders, and assessing and evaluation FAIR compliance. The focus was on determining whether the designed framework was effective and useful in addressing the identified challenge.

3.6 Data Sources

3.6.1 Community Health Unit Data

The first data source was the Community Health Unit dataset from Kamakuji, Kamukunji Community Empowerment Initiative (KCEI) in Nairobi.

The CHU schema had information relating to CHUs, household registration, individual clients, service delivery activities, health surveillance, referrals, maternal and child health, environmental and WASH indicators. The dataset represented the information collected by Community Health workers during their routine visits in the community.

3.6.2 Hospital Data

The second data source was the maternal healthcare information dataset from Pumwani hospital also in Nairobi.

This dataset included information about patients, pregnancy episodes, ANC visits, diagnosis, laboratory screening and treatments. This provided clinical perspectives compared to the community generated healthcare information.

3.6.3 Secondary Data Sources

Secondary sources include published literature, healthcare interoperability standards, healthcare vocabularies, FAIR implementation guidelines, and Semantic Web specifications. These influence the theoretical and technical aspects of the research.

Since real patient data from either organisations couldn't be accessed for this research, both datasets are mock datasets built to mirror their actual structure.

3.7 Stakeholder Engagement

Stakeholder participation was a crucial part of this thesis. It adopted a user centric approach and the stakeholders contributed with requirement identification, providing data schema, and evaluation of the architecture.

Stakeholders included Community health workers, and hospital administrators. The discussions were conducted through face to face and online interviews for discussions and feedback.

The main objective for stakeholder engagement was to understand the current workflow, identify the interoperability challenges, define requirements, and evaluate the proposed solution. The input from them insured the developed architecture addressed real world healthcare needs and requirements.

3.8 Architecture Development

The architecture developed for this research consists of 4 layers, a relational data layer holding synthetic CHU and hospital data in H2, a shared maternal health ontology developed in Protégé, an Ontology-Based Data Access mapping layer implemented in Ontop, and a knowledge graph layer consisting of two independently deployed AllegroGraph repositories, one for each organisation, tied together by a base pipeline that runs the process end to end.

3.8.1 Mock Data Generation

Access to real patient data was not available for this research. The records held in KCEI's mobile application and in Pumwani's KenyaEMR system are confidential clinical data, and working with either live system would not have been ethically justifiable. Both datasets used throughout this thesis are therefore synthetic. They are mock records constructed to mirror the structure and content of the real systems rather than extracts from them.

What was replicated is the schema rather than the data. The CHU side was built to mirror the complete structure of KCEI's mobile application data model. The hospital side was derived from the MOH 405 Antenatal Care Register and KenyaEMR, but reduced rather than reproduced in full. The real register carries fields such as breast examination findings and FGM status that fall outside this research's scope, so only the columns relevant to antenatal care, referral tracking, and syphilis screening were retained. Both schemas were implemented as relational tables in a local H2 database.

The mock records were generated with known ground truth. Rather than populating both sides with records that all correspond, a proportion were deliberately constructed not to match. This ensured that the architecture could be assessed on whether it correctly declines to link records as well as on whether it correctly links them. The specific composition of the dataset, and the reasoning behind each of these choices, is set out in Section 8.2 and Section 8.2.1.

3.8.2 Maternal Health Ontology

The ontology was developed on Protégé. Rather than creating it from scratch, it was extended from an existing hospital ontology, developed by a fellow student, Lina Dadic working on interoperability between two hospitals. [32] That work is unpublished at the time of writing, hence the ontology was obtained directly from the author in Turtle format. Because it is not publicly available, the class and property counts reported for it in this thesis were verified against the file as supplied rather than taken from a published description. Classes and properties relevant to

CHU level data were added to this ontology. This provided a shared semantic representation of maternal healthcare concepts from both the hospital's existing clinical structure and the community level concepts this research added to it.

3.8.3 Ontology Mapping

Ontology mapping was done using Ontop. This was done with the help of the Protégé plugin during the development and testing of mappings, and also as a tool within the automated pipeline. It enabled relational healthcare data to be mapped to ontology classes and transform into triples. This allowed different datasets to be integrated semantically without modifying the existing data systems.

3.8.4 Knowledge Graph Construction

The triples were materialised separately for each organisation and uploaded into its own AllegroGraph repository. Each repository holds only its own organisation's records, together with the correspondence links that connect them to the other side. A unified semantic representation is therefore produced at query time through SPARQL federation rather than by combining both organisations' data into one store.

3.8.5 Pipeline Orchestration

The H2 databases (Section 3.8.1) are connected to the rest of the architecture by a base pipeline, implemented in Python, that runs the process end to end. The pipeline first checks that the expected database tables are present, then validates record counts across the client, visit, laboratory, and referral tables, and then runs materialisation, post processing, and upload once per organisation, producing and deploying the Community Health Unit and hospital graphs independently of one another. Each side's success or failure is tracked separately, so a problem on one side does not prevent the other from completing. Running these steps as a single pipeline means the framework can be rerun as new data is added, rather than functioning only as a one off demonstration.

3.8.6 Deployment

The two knowledge graphs are not held on research controlled infrastructure. Both AllegroGraph repositories are deployed on the participating institutions' own servers. One holding the Community Health Unit graph and one holding the hospital graph, each with its own endpoint and its own access credentials. Therefore, each organisation's data resides physically within the jurisdiction that produced it and under the

operational control of the facility responsible for it. This was a deliberate decision, rather than a hosting decision. This thesis's architecture argues for Localisation (Section 2.1.6). Hence, holding both organisation's data on a researcher's own infrastructure would not have demonstrated the principle it proposes.

Deployment was carried out by the VODAN Africa engineers rather than by the researcher. The researcher generated the materialised RDF datasets, the accompanying metadata, and the semantic pipeline used to upload the datasets to their respective AllegroGraph repositories. Repository information, including the repository endpoints was also provided to support dataset registration and discovery. These repositories were then deployed within an existing FAIR Data Point (FDP) that forms part of the Africa Health Data Space (AHDS). AHDS is a broader federated ecosystem that connects multiple FAIR Data Points, enabling datasets to be discovered and queried while allowing each organisation to retain control over its own data. In this architecture the AllegroGraph repositories act as the underlying data stores. While the FDP provides the metadata required to make the repositories discoverable. This reflects the intended operating model of the architecture. The researcher provides the semantic artefacts required for deployment, and infrastructure partners manage the operational deployment and long term maintenance of the system.

Deployment details and access points are given in Appendix F.

3.9 Evaluation

The evaluation method consisted of four components.

3.9.1 Technical Evaluation

The technical assessment was to ensure the functionality of the architecture. It included ontology consistency, mapping correctness, RDF generation success, knowledge graph construction and query execution.

3.9.2 Use case Evaluation

The architecture was evaluated using maternal healthcare use cases identified during stakeholder engagement. These included continuity of care, referral tracking, maternal risk identification, and syphilis related analysis. The aim was to demonstrate how semantic interoperability was beneficial to real practical healthcare situations.

3.9.3 Stakeholder Feedback

Stakeholders evaluated the usefulness of the architecture. The criteria for the evaluation were perceived usefulness, usability, clinical relevance, and betterment of healthcare workflows. The feedback collected was the primary assessment of the architecture.

This was done in two rounds. The first round was conducted with members of the co-creation team at KCEI and Pumwani Maternity Hospital who had been involved in the research from the field visit onwards. They were given authenticated access to their own organisation's repository and asked to assess the framework on perceived usefulness, usability, clinical relevance, and whether it would improve existing healthcare workflows. This round took place before any demonstration of the deployed system.

The second round was conducted after a recorded demonstration of the working architecture. The demonstration walks through the localisation of the data in the two locally deployed repositories, the FAIR Data Point, and the AHDS. It explains how these components together enable interoperability, discoverability, and querying across both organisations without the underlying datasets being revealed. The recording was shared the VODAN Africa, KCEI, and Pumwani groups, and responses were collected from the resulting discussions.

Feedbacks were gathered through WhatsApp group discussion rather than individual interview. Responses are therefore reported as WhatsApp Group Feedback loops. Each respondent is assigned a number, and every response drawn on in Chapter 9 is attributed to that number together with the date and the channel it came from.

3.9.4 Evaluation Against the Humanitarian Data Space Framework

The architecture is additionally assessed against an independent evaluation of the data transformation processes within the Humanitarian Data Space, conducted by academic reviewers. [33] This sets out the design decisions any application converting structured input into RDF must resolve, a strengths and weaknesses assessment of the systems evaluated, and a staged process toward a governed, federated data space. COMPASS (Section 3.3) is one of the applications it assesses, as it stood at the end of the course project and before the thesis work that followed it.

Unlike the other three components, this assessment is reported in Chapter 10 rather than Chapter 9. The other three evaluate whether the architecture does what this research set out to build. This one compares it against a framework external to the thesis. Hence, it belongs with the positioning against the literature gaps rather than with the interpretation of results.

3.10 Ethical Considerations

This research is based on maternal healthcare data and hence has important ethical, legal, privacy, and data management considerations. Healthcare data has highly sensitive patient data including antenatal records, laboratory observations, referral information, diagnoses, medications, and maternal treatment histories. Therefore, it must be handled in a secure, privacy preserving ethical way.

This thesis acknowledges the importance of protecting patient confidentiality and hence complies with applicable ethical requirements for healthcare related research. No personally identifiable patient information will be disclosed. Any healthcare data used during development and evaluation will be artificially generated synthetic data and a representation of the real data. Stakeholder participation will be voluntary, and consent will be obtained prior to interviews or feedback sessions.

The study moreover recognises that the healthcare data interoperability has to align with healthcare data protection principles, FAIR data practices, and regulatory requirements relevant to sensitive health information. Therefore, the proposed architecture supports semantic interoperability while maintaining healthcare data ownership, privacy, and restricted accessibility within healthcare environments.

The research involves stakeholder engagement like discussions and interviews with Community Health Workers and hospital staff. Stakeholder participation will be voluntary, and consent will be obtained prior to interviews or feedback sessions. Interviews, research observations, or any documentation will be securely stored and used solely for academic research purposes.

This research additionally follows the broader principles promoted within FAIR healthcare infrastructures and VODAN-Africa regarding responsible healthcare data management, data sovereignty, and privacy preserving interoperability. It will also adhere to the relevant institutional ethical guidelines and applicable healthcare data protection principles throughout the duration of the study.[12, 34]

3.11 Use of AI in this research

Generative AI tools were used in the early stage of this thesis, particularly in the writing of the research proposal. The initial set of research questions were broader and chaotic, numbering 7. AI was used as an organisational tool to restructure them into 4. It was used as a tool to help frame sentences clearly, and reduce overlap.

Chapter 4

Literature Review

Healthcare interoperability is becoming a very important research field due to the rapid increase of digitization of the healthcare systems across the globe. Existing literature shows that significant progress has been made in the healthcare data exchange technologies, semantic technologies, FAIR compliant healthcare data management, and interoperability frameworks. However, there remain gaps in research for semantic interoperability between community healthcare and hospitals, specifically in the maternal healthcare environments in low resource areas like Africa.

4.1 FAIR Healthcare Data and FAIRification

FAIR Guiding Principles paper by Wilkinson et al. [1] plays a major role in healthcare data management. The FAIR framework introduced in the paper provide guidelines for better findability, accessibility, interoperability, and reusability of data. The principles emphasize the importance of machine readable metadata, standard identifiers, semantic interoperability, and reusable data systems.

Since 2016 FAIR has become quite significant in bio medical informatics and healthcare data management. A lot of research has been done on FAIRification of healthcare data systems. The FAIR4Health project showed that healthcare data can be FAIRified with semantic metadata, interoperability standards, and semantic healthcare models [35].

Several studies have been done on specialized healthcare domains with interoperability and FAIRification. For example, Lazarova et al. (2022) did research within FAIR healthcare in cardiovascular and cardiology environments [36]. He developed an interoperable Electronic Health Record system specifically for cardiology wards using HL7 CDA, HL7 V2, and FHIR standards to improve integrated clinical workflow. These are important evidences that FAIR principles lead to improved healthcare interoperability in real world healthcare systems and healthcare interoperability.

However, the current literature primarily focuses on hospital generated clinical data and institutional research infrastructures. Little research has been done on interoperability between different hierarchies of healthcare systems, like community healthcare systems and hospitals. This can be specially be observed in low resource maternal healthcare environments.

4.2 FHIR and healthcare interoperability

A major research area of healthcare interoperability is FHIR based integration. FHIR is one of the most significant healthcare interoperability standards, as it simplifies healthcare data exchange through modular healthcare resources and API based communication.

Mandel et al. proposes SMART on FHIR as a standards based interoperable platform for electronic health records [37]. This shows that healthcare applications can be developed independently and integrated through multiple healthcare systems using standardised APIs.

FHIR study show better interoperability between electronic health records (EHR), laboratory systems, and hospital systems. These research demonstrate that FHIR improves syntactic interoperability and facilitates healthcare data exchange.[38]

Literature shows that FHIR primarily solves interoperability at the structural and communication level instead of full semantic interoperability. Healthcare systems can still use different terminologies, coding systems, structures, and healthcare semantics even while using FHIR resources.

Therefore, FHIR alone can not sufficiently solve semantic interoperability between community maternal systems and hospitals.

This limitation is not just theoretical. Chukwu et al. (2022) surveyed practitioners and reviewed referral forms to design FHIR resource profiles for maternal and child health referrals between primary health care centres and hospitals in Ebonyi State, Nigeria [39]. It's a low and middle income country setting comparable to Kenya. The resulting profiles standardised how referral data is structured and exchanged between facilities, but the study stopped at that structural layer. No ontology or semantic mapping was introduced to reconcile the informal terminology recorded at the community level with the clinical coding used at the hospital, leaving the same terminology gap this thesis addresses unresolved.

Gebreslassie et al. (2023) addressed the same structural to semantic gap directly. They described how FHIR can be leveraged towards FAIR compliance within the VODAN-A implementation rather than treated as a competing framework. Their architecture uses FHIR as the exchange format at each facility's own data editor, with AllegroGraph functioning as a local FAIR Data Point queried through SPARQL rather than exported wholesale.[40] This supports the position taken in Section 2.2.3.

FHIR provides the mechanism for structured exchange. While a FAIR aligned semantic layer the kind this thesis builds through ontology and OBDA, provides the meaning FHIR alone does not guarantee.

4.3 Semantic Interoperability and RDF Based Healthcare Systems

Semantic interoperability research aims to resolve the limitations using semantic web technologies, ontologies, linked data, and RDF based healthcare systems.

Berners-Lee et al. introduces the Semantic Web as a machine readable web of linked information [20]. RDF was one of the foundational semantic web technologies for representing interconnected information through subject-predicate-object triples.

Studies on RDF based interoperability show that semantic technologies help improve data linkage, semantic querying, machine interpretation, and knowledge representation.[4] Knowledge graphs enabled better and visual representation of healthcare relationships between patients, diagnoses, referrals, laboratory observations, and treatments than traditional relational systems.

Research shows that semantic healthcare models have better interoperability through fragmented healthcare environments. Machine readable healthcare analytics and distributed healthcare integration benefits from semantic approaches.[41]

However, most of the RDF based research is focused on institutional healthcare systems, biomedical ontologies, or research repositories. Very limited research is conducted on RDF based semantic interoperability in maternal healthcare between CHWs and hospitals.

Zhang et al. (2018) demonstrate the importance of this technique outside the maternal health domain. Using the same Ontology Based Data Access approach adopted in this thesis, they linked heterogeneous relational cancer registries in the United States into a single ontology guided RDF dataset for integrative survival analysis [42]. Their study confirms that OBDA is a proven method for semantic integration of institutional health data. But it was applied to a single country, high resource cancer registry setting, not to community and hospital maternal health records in a low resource context.

4.4 Healthcare Ontologies and Controlled Vocabularies

Healthcare ontologies and controlled vocabularies are often discussed with semantic interoperability because they help with standardised representation of healthcare concepts.

SNOMED CT represents a comprehensive clinical terminology for representing diagnoses, symptoms, procedures, and healthcare observations [43]. ICD-10 and ICD-11 are standardized disease classification systems used worldwide for reporting and analytics [27].

Studies show that standardized vocabularies improve semantic consistency in fragmented healthcare systems. However, healthcare systems in low resource settings use mixed terminologies, paper documented records, local abbreviations, and informal healthcare descriptions.

Community health workers in healthcare systems in countries like Kenya record maternal observations like swelling, fever, or weakness using descriptive language, whereas hospitals use clinical diagnostic terminology like edema, hypertension, and preeclampsia.

This semantic mismatch creates major interoperability challenges and highlights the need for semantic mapping mechanisms.

4.5 Maternal Healthcare Interoperability in Kenya

Research on maternal healthcare interoperability remains relatively limited compared to the broader hospital interoperability research.

Current literature on maternal healthcare studies focus on antenatal care digitization, mobile health applications, reporting systems, and healthcare access rather than semantic interoperability frameworks.

Studies in African healthcare systems, including Kenya, show fragmented maternal health records, inconsistent reporting mechanisms, weak linkage between healthcare levels, and limited interoperability between community healthcare systems and hospitals [44]. This has been measured directly across Kenya and three neighbouring countries. A comparative analysis of antenatal care data collected in Ethiopia, Kenya, Nigeria, and Uganda found that of 228 variables recorded across the four countries' DHIS2 templates, only 4 were captured consistently in all four. This confirmed that fragmentation is not unique to Kenya but a shared feature of maternal healthcare data collection across the region [44].

Kenya's healthcare infrastructure relies heavily on Community Health Workers for maternal healthcare outreach and antenatal monitoring. CHWs conduct household visits, maternal follow up, referral activities, and antenatal care monitoring. They typically gather information like pregnancy status, antenatal attendance, referrals, maternal risk indicators, and follow up information. However, this data is isolated from hospital systems.

Referral hospitals maintain detailed clinical records like laboratory observations, syphilis test results, medications, HIV indicators, diagnoses, delivery outcomes, and treatment information.

The interoperability gap between these healthcare systems create a major challenge in continuity of care. For instance, pregnant women may be initially identified at community level before visiting hospitals for lab testing and diagnosis followed by treatment. This fragmented healthcare system makes it difficult to keep a track of maternal healthcare information consistent across different healthcare levels.[5, 6, 45–47]

A similar trend has been documented through empirical research elsewhere in East Africa. In Ethiopia, a study comparing paper based referral logbooks from a health centre against the corresponding patient files at the receiving hospital found that only 13.9% of referred patients could be perfectly matched between the two records. A solid proof of how severe the record linkage problem this thesis addresses becomes once it is actually measured rather than assumed [48]. A ten year scoping review of electronic health record implementation across Sub-Saharan Africa, covering 30 studies published between 2014 and 2024, further confirmed that the regional literature has focused on implementation barriers, infrastructure, and data accuracy rather than on semantic interoperability between different tiers of the healthcare system [49].

The closest existing work to this thesis addresses a related but distinct problem. Amare et al. (2025b) apply FAIR principles specifically to maternal syphilis data [50]. They developed a cross border health data space to monitor and prevent mother to child syphilis transmission across African countries. Their work operates at the level of data sovereignty and data visiting across national borders, rather than at the level of reconciling two independently operated systems within a single country through ontology based entity resolution. This is where the present thesis situates its own contribution.

Chapter 5

Context Description

This thesis focuses on Kenyan healthcare environments and specifically on interoperability challenges between community healthcare systems and referral maternity hospitals.

Kenya is located in East Africa, bordered by Ethiopia, Somalia, South Sudan, Uganda, and Tanzania, with a coastline on the Indian Ocean. Its capital, Nairobi, is the country's largest city and its primary administrative, economic, and healthcare hub, home to several of Kenya's national referral facilities. Both organisations this thesis works with are based in Nairobi.

The focal points are the community health care centre, Kamukunji Community Empowerment Initiative (KCEI) and Pumwani Maternity Hospital in Nairobi, Kenya.

5.1 Kenyan Healthcare Environments

Kenya has a fragmented healthcare environment having levels like community healthcare programs, health centres, county referral hospitals, and national referral hospitals. Access to healthcare varies between urban and rural areas. A lot of communities are dependent on community based healthcare units and workers for healthcare outreach and antenatal followup. As discussed in the previous chapter, this multi level structure is a recurring theme in the literature on Kenyan healthcare interoperability. This chapter focuses on how it unfolds at the two organisations this thesis works with.

5.2 Community Health Workers in Kenya

Community Health Workers play a crucial role in the Kenyan healthcare system. CHWs are the frontline healthcare service which includes maternal health education, antenatal monitoring, household visits, maternal risk identification, referrals,

and postnatal followup. They are often the first point of contact for pregnant women within a community. They collect information like demographic information, pregnancy status, antenatal attendance, estimated delivery dates, maternal danger signs, referrals and followup observations. These records are maintained through paper, household registers, mobile applications, or semi structured systems. Therefore, the generated data is frequently disconnected from hospital systems.

It's important to address a terminology gap here. Kenya's community based health workers were officially known as Community Health Volunteers (CHVs) until 2022. Then the government renamed them as Community Health Promoters (CHPs) as part of a broader professionalisation effort. They also included a stipend, standardised equipment kits, and formal recognition under the Community Health Policy 2020–2030 and the Primary Health Care Act 2023.[51, 52] This explains the mixed terminology that was encountered while working on this thesis. Specifically in the MOH 100 referral form which reads 'CHV', while stakeholders would usually use the term 'CHP'. This thesis uses 'CHW' throughout as the generic academic term covering both.

5.3 Community Health Centre

For this thesis we are working with Kamukunji Community Empowerment Initiative (KCEI) community health center. KCEI is a youth led organization established in 2011 and based in Majengo, Pumwani, Kamukunji Sub-County, Nairobi. [53] KCEI's stated mission is to empower vulnerable groups particularly children, youth, and women through skills development, knowledge sharing, and livelihood support. It organises its work across 5 areas: Economic Empowerment & Livelihoods, Education & Digital Skills, Health & Humanitarian Response, Peace Building & Prevention of Violent Extremism, and Social Protection & Human Rights Advocacy. In Health and Humanitarian Response, sexual and reproductive health outreach is a sole focus among several others. This is also consistent with what the CHW described as the day to day role at the organisation (discussed in the next chapter). They provide primary maternal healthcare services including ANC registration, maternal monitoring, referrals etc. They act as an intermediate healthcare layer between community healthcare workers and referral hospitals.

Healthcare information at this community centre is recorded on a mobile application and is digitised. But their system uses informal terminology, inconsistent identifiers, and heterogeneous reporting formats.

5.4 Pumwani Maternity Hospital

Pumwani Maternity Hospital is a level 5 maternity referral hospital in Kenya and one of the largest. It was founded in 1926 as Lady Grigg Maternity by the Lady Grigg Welfare League. It was then taken over by Nairobi Municipal Council in 1944. [54] It now operates 354 obstetric beds, 144 baby cots, and 2 theatres, handling 50–100 normal deliveries and 10–15 Caesarean sections daily, and roughly 18,000–20,000 deliveries a year. Reporting on a 2024 facility improvement grant records the hospital receiving approximately 317,000 patient visits annually against this bed capacity. The high number of patients increases the practical difficulty and value of unassisted manual approach in linking a client's community level record to their hospital record.

The hospital offers services like antenatal care, emergency obstetric care, maternal laboratory testing, delivery services, neonatal care, and specialized maternal treatment. It maintains more detailed clinical healthcare records like laboratory observations, syphilis test results, medications, diagnoses, procedures, delivery outcomes, HIV testing, and treatment histories.

Unlike KCEI, where records are directly fed into a mobile application, Pumwani first collects health records on paper. This data is subsequently fed into KenyaEMR, Kenya's electronic medical record platform. After that Pumwani's records become more closely coded to a national standard, in contrast to KCEI's informal records.

Pumwani's connection to this thesis's methodological approach also has a direct precedent. Between 2021 and 2022, Pumwani was one of 4 Kenyan health facilities in which an earlier VODAN Africa FAIR-OLR implementation study deployed a semantic, machine actionable patient data architecture built on AllegroGraph. This was done under a formal Data Use Agreement restricting any external computation to deidentified, aggregate metadata.[13] This thesis's own architecture (Chapter 8) extends the same underlying approach to the specific problem of linking Pumwani's records with those of a community health unit. An ontology mapped, locally held knowledge graph queried through SPARQL rather than exported wholesale.

5.5 Maternal Syphilis as a Continuity-of-Care Use Case

Maternal syphilis is a relevant interoperability use case because continuity of care is very important across multiple healthcare levels. WHO estimates that close to one million pregnant women have active syphilis annually worldwide, resulting in over 350,000 adverse birth outcomes.[55, 56] The majority of these do not occur in women who are not screened, but rather in women who were but didn't receive complete treatment or follow up. This is the systemic pattern, data collected but

not acted on or carried forward.

Within East Africa specifically, antenatal syphilis prevalence has been estimated around 6.5% by RPR/VDRL testing. [56]This is significantly higher than the global antenatal average of under 1%. In Kenya, antenatal syphilis prevalence is estimated at 3.8%, with more than half of positive cases associated with unfavourable obstetric outcomes including maternal death, prolonged morbidity, and congenital syphilis in the infant.[57] Syphilis screening is compulsory in Kenya under the national reproductive health program.[57] This means the screening step is well established. What remains fragile is the continuity between screening result and community level follow up.

This pattern is why syphilis was selected as the use case for this thesis. In the context of our case of CHUs and hospitals, pregnant women initially contact CHWs during ANC outreach and later visit the hospital for laboratory testing, diagnosis, and treatment. However, the different levels of healthcare systems make it very difficult to track referrals, link patient records, preserve healthcare history, maintain continuity of care, and proper communication between CHU and hospital systems. For example, a CHW may notice some danger signs and refer a pregnant woman for syphilis screening. And the hospital will record laboratory confirmed diagnoses and treatment information with different identifiers and clinical terminology. This fragmentation leads to interoperability challenges and reduce the continuity of care in maternal health.[50]

5.6 Existing Digital Health Infrastructure

Kenya currently has implemented several digital health initiatives like DHIS2 (District Health Information Software 2), OpenMRS-based systems, and electronic medical record platforms.

DHIS2 is deployed across all 47 counties and aggregates routine health facility data for the Ministry of Health's epidemiological surveillance, planning, and resource allocation.[58] Kenya was the first country in Sub-Saharan Africa to run a fully online DHIS2-based health information system, completed in 2011.[59] But it primarily supports aggregate reporting rather than semantic interoperability between distributed healthcare systems.

KenyaEMR is a customised distribution of the open-source OpenMRS platform, built to the Kenya Ministry of Health's 2011 EMR Standards and Guidelines.[60] It is deployed across 2,300 facilities and covers over 90% of Kenya's HIV-treatment patient population. This connects it to DHIS2 through a national interoperability layer that transmits data upwards while patient level records remain at the facility. This is the system Pumwani's workflow feeds data into.

Kenya has separately been rolling out eCHIS (the Electronic Community Health

Information System) since 2023 for CHUs, alongside the national professionalisation of Community Health Promoters.[61]This gives the CHPs a standard digital tool for household level data collection. This is a national level digitisation effort. KCEI's mobile application operates within this national system. Therefore, the CHU side data of this thesis is not a local NGO record collecting tool, but an instance of Kenya's own community health digitisation strategy.

Despite the effort on both the hospital and the community side, these systems don't interoperate semantically with one another at a patient level. Each system solves data collection and aggregate reporting within its own tier. [46]

5.7 Interoperability Challenges

Despite the efforts in the healthcare infrastructure, interoperability challenges exist because of fragmented healthcare systems, inconsistent healthcare identifiers, heterogeneous reporting structures, mixed systems of paper and digital, and limited semantic interoperability.

Community healthcare systems are usually household centred and focus on prevention, whereas hospital systems are encounter centered, clinically structured and focus on treatment.

These differences create major integration challenges. This thesis attempts to solve this challenge by proposing a FAIR-compliant semantic interoperability framework that combines standardised vocabularies, RDF based semantic modeling, knowledge graphs, and healthcare interoperability principles.

Chapter 6

Current Maternal Healthcare Data Workflows and Interoperability Challenges

6.1 Introduction

This chapter wants to address the research questions of what are the current interoperability challenges affecting maternal healthcare data integration between Community Health Workers and referral hospitals in Kenya? This chapter draws its observation from two sources; observations and interviews from the January 2026 field visit, and a structural comparison of the two organisation's data schema.

6.2 Current Data Collection Workflow

6.2.1 Community Health Unit (KCEI)

CHWs conduct household visits as part of KCEIs outreach. A CHW is responsible for visiting 100 households a month. When a CHW identifies a maternal danger sign, it is recorded in the mobile application, and the client is referred to the hospital. Rather than sending the client to the hospital alone, they accompany them there. According to a CHW the hospital usually sends a report back once the client has been seen. Also it was noted that no specific case where this feedback loop failed was recalled.

This is also supported by the MOH 100 community referral form itself. Section B in this form is explicitly designed as a report back mechanism. This means the feedback loop described by the CHW is not just an informant's recollection but a documented, designed part of the referral process. Whether this designed mechanism is consistently executed in practice is a separate question and this account should

be read alongside the hospital side perspective which does not fully agree with it.

6.2.2 Pumwani maternity Hospitals

At Pumwani, the first point of clinical record is paper. Client details are recorded in the MOH 405 Antenatal Care Registry at the point of care, before being entered into KenyaEMR. Individual patients also have their own paper case files, separate from the shared register. A hospital side contact indicated that they are not routinely communicated back to CHW's. He described it as a known weakness in the current referral pathway.

6.3 Structural Comparison of The Two Data Sources

This section wants to focus on the two system's actual recorded data, and compare them to find concrete interoperability gaps.

6.3.1 No shared identifier

The CHU dataset has its own individual_id and also notes down a national ID. The MOH 405 register instead identifies client by an ANC number, a sequential number for that specific pregnancy and facility; and the National Unique Patient Identifier (NUPI), a system generated identifier for all facilities nationally. The ANC number wouldn't follow a client to a different facility or another pregnancy, while NUPI will. Either of these are not shared with CHU system. The CHU never records NUPI, and the hospital never records the national ID that CHU does. This means that when a referral is successfully made, there is currently no identifier confirmed to link the two systems.

6.3.2 Asymmetric contact information

Although the two systems have a sort of shared field, phone numbers, they don't do so at the same level. The CHU schema records phone numbers of the household (generally the household head), not of the individual client. The hospital's MOH 405 register saves the phone number of the individual person. If the pregnant person is not the household head, which the household registration can't confirm, the CHU phone number won't match the hospital because it won't be of the patient.

6.3.3 A referral field that doesn't link

The MOH 405 register has a coded option to put CHU if the referral comes from one, but not which CHU. This means the form already has a section for the referral

this thesis is concerned with. However, the register doesn't capture the referral_id, individual_id, or the CHW recorded on the CHU side. A completed referral field at Pumwani can't automatically trace back to its CHU record.

6.3.4 Pregnancy Tracked Twice

Both CHU and the hospital maintain their own record of the same pregnancy. CHU's Maternal and Child Health dataset independently tracks pregnancy_status, anc_visits, and expected_delivery_date. Pumwani's register independently records its own Last Menstrual Period date, Expected Date of Delivery, and visit count. Since there is no identified link between the two, there are two records of the same pregnancy and are parallely updated without ever being recognised as the same event.

6.3.5 Syphilis Data

The hospital register has a dedicated field for syphilis screening. This records the type of tests performed (RPR/VDRL/Duo Test) and the results of the test (Positive/Negative/Not Applicable). This field, however, doesn't record whether treatment was received. The register's only treatment field is for Hepatitis B screening, not syphilis. Syphilis treatment, if recorded, will go under "Other Conditions and Treatment" field which is quite generic. The CHU Health Surveillance data has no such field. A CHW after observing danger signs consistent with Syphilis would record it under generic symptoms, diagnosis (if known), or danger_sign field. There is no dedicated screening result, or treatment field. So a CHW can flag and refer a case, but no structure is present for the outcome of that referral, whether the client was screened, or what the results were to be recorded on the CHU side.

A further piece of information that came from an interview with the CHW is relevant to this gap. When a hospital is notified that a client is a survivor of sexual violence, it generally administers a full sexually transmitted infection screening panel. Syphilis being one of them. This is significant for this thesis for two reasons. First, it means a sexual violence disclosure recorded at the CHU level functions as a genuine, real world trigger for hospital side syphilis testing, separate from the routine antenatal screening already discussed (Section 5.5). Second, it confirms why syphilis specifically was selected as this thesis's use case. Syphilis detection is not reached through compulsory screening alone, but also through a community originated pathway that current CHU records have no structured way to capture. A CHW recorded observation of this kind is therefore meaningful to a hospital confirmed syphilis case, even though the CHU has no dedicated field to record it as a syphilis specific concern.

6.4 Data Access Constraints

The actual patient level records held in the CHU mobile application and in Pumwani's KenyaEMR system are confidential clinical data. And access to either live system for this research would have been unethical. The structural comparison presented in this chapter is therefore grounded in the real schemas and forms used by KCEI and Pumwani, rather than in real patient records. A synthetic replica of both schemas, populated with mock data was constructed instead to develop and test a working architecture. How this replica was built, and the reasoning behind its structure is presented in Section 8.2.

6.5 Informal data

Referrals run through a physical paper form, the MOH 100 community referral form. MOH 100 is the form used at the point of referral itself, filled in by the CHW making the referral and by the receiving officer at the facility. The form is structured in two parts. Section A is completed by the referring CHW and captures the patient's name, sex, and age, the CHU name, the receiving link facility, and the reason for referral. Section B "Referral back to the Community," is completed by the receiving hospital officer and captures their name, the referring CHV's name, a "call made" yes/no field, and any followup instructions for the CHU. It is worth noting that the form itself uses the term "CHV" (Community Health Volunteer) rather than "CHW" or "CHP". This reflects the form predating Kenya's 2022 renaming to Community Health Promoters. However, none of this is digitised and hence not visible in the data used for this research. The MOH100 exists only as a paper artefact circulating between CHW and the facility. Whatever coordination and feedback the form enables in practice leaves no trace in either the CHU dataset or the hospital's data schema, and cannot be queried, linked, or reasoned over by the framework this thesis proposes. It is treated here as informal context that helps explain the current workflow, rather than as a data source for the interoperability framework itself.

6.6 Synthesis

The five structural gaps identified are five expressions of the same underlying condition. The CHU and the hospital each maintain a partial, independently held account of the same patient. Nothing in either system's current design makes those two accounts verifiably describe the same person, referral, or pregnancy. This was not only a conclusion drawn from comparing schemas. It surfaced directly in the field, as a disagreement between CHWs and hospital staff over whether referral outcomes are ever actually communicated back to the community level. A disagreement neither

side could settle from their own records alone (Section 6.3.3).

Table 6.1 sets out how each gap identified above is carried forward.

Table 6.1: Structural gaps and the requirements that resolve them

Gap (Chapter 6)	**Missing information**	**Requirement (Chapter 7)**	**Resolved (Ch. 8)**	**Evaluated (Ch. 9)**
No shared identifier (6.3.1)	Neither system records an identifier the other also holds	FR1 – Cross System Client Matching (7.4.1)	8.9.1	9.3.1
Referral field that doesn't link (6.3.3)	A CHU referral has no field connecting it to a hospital record	FR3 – Referral Traceability (7.4.3)	8.9.4, 8.9.6	9.3.3, 9.3.6
Pregnancy tracked twice (6.3.4)	Both systems independently record an EDD for the same pregnancy	FR4 – Pregnancy Record Reconciliation (7.4.4)	8.9.3	9.3.4
Syphilis data gap (6.3.5)	CHU has no structured syphilis field, only free-text danger signs	FR2 – Syphilis Screening and Diagnosis Traceability (7.4.2)	8.9.2	9.3.2
The above three, combined	No single record shows one client's full community-to-hospital pathway	FR5 – Continuity of Care (7.4.5)	8.9.5	9.3.5

Two of the gaps do not follow this pattern directly. The asymmetric contact information gap (Section 6.3.2) produces no requirement of its own. Neither the household phone number the CHU records nor the patient's own number the hospital records reliably identifies the same individual across both systems. Instead of engineering around them, this is ruled out as a candidate identifier which is what leaves full name and date of birth as the only descriptive information both systems independently and reliably hold about the same person, and therefore the only viable basis for FR1 (Section 7.4.1).

Continuity of care follows the opposite pattern. It is not a sixth gap alongside the other four, resolved by its own mapping. It is what becomes visible once the identity gap (Section 6.3.1), the syphilis data gap (Section 6.3.5), and the referral gap (Section 6.3.3) are each individually resolved and then read together as a single pathway for one client, from a community-level observation through to a hospital diagnosis.

Chapter 7

Ontology Design

7.1 Introduction

This chapter answers the research question "What specifications and interoperability requirements are necessary for integrating maternal healthcare data between community and hospital healthcare systems?" Last chapter talked about interoperability gaps in the current system. This chapter focuses on the two artefacts built to close them, a shared maternal health ontology, and an Ontology-Based Data Access (OBDA) mapping layer connecting both organisation's relational data to it.

7.2 Requirements Derived from Stakeholders and Analysis

The four core goals established for this research are semantic interoperability between CHU and hospital systems, maternal syphilis detection, referral tracking, and continuity of care.

7.2.1 Functional Requirements

Functional requirements, in order of priority:

- **Cross System client matching (FR1)** in the absence of any identifier shared between CHU and hospital records. This is the foundational requirement all others depend on.
- **Syphilis screening and diagnosis traceability (FR2)** across systems. A CHU recorded danger sign can be connected to a hospital recorded laboratory result.

- **Referral traceability (FR3)**. A CHU referral should be able to be linked to its corresponding hospital referral record. This will solve the disagreement between CHW and hospital accounts of whether referral outcomes are actually communicated back.
- **Pregnancy record reconciliation (FR4)**. A pregnancy can be tracked independently by both the systems. It is also important that it is recognised as a single event rather than two unconnected records.
- **Continuity of care (FR5)**. Not a requirement produced by any single gap. Continuity of care across the full community to hospital pathway for a single client. It makes the combined result of FR1, FR2, and FR3 traceable from one starting point. So, a community-level observation, a referral, and a hospital diagnosis are readable as a single connected record rather than three separate facts.

7.2.2 Semantic Interoperability Requirements

The architecture must represent CHU and hospital concepts using a common set of classes and properties, rather than treating each system's terminology as a closed local vocabulary. It also needs to preserve each side's original structure. CHU concepts should not be forced into clinical coding designed for hospital use, and vice versa.

7.2.3 Non functional requirements

- **(NFR1)** The mapping layer must operate without modifying either organisation's existing database.
- **(NFR2)** The knowledge graph must be rerunnable against updated data without manual rework.
- **(NFR3)** The confidentiality of the data should be maintained by authenticated access to the resulting graph.
- **(NFR4)** The semantic model should allow new maternal health concepts to be added without major redesign.
- **(NFR5)** Healthcare information should be represented in machine-readable formats suitable for automated processing.
- **(NFR6)** The architecture should support future expansion to additional healthcare facilities and datasets.

- **(NFR7)** Each organisation's data must remain on infrastructure it operates and controls. The other organisation may query it only through federation, authorised and scoped to what that query returns, and gains no copy of or claim over the underlying data.

7.3 Ontology

The previous section defines what the architecture must achieve. It doesn't show how. This section describes the shared ANC ontology, the vocabulary the architecture uses to represent CHU and hospital concepts on common ground, and the foundation that every requirement ultimately depends on.

7.3.1 Origin and Extension Approach

This thesis extends on an existing hospital ontology developed by a collaborating student Lina, instead of building one from scratch (Section 3.8.2). We add classes and properties specific to the Community Health Unit setting. This means that CHU and hospital concepts are two layers in one shared concept. The hospital side already carries a clinically grounded structure built with reference to standard terminologies including SNOMED CT and LOINC. While the CHU adds household, community, and referral level concepts that this thesis is responsible for.

The reason to use an already built ontology for hospitals and adding on to it was done for two reasons.

First, both the thesis and the original hospital ontology's development sit within the same COMPASS project context (Section 3.3). Reusing ontology avoids two incompatible hospital domain ontologies existing side by side. Developing two independently created hospital ontologies would represent a failure of the semantic infrastructure we are trying to showcase here. Extending the already existing ontology also makes this thesis's work interoperable with works related to COMPASS or hospitals in Kenya.

Secondly, it also is a demonstration of this thesis's argument that semantic interoperability should let independently developed systems connect through a shared model. They shouldn't require either one to rebuild to match the other. If we build an ontology, specifically tailored for Pumwani, we would have contradicted ourselves by solving the hospital side by rebuilding it.

This layered extension approach is also what resolves the semantic interoperability requirement (Section 7.2.2). CHU and hospital concepts are represented using a common set of classes and properties without either side's terminology being treated as a closed local vocabulary, while each side's original structure is preserved rather than forced into the other's.

7.3.2 Ontology Composition

Lina's original ontology has 43 classes, 88 object properties, and 162 datatype properties. The ontology represents a clinically grounded hospital data model built with reference to standard terminologies including SNOMED CT and LOINC. The existing classes cover the internal structure of a hospital encounter, like `Patient`, `Visit`, `Consultation`, `Admission`, `Diagnosis`, `LaboratoryResult`, `Medication`, `Referral`, `Provider`, `Department`, and `Facility` etc. This models how a hospital records and organises care for a patient who is already inside its own system.

This hospital focused model had no way to represent anything outside a hospital. It had no concept of a household, a community health unit, a community health worker's observation, or any mechanism for describing a different, independently run system that might also have patient data. This was because it was built for the purpose of hospital interoperability and it reasonably assumed that anything it needed to connect to would already look structurally like a hospital record. It didn't have answers for the structural incompatibility of CHU and hospital records while describing the same person.

This thesis adds 8 classes specifically to fill that gap.

`CommunityHealthUnit`, `HouseholdRegistration`, `Client`, `ServiceDelivery`, and `EnvironmentalWASH` together represent KCEI's actual household based data model. These have no hospital equivalent, since hospitals do not organise records around households at all.

`HealthSurveillance` exists to hold the CHW's own recorded symptoms, diagnosis, and danger signs. As discussed in Section 6.3.5, CHU has no structured way to record a syphilis relevant observation, only generic free text fields.

`MaternalChildHealth` exists because the CHU tracks pregnancy status, ANC visit count, and expected delivery date entirely independently of the hospital's own pregnancy record (Section 6.3.4). The class name is deliberatley close to the inherited `MaternalChildHealthRecord` class. That class represents the hospital's own maternal and child health record. `MaternalChildHealth` represents the Community Health Unit's independently maintained tracking of the same pregnancy. The two are kept separate rather than merged because they are two records of the same event held by two organisations, and the framework's purpose is to link them while preserving that distinction (Section 6.3.4).

`CHU_Referral` has CHU's own referral record (who was referred, when, and why). This has no counterpart in Lina's hospital modelled ontology, and is the CHU side of the referral field that doesn't link problem (Section 6.3.3).

There are 12 added object properties split into two purposes, which address the problems generated because of adding a non hospital system into the picture. Seven are structural links holding the new CHU classes together internally (`hasClient`, `hasHealthSurveillance`, `hasMaternalChildHealth`, `hasServiceDelivery`, `ha`

sEnvironmentalWASH, hasHouseholdRegistration, hasClientReferral). The remaining 5 are cross linking properties (hasCorrespondingPatient, hasCorrespondingPregnancy, hasCorrespondingReferral, hasCorrespondingSyphilisRecord, referredToFacility) that answer the functional requirements mentioned earlier (Section 7.2.1). referredToFacility is an exception though. It crosses from a CHU class (CHU_Referral) to a hospital class (Facility) in the same way the other 4 do. But it doesn't resolve any specific functional requirement. It instead confirms a referral's destination as a graph link. This supports the machine readability non functional requirement (NFR5) (Section 7.2.3).

The 58 added datatype properties attach the concrete attributes each new class needs, drawn directly from the fields actually present in KCEI's schema (Table 8.1). Some examples are hasDangerSigns, hasSymptom, and hasSurveillanceDiagnosis on HealthSurveillance; hasGPSLocation, hasHouseholdHeadName, and hasWaterSource on the household side; hasFullName, hasNationalID, and hasAge on Client. Of the 58 added datatype properties, 53 attach the concrete attributes the 8 new CHU classes need. The remaining 5 extend inherited hospital classes with attributes the MOH 405 register records but the original ontology did not model. For example, hasANCNumber on Patient, hasFacilityName, hasMFLCode and hasServiceDeliveryPoint on Facility, and hasReferralSource on Referral. These are additions to the hospital layer rather than the community layer. No inherited property was redefined to accommodate them.

In total, this thesis added 8 classes, 12 object properties, and 58 datatype properties to Lina's inherited 43 classes, 88 object properties, and 162 datatype properties to make a final combined ontology of 51 classes, 100 object properties, and 220 datatype properties within the ontology's own https://w3id.org/dhs/ontology# namespace. The serialised file additionally declares 17 external HL7 FHIR StructureDefinition URIs as classes. These are the targets of the inherited owl:equivalentClass and rdfs:subClassOf alignments rather than classes authored here, and are not counted among the 51.

7.4 Functional Requirement Specifications

This section works through every functional requirement individually, identifying the specific ontology specification (a class, property, or mapping) that satisfies it (Section 7.2.1). Each specification defines conditions under which a CHU record and a hospital record are considered to represent the same person or event. A link is created only when those conditions are met. If there is not enough evidence, no link is created. This prevents incorrect connections between unrelated records. How these specifications are technically executed against the relational data is described in Chapter 8.

7.4.1 FR1 – Cross System Client Matching

The core interoperability problem was that the CHU and hospital records share no common identifier (Section 6.3.1). This is resolved by `hasCorrespondingPatient` (`Client` → `Patient`) – the primary entity resolution link, resolving the no shared identifier gap.

Two records in the two systems are treated as the same client when the CHU's `hasFullName` and `hasClientDateOfBirth` are identical to the hospital's `hasFirstName`, `hasLastName` and `hasDateOfBirth`. This is the only descriptive information both systems independently record about the same person. This is because neither the CHU's individual ID nor the hospital's NUPI is shared between them, name and date of birth are what remain to resolve identity (Section 6.3.1).

Because every other cross system requirement depends on first resolving which CHU client corresponds to which hospital patient, this specification is the foundation the remaining functional requirements build on.

7.4.2 FR2 – Syphilis Screening and Diagnosis Traceability

The CHU's Health Surveillance dataset has no dedicated field for syphilis screening results or treatment, only generic symptom, diagnosis, and danger sign fields (Section 6.3.5). This is resolved in two ways, a general and a direct.

This gap is resolved not by a single new field, but by making the hospital's own structured result reachable from the CHU side once a client is matched. Once a client is resolved to their corresponding patient, a syphilis case becomes traceable by chaining from that resolved identity through the hospital's own clinical structure. `hasCorrespondingPatient` leads to `hasVisit`, which leads to `hasHIVCareRecord`, which carries `hasVDRLResult`. Where available, this is paired with the CHU's own recorded danger signs for that same client via `hasHealthSurveillance` and `hasDangerSigns`. This allows a positive result to be read alongside whatever prompted the original community level concern.

This direct approach is supported by `hasCorrespondingSyphilisRecord` (`HealthSurveillance` → `HIVCareRecord`), which gives a specific CHU observation a direct path to the specific hospital result it corresponds to. Two records are treated as corresponding here when the CHU's `hasSymptom`, `hasSurveillanceDiagnosis`, or `hasDangerSigns` names syphilis directly, or signals sexual violence, combined with a `hasVDRLResult` confirmed positive at the hospital. Sexual violence is included as a matching signal because a disclosure of this kind is itself a real world trigger for hospital syphilis screening (Section 6.3.5). Therefore, a CHW recorded observation of this kind is a meaningful indicator to a hospital confirmed case.

The direct path is not a dependency, but a refinement. Where its conditions are not met, a positive result remains fully traceable through the general route above. A

syphilis case is therefore never left unconnected purely because no CHU observation happened to use qualifying language.

7.4.3 FR3 – Referral Traceability

The disagreement between the CHW and hospital accounts of whether referral outcomes are actually communicated back is resolved by `hasCorrespondingReferral` (`CHU_Referral` → `Referral`) (Section 6.6). It gives that outcome a verifiable, queryable record rather than leaving it dependent on informal reporting.

A CHU referral and a hospital referral are treated as corresponding when the same client has already been resolved (FR1) and the CHU's `hasCHUReferralDate` falls within a defined window of the hospital's recorded visit date. A window (7 days) is used rather than requiring an exact date match because a referral being made and a client being seen at the hospital are not necessarily recorded on the same day.

Because this correspondence is conditional, a CHU referral could otherwise become effectively invisible whenever no matching hospital record is found. `hasClientReferral` (`Client` → `CHU_Referral`) prevents this. It links a client directly to their own CHU referral record independent of any hospital match, and is populated for every CHU referral regardless of whether one is later found. Where the cross system condition is not satisfied, the referral's own reason, date, and identity remain fully intact and queryable through this link. The referral is not lost, only its link to a specific hospital record is unconfirmed.

7.4.4 FR4 – Pregnancy Record Reconciliation

The pregnancy tracked twice problem (Section 6.3.4) is resolved by `hasCorrespondingPregnancy` (`MaternalChildHealth` → `MaternalProfile`).

Two pregnancy records are treated as corresponding when the same client has already been resolved (FR1) and the CHU's `hasCHUExpectedDeliveryDate` falls within a defined window of the hospital's independently recorded EDD. A window (14 days) is used rather than an exact match because the two systems may each estimate or record this date slightly differently for the same pregnancy, without that difference meaning they describe different events.

7.4.5 FR5 – Continuity of Care

Unlike FR1 through FR4, continuity of care has no dedicated correspondence property of its own. What distinguishes it from the other four requirements is that it is not a claim about how two records from different systems relate to each other. It is a claim about whether the records already reconciled by FR1, FR2, and FR3 remain

reachable from a single starting point as one connected account of a patient's care, rather than as three separate facts a reader would have to look up and assemble by hand.

A client satisfies FR5 when four conditions hold simultaneously from that client's own node in the graph. The client resolves to a corresponding hospital patient (FR1, hasCorrespondingPatient). A CHU recorded sign for that client is present (hasHealthSurveillance, hasDangerSigns). The client's referral is linked both directly (hasClientReferral) and where a match exists, to its corresponding hospital referral (FR3, hasCorrespondingReferral). At the resolved patient's hospital visit, a confirmed positive VDRL result is reachable through FR2's general path (hasVisit, hasHIVCareRecord, hasVDRLResult). No single one of these is sufficient on its own. A client resolved under FR1 alone has an identity but no demonstrated pathway. A client with a positive VDRL result reachable under FR2 alone has a diagnosis with no visible predecessor. FR5 is satisfied only when the whole chain, from a community level observation to a confirmed hospital outcome is traversable for the same client without any step being assumed or supplied from outside the graph.

Where the chain is incomplete for a given client, FR5 is simply not satisfied for that client, but nothing about the records that do exist is hidden or lost. The client remains queryable under whichever of FR1 through FR4 apply to them. Only the combined pathway is unavailable.

7.5 Conclusion

This chapter identifies the functional, semantically interoperable and non functional requirements (Section 7.2) derived from Chapter 6. While Section 7.3 and 7.4 give the specifications required to meet the functional and semantically interoperable requirements. We are yet to resolve the non functional requirements. We will address them in Chapter 8 by introducing OBDA and AllegroGraph.

Chapter 8

Knowledge Graph Construction and Pipeline Implementation

8.1 Introduction

The previous chapter defined the maternal health ontology and identified the functional, semantic interoperability, and non-functional requirements it must satisfy. This chapter answers sub Research Question C – how a FAIR compliant methodology supports integration between Community Health Workers and hospital systems in practice, by describing a working system rather than only a conceptual one.

How the ontology is connected to relational data through OBDA. How the mapped data is materialised into RDF, loaded into a queryable knowledge graphs in AllegroGraph, and kept up to date through a single pipeline. This chapter also resolves the 7 non-functional requirements (Section 7.2.3) that Chapter 7 left open.

8.2 Constructing the Synthetic CHU and Hospital Databases

To develop and test a working architecture for this thesis, both organisation's data schemas were instead replicated as relational tables in a local H2 database. They were populated with synthetic mock data records that mirrored the real structure and content of each system.

Table 8.1 lists the specific H2 tables built for each side of the architecture. The CHU tables mirror KCEI's full mobile application data model. The hospital tables were derived from the MOH 405 register fields, keeping the column most relevant to ANC, referral tracking, and syphilis screening that are important for this thesis. The hospital side was a deliberate filtered subset, rather than a complete replica. The actual register has over 40 columns spreading across demographics, laboratory

Table 8.1: H2 Tables Constructed for Each Data Source

CHU side H2 Table	**Hospital side H2 Table**
COMMUNITY_HEALTH_UNIT	FACILITY_REGISTER
HOUSEHOLD_REGISTRATION	ANC_CLIENT
CLIENT	ANC_VISIT
SERVICE_DELIVERY	PREGNANCY_DETAILS
HEALTH_SURVEILLANCE	LABORATORY_RESULTS
REFERRAL	HIV_SCREENING
MATERNAL_CHILD_HEALTH	TREATMENT_AND_PREVENTION
ENVIRONMENTAL_WASH	HOSPITAL_REFERRAL

results, HIV care, and treatment tracking. Fields such as breast examination findings and FGM status, present in the real register but outside this thesis's scope, were excluded.

H2 was used because it is a lightweight, embeddable, standards compliant relational database that requires no separate server process. It is fully accessible over a standard JDBC connection. Ontop, the mapping tool used, operates over any JDBC accessible relational source. It has no dependency on which specific database engine is behind that connection. H2 served purely as a stand-in data source during development, letting the ontology, mappings, and pipeline be built and tested end to end without needing to be rebuilt later for a different database engine.

The mock H2 database used here though is not a permanent part of the proposed framework, but a placeholder for whichever real system a deploying organisation already uses. This will work because Ontop mappings are defined against a schema and a JDBC connection string rather than against H2 specifically. Replacing the mock database with a live system (like KenyaEMR) would require updating the JDBC connection details in the pipeline to point towards the live database. It would be wise to revisit and review the mappings for any minor change in the live system schema against the mock dataset, although it's been made to closely mirror it. No change to the ontology or to the underlying pipeline logic would be required for this substitution. This is a deliberate design property of the architecture. It is intended to sit alongside an organisation's existing systems rather than to require replacing them with a new database of its own.

8.2.1 Population Design

The mock dataset was populated with 20 CHU clients and their associated households, health surveillance records, and referrals. Of these, 17 clients were given matching hospital side records. Three were deliberately left without a hospital match, to test whether FR1 correctly leaves genuinely unmatched clients unconnected rather than forcing a link (Section 7.4.1).

Among the 17 matched clients, 15 pregnancy records were given Expected Delivery Dates within the 14 day matching window used by FR4. While 2 were deliberately given dates more than 14 days apart, to confirm that the window correctly rejects records that describe genuinely different estimates rather than linking every matched client's pregnancy regardless (Section 7.4.4).

From the matched clients, 6 were given a positive VDRL result at the hospital. Of these 6, 4 were also given a CHU recorded danger sign referencing sexual violence. This directly reflected the field finding that a hospital notified of a sexual violence disclosure administers a full STI screening panel, syphilis included (Section 6.3.5). This was an intentional design choice intended to test whether FR2's direct matching path correctly identifies this specific real world pathway, rather than relying solely on the general Client-to-Patient chain. The remaining 2 positive cases were left without any qualifying CHU side to confirm that a syphilis case is still fully traceable through the general route even when no such disclosure exists.

Of 7 CHU referrals, 6 were given a matching hospital referral record within the 7 day window used by FR3. One was deliberately left without a match, to test whether `hasClientReferral` keeps a referral fully traceable even when no cross system match is found (Section 7.4.3). This mirrors the unresolved disagreement between the CHW and hospital accounts of whether referral outcomes are reliably communicated back (Section 6.6).

8.3 Ontology Based Data Access (OBDA)

Both the organisation's data is saved in ordinary relational tables. Converting either database directly and permanently into RDF would require duplicating and maintaining a second copy of the data. OBDA avoids this by defining a set of mappings describing how existing relational tables and columns correspond to ontology classes and properties, without changing the source database. Ontop is the OBDA tool we used to execute the mappings, expressed in the standard R2RML mapping language, and to materialise the result into RDF file.

The mappings are held in two separate files, `chu-mapping.ttl` and `hospital-mapping.ttl`. One per organisation, so that each side can be materialised and deployed without reference to the other. Between them the mapping file consists of 30 TriplesMaps. 15 on the CHU side, 10 on the hospital side, and 5 performing cross-system entity resolution. The 5 entity resolution TriplesMaps are present in both files rather than in one, giving 35 TriplesMap instances in total across the two files. The reason for this deliberate duplication is set out in Section 8.4.2. Table 8.2 summarises this mapping file in full.

Table 8.2: Mapping File Summary

Source Table	**Ontology Class Created**	**Linked From**	**Linking Property**
		CHU side	
COMMUNITY_HEALTH_UNIT	CommunityHealthUnit	—	—
HOUSEHOLD_REGISTRATION	HouseholdRegistration	CommunityHealthUnit	hasHouseholdRegistration
CLIENT	Client	HouseholdRegistration	hasClient
SERVICE_DELIVERY	ServiceDelivery	Client	hasServiceDelivery
HEALTH_SURVEILLANCE	HealthSurveillance	Client	hasHealthSurveillance
MATERNAL_CHILD_HEALTH	MaternalChildHealth	Client	hasMaternalChildHealth
ENVIRONMENTAL_WASH	EnvironmentalWASH	HouseholdRegistration	hasEnvironmentalWASH
REFERRAL	CHU_Referral	Client	hasClientReferral
		Hospital side	
FACILITY_REGISTER	Facility	—	—
ANC_CLIENT	Patient	—	—
ANC_VISIT	Visit	Patient	hasVisit
PREGNANCY_DETAILS (joined with ANC_VISIT)	MaternalProfile	Visit	hasMaternalProfile
LABORATORY_RESULTS (joined with ANC_VISIT)	HIVCareRecord	Visit	hasHIVCareRecord
HOSPITAL_REFERRAL (joined with ANC_VISIT)	Referral	Visit	hasReferral
		Cross-system entity resolution	
CLIENT × ANC_CLIENT (on FULL_NAME + DATE_OF_BIRTH)	—	Client	hasCorrespondingPatient
MATERNAL_CHILD_HEALTH × PREGNANCY_DETAILS (via resolved client, EDD ± 14 days)	—	MaternalChildHealth	hasCorrespondingPregnancy

Source Table	Ontology Class Created	Linked From	Linking Property
REFERRAL × HOSPITAL_REFERRAL (via resolved client, dates ± 7 days)	—	CHU_Referral	hasCorrespondingReferral
HEALTH_SURVEILLANCE × LABORATORY_RESULTS (via resolved client)	—	HealthSurveillance	hasCorrespondingSyphilisRecord
Referred to facility	—	CHU_Referral	referredToFacility

8.4 Mappings

8.4.1 The Structural Mappings

The 7 structural object mappings (Section 7.3.2) hold the newly added CHU classes together internally. All are implemented as working Ontop mappings, executed as a pair of TriplesMaps per class. One populates the class's own datatype properties from its source table, and the other produces the object property that links it to its parent class.

CommunityHealthUnit generated from COMMUNITY_HEALTH_UNIT. It has no parent link. Its datatype properties (hasCHUID, hasCHUName, hasLinkedFacility, hasCHUWard, hasCHUSubCounty, hasCHUCounty) are populated directly from the table's columns.

HouseholdRegistration generated from HOUSEHOLD_REGISTRATION. It is linked from CommunityHealthUnit via hasHouseholdRegistration. Its datatype properties (hasHouseholdID, hasHouseholdHeadName, hasHouseholdPhoneNumber, hasGPSLocation, hasNumberOfMembers, hasDateRegistered, hasHouseholdVillage) are populated from the table directly.

Client generated from CLIENT. It is linked from HouseholdRegistration via hasClient. Its datatype properties (hasIndividualID, hasFullName, hasSex, hasClientDateOfBirth, hasAge, hasClientDisabilityStatus, hasNationalID, hasRelationshipToHead) are populated from the table directly. Client is the most heavily connected of the eight CHU classes. It is the subject of 4 other structural links (to ServiceDelivery, HealthSurveillance, MaternalChildHealth, and CHU_Referral) and the object of the cross-system hasCorrespondingPatient (Section 7.4.1).

ServiceDelivery generated from SERVICE_DELIVERY. It is linked from Client via hasServiceDelivery. Its datatype properties (hasServiceID, hasServiceType, hasProgramArea, hasServiceDeliveryOutcome, requiresFollowUp, hasServi

ceDeliveryDate) are populated from the table directly.

HealthSurveillance generated from HEALTH_SURVEILLANCE. It is linked from Client via hasHealthSurveillance. Its datatype properties (hasRecordID, hasSymptom, hasSurveillanceDiagnosis, hasChronicCondition, hasPregnancyStatus, hasImmunizationStatus, hasSurveillanceNutritionStatus, hasDangerSigns, hasDateRecorded) are populated from the table directly. In this class we store CHU recorded danger signs relevant to syphilis (Section 6.3.5, Section 7.4.2).

MaternalChildHealth generated from MATERNAL_CHILD_HEALTH. It is linked from Client via hasMaternalChildHealth. Its datatype properties (hasMCHID, hasPregnancyStatus, hasANCVisits, hasCHUExpectedDeliveryDate, hasDeliveryPlace, hasPostnatalVisits, hasChildImmunizationStatus, hasGrowthMonitoring) are populated from the table directly.

EnvironmentalWASH generated from ENVIRONMENTAL_WASH. It is linked from HouseholdRegistration via hasEnvironmentalWASH. Its datatype properties (hasWaterSource, hasSanitationType, hasLatrineAvailable, hasHandwashingFacility, hasWasteDisposalMethod) are populated from the table directly.

CHU_Referral generated from REFERRAL. It is linked from Client via hasClientReferral. Its datatype properties (hasReferralID, hasCHUReferralDate, hasReasonForReferral, hasFacilityReferredTo, hasReferralType) are populated from the table directly. This link is populated for every CHU referral regardless of the existence of a corresponding hospital referral (Section 7.4.3). This keeps a CHU referral traceable even where cross system matching does not succeed (Section 8.9.6).

The same structural pattern connects Facility, Patient, Visit, MaternalProfile, HIVCareRecord, and Referral, inherited from Lina's original ontology and mapping work. FACILITY_REGISTER and ANC_CLIENT sit at the root of their respective hierarchies (Facility and Patient), with no parent link. ANC_VISIT produces Visit, linked from Patient via hasVisit. PREGNANCY_DETAILS, LABORATORY_RESULTS, and HOSPITAL_REFERRAL are each joined against ANC_VISIT on ANC_NUMBER to recover the VISIT_ID needed to link MaternalProfile, HIVCareRecord, and Referral back to their parent Visit via hasMaternalProfile, hasHIVCareRecord, and hasReferral respectively.

8.4.2 Entity Resolution Mapping

The 5 entity resolution mappings (Section 7.4) [LinkingMapping, SyphilisLinkMapping, PregnancyLinkMapping, ReferredFacilityLinkMapping, and ReferralLinkMapping] are defined identically in both mapping files and run automatically as part of Ontop's materialisation step. Each produces a persistent RDF relationship every time the pipeline materialises. The relationship is wrritten in both repositories rather than into one.

This duplication is deliberate. A correspondence link is not the property of

either organisation alone. It is a statement about a pair of records, one held by each side. Writing it into both repositories means either side can answer a cross system question from its own entry point. A Community Health Worker querying the CHU repository can reach the corresponding hospital patient, and a clinician querying the hospital repository can reach the community record that preceded a referral without either having to know which repository the link belongs to. The 49 correspondence triples produced per materialization are consequently present twice across the deployment, which is why the two repositories' triple counts sum to more than a single merged graph would contain (Section 8.5).

Client To Patient Linking (FR1) `LinkingMapping` joins the CHU `CLIENT` table to the hospital `ANC_CLIENT` table on `FULL_NAME` and `DATE_OF_BIRTH`, since no shared identifier exists between the two systems. Every match produces a `hasCorrespondingPatient` triple connecting a `Client` individual to its corresponding `Patient` individual. This is a permanent, queryable edge in the knowledge graph, present every time the graph is materialised, and traceable by anyone querying it later. Because every other cross system requirement depends on first resolving which CHU client corresponds to which hospital patient, this specification is the foundation the remaining functional requirements build on.

```
SELECT c.INDIVIDUAL_ID, ac.NUPI
FROM CLIENT c
JOIN ANC_CLIENT ac ON c.FULL_NAME = ac.FULL_NAME AND
    c.DATE_OF_BIRTH = ac.DATE_OF_BIRTH
```

Syphilis Record Linking (FR2) `SyphilisLinkMapping` resolves the link where the CHU's own recorded symptom, diagnosis, or danger sign text names syphilis directly or names sexual violence, with a confirmed positive VDRL result at the hospital (Section 6.3.5).

This produces a `hasCorrespondingSyphilisRecord` triple (`HealthSurveillance` → `HIVCareRecord`) giving relevant CHU observations a direct edge to their corresponding hospital screening record, rather than relying solely on the general Client to Patient chain.

```
SELECT hs.RECORD_ID, lr.ANC_NUMBER
FROM HEALTH_SURVEILLANCE hs
JOIN CLIENT c ON hs.INDIVIDUAL_ID = c.INDIVIDUAL_ID
JOIN ANC_CLIENT ac ON c.FULL_NAME = ac.FULL_NAME AND
    c.DATE_OF_BIRTH = ac.DATE_OF_BIRTH
JOIN ANC_VISIT av ON av.NUPI = ac.NUPI
JOIN LABORATORY_RESULTS lr ON lr.ANC_NUMBER = av.ANC_NUMBER
```

```
WHERE (LOWER(hs.SYMPTOM) LIKE '%syphilis%' OR LOWER(hs.DIAGNOSIS) LIKE '%
    syphilis%'
    OR LOWER(hs.DANGER_SIGNS) LIKE '%syphilis%' OR LOWER(hs.DANGER_SIGNS)
        LIKE '%sexual%'
    OR LOWER(hs.DANGER_SIGNS) LIKE '%assault%' OR LOWER(hs.DANGER_SIGNS)
        LIKE '%gbv%')
AND LOWER(lr.RPR_VDRL_DUO_TEST_RESULT) = 'positive'
```

Pregnancy Linking (FR4) `PregnancyLinkMapping` addresses the pregnancy tracked twice problem. It joins the CHU's `MATERNAL_CHILD_HEALTH` records to hospital `PREGNANCY_DETAILS` records via the resolved client match. It also additionally requires the two systems to independently record Expected Delivery Dates to fall within 14 days of each other. A successful match produces a `hasCorrespondingPregnancy` triple, representing the two systems separately tracked pregnancy records as the same event.

```
SELECT mch.MCH_ID, pd.ANC_NUMBER
FROM MATERNAL_CHILD_HEALTH mch
JOIN CLIENT c ON mch.INDIVIDUAL_ID = c.INDIVIDUAL_ID
JOIN ANC_CLIENT ac ON c.FULL_NAME = ac.FULL_NAME AND
    c.DATE_OF_BIRTH = ac.DATE_OF_BIRTH
JOIN ANC_VISIT av ON av.NUPI = ac.NUPI
JOIN PREGNANCY_DETAILS pd ON pd.ANC_NUMBER = av.ANC_NUMBER
WHERE ABS(DATEDIFF('DAY', pd.EDD, mch.EXPECTED_DELIVERY_DATE)) <= 14
```

Referral Linking (FR3) `ReferralLinkMapping` addresses the referral field that doesn't link problem (Section 6.3.3). It joins CHU `REFERRAL` records to hospital `HOSPITAL_REFERRAL` records, again via the resolved client match. The CHU referral date and the hospital ANC visit date are required to fall within 7 days of each other. This produces, if successful, a `hasCorrespondingReferral` triple.

There is an additional important property, `hasClientReferral` (`Client` → `CHU_Referral`). It links a client directly to their own CHU referral record, independent of any hospital match. This is a within-CHU-system link, not a cross system one. It is populated regardless of whether a hospital match is later found.

```
SELECT r.REFERRAL_ID, hr.ANC_NUMBER
FROM REFERRAL r
JOIN CLIENT c ON r.INDIVIDUAL_ID = c.INDIVIDUAL_ID
JOIN ANC_CLIENT ac ON c.FULL_NAME = ac.FULL_NAME AND
    c.DATE_OF_BIRTH = ac.DATE_OF_BIRTH
JOIN ANC_VISIT av ON av.NUPI = ac.NUPI
JOIN HOSPITAL_REFERRAL hr ON hr.ANC_NUMBER = av.ANC_NUMBER
```

```
WHERE ABS(DATEDIFF('DAY', av.DATE_OF_VISIT, r.REFERRAL_DATE)) <= 7
```

Referred Facility Linking `ReferredFacilityLinkMapping` (`CHU_Referral` → `Facility`) resolves the CHU's own `hasFacilityReferredTo` value to an actual link to the corresponding facility. While `hasFacilityReferredTo` only records the name of the facility a referral was sent to, the `referredToFacility` triple lets the whole linkage be queried and traversed as part of the graph itself (Section 2.2.3).

```
SELECT r.REFERRAL_ID, fr.MFL_CODE
FROM REFERRAL r
JOIN FACILITY_REGISTER fr ON r.FACILITY_REFERRED_TO = fr.HEALTH_FACILITY
```

Therefore, entity resolution across CHU and hospital happens every time the pipeline is rerun against updated data. More about the results of the entity resolution mapping is presented in Section 8.9.

8.4.3 Mapping Scope

Several columns in the CHU and hospital H2 tables are not currently carried into the mapping layer, including two hospital tables. The reason for so being, they weren't required for the scope of this thesis (interoperability between CHU and hospital systems, syphilis detection, referral tracking, and continuity of care). This doesn't in any way affect the results of the mapping (Section 8.9). They are included in the H2 schema for future relevance, if the framework later is extended beyond its current scope.

Table 8.3: Fields and Tables outside current mapping scope

Source Table	Unmapped Field(s)
HOUSEHOLD_REGISTRATION	COUNTY, SUB_COUNTY, WARD
FACILITY_REGISTER	START_DATE, END_DATE
ANC_CLIENT	AGE, SUBCOUNTY_COUNTY, VILLAGE_ESTATE_LANDMARK
ANC_VISIT	ANC_NUMBER_REVISIT, NUMBER_OF_VISITS
PREGNANCY_DETAILS	MUAC_CODE, HEIGHT_CM, WEIGHT_KG, BLOOD_PRESSURE
LABORATORY_RESULTS	HAEMOGLOBIN, BLOOD_SUGAR_TESTING, BLOOD_GROUP_RHESUS, URINALYSIS, RPR/VDRL/Duo Test Type, HEPATITIS_B_RESULT, HEPATITIS_B_TREATMENT, TB_SCREENING_CODE
HOSPITAL_REFERRAL	REFERRAL_TO, REMARKS
HIV_SCREENING	Entire table not mapped

Source Table	Unmapped Field(s)
TREATMENT_AND_PREVENTION	Entire table not mapped

It's worth noting that in LABORATORY_RESULTS, only RPR_VDRL_DUO_TEST_RESULT is mapped out of the 10 columns. This is due to the thesis's syphilis focused scope (Section 5.5). The remaining laboratory fields belong to maternal health areas this thesis does not address. HIV_SCREENING and TREATMENT_AND_PREVENTION follow the same reasoning at the table level.

8.5 RDF Materialisations

Once we identify relevant data, we call on Ontop from the pipeline to materialise the mapped data into RDF. The pipeline calls Ontop's materialize command twice, passing each mapping file, the shared ontology file, and a properties file describing the H2 connection, and writing the result to each organisation's individual RDF output file. This is the same Ontop we used in Protégé as a plugin during mapping development. Here it is used as a standalone step so that materialisation can happen automatically rather than requiring manual export every time from Protégé.

The current mock dataset, containing 20 CHU clients and 17 hospital ANC clients, produces 1450 triples in the Community Health Unit graph and 2004 triples in the hospital graph. The two figures sum to 3454 rather than to the size of a single merged graph, because the 49 cross-system correspondence triples are written into both graphs rather than into one (Section 8.4.2). This confirms that the mapping layer executes successfully end to end against real, populated H2 tables, not only against the ontology's design.

8.6 Pipeline Overview (ETL process)

The process of automated Extract, Transform, and Load (ETL) used to to implement the proposed semantic interoperability framework is ran by a single pipeline. Rather than manually executing each processing step, the framework employs a Python based FAIRification pipeline that orchestrates the complete workflow. It begins with relational healthcare data stored in the H2 database and ending with semantically interoperable RDF knowledge graphs stored in AllegroGraph. The pipeline integrates database validation, semantic transformation using Ontop, semantic post-processing, RDF deployment, and execution reporting into a single automated workflow.

Unlike traditional ETL pipelines that primarily transfer data between relational databases, the proposed pipeline performs a semantic ETL process in which relational healthcare records are transformed into RDF knowledge graphs according to a shared ontology while preserving semantic interoperability and FAIR principles.

Figure 8.1 illustrates the overall ETL workflow implemented by the framework.

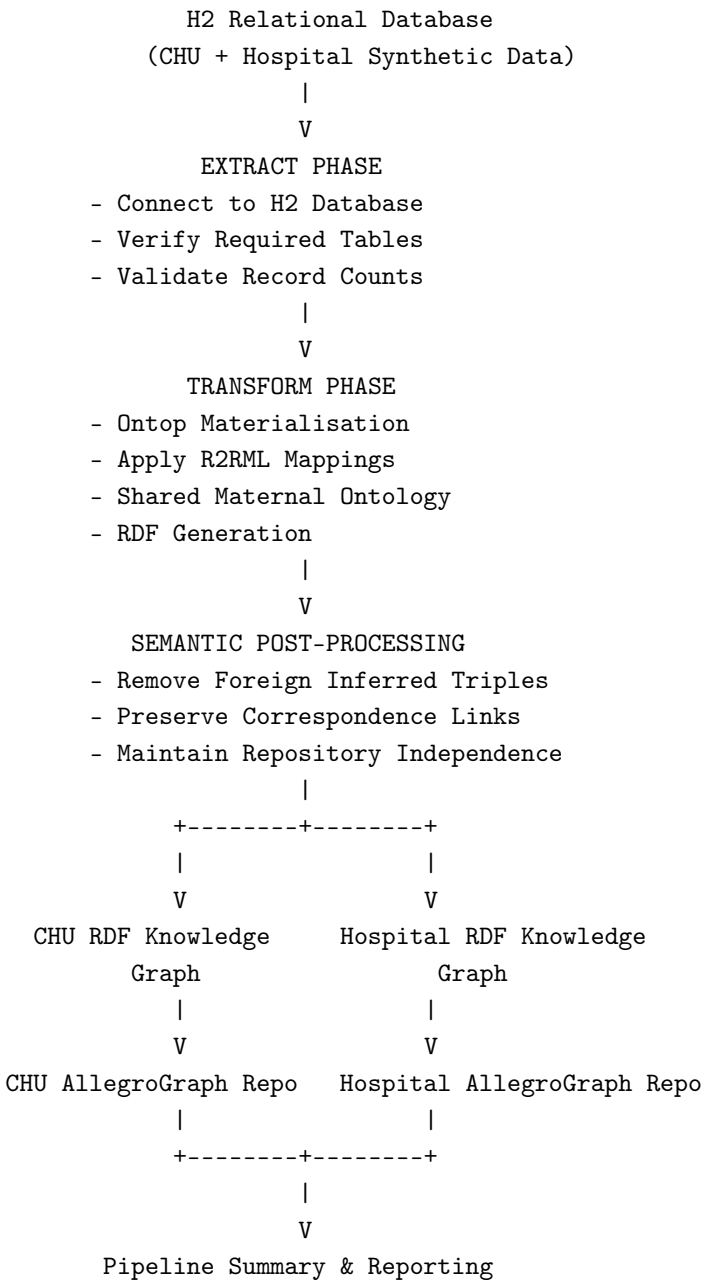


Figure 8.1: Overview of the ETL Process and FAIRification Pipeline

8.6.1 Extraction Process

The ETL process begins by extracting data from the relational H2 database containing the synthetic Community Health Unit (CHU) and hospital datasets. Before semantic transformation is initiated, the pipeline establishes a JDBC connection to the database and performs validation checks to ensure that the required source data are available.

The pipeline first verifies that all expected database tables are present. It then validates the number of records contained within each table, including clients, antenatal care visits, pregnancies, referrals, laboratory results, HIV screening records, and treatment records. Performing these validation checks before transformation prevents incomplete or inconsistent datasets from propagating through the remainder of the FAIRification process.

8.6.2 Transformation Process

During the transformation phase, Ontop materialises the relational healthcare data into RDF by applying the R2RML mapping files developed for the CHU and hospital datasets. The mappings establish correspondences between relational tables and the classes and properties defined within the shared maternal healthcare ontology.

Instead of modifying the original relational databases, Ontop performs a semantic transformation that converts relational records into RDF triples. There are separate mapping files for CHU and hospital datasets. This allows each dataset to be materialised independently while using the same ontology to achieve semantic interoperability. The output of this phase consists of RDF knowledge graphs that provide a common semantic representation of healthcare information originating from both organisations.

8.6.3 Post-Processing

After RDF materialisation, the generated RDF graphs undergo an additional semantic post-processing stage before deployment. This stage is unique to this framework and ensures that the resulting repositories remain organisationally independent.

Because both mapping files are materialised against the shared ontology, each side's output initially carries some inferred type and property triples describing entities that belong to the other organisation. These are removed before deployment. The rule applied is specific. A triple is removed when its subject is a URI belonging to the opposite side, unless its predicate is one of the five cross system correspondence properties. Foreign URIs appearing as the object of a triple are always retained.

This distinction is what makes the architecture work. Removing subject-position foreign triples means the CHU repository holds no descriptive statements about

hospital patients and the hospital repository holds no descriptive statements about CHU clients. Each organization's repository describes only its own records. Retaining foreign URIs in object position and exempting the correspondence predicates from removal means the link between the two sides survives in both repositories and remains traversable. The result is that each repository is independently deployable and independently meaningful, while the correspondence between them is queryable from either side.

8.6.4 Loading Process

In the final ETL stage, the generated RDF knowledge graphs are uploaded to AllegroGraph through its REST interface. The CHU graph and the hospital RDF graph are uploaded independently into separate AllegroGraph repositories. Maintaining separate repositories reflects the federated architecture proposed in this research. Each organisation retains ownership of its own FAIR data while exposing semantically interoperable knowledge graphs that can be queried through SPARQL queries.

8.7 The Knowledge Graph (AllegroGraph)

The materialised RDF are uploaded on AllegroGraph over its HTTP interface. They are posted directly to the target repository's statements endpoint using basic authentication. This happens twice, into two separate repositories. The Community Health Unit graph is uploaded to KCEI's repository and the hospital graph to Pumwani's. Each has its own base URL and its own credential set held in the pipeline's environment configuration rather than in its source. Both repositories are deployed on the participating institution's own server rather than on commercial cloud hosting (Section 3.8.6). Each organisation's data therefore physically resides within the jurisdiction and under the operational control of the facility that produced it. This is the condition Localisation requires (Section 2.1.6) and the property that distinguishes this deployment from a conventional hosted database. Deployment details and access points are given in Appendix F. Once uploaded, the data becomes queryable through SPARQL and explorable through AllegroGraph's graph visualisation tools.

This is the architectural expression of the Localisation and Ownership principles described (Section 2.1.6). Each organisation holds, controls, and authenticates access to its own data. Neither organisation's records are copied into a combined store held by a third party, and neither organisation can read the other's clinical detail without a federated query explicitly authorised at that other side's own endpoint. This follows the same data visiting pattern VODAN Africa deploys, and the same separation Kawu et al. recommend between externally generated and facility collected data (Section 6.3.1).

The pipeline confirmed a successful upload (HTTP 200). 1450 statements were confirmed live and queryable in the Community Health Unit repository and 2004 in the hospital repository.

Interoperability is then achieved at query time rather than at storage time. Because both graphs are materialised against the same ontology and mint URIs in the same `https://w3id.org/dhs/ontology#` namespace, a `Patient` URI produced by the hospital mapping is the same identifier the CHU repository's `hasCorrespondingPatient` triple points to. A query issued against one repository can therefore delegate part of its graph pattern to the other through SPARQL's `SERVICE` keyword and join the two result sets, without either dataset being transferred or merged (Section 8.9).

This is when the entity resolution becomes visible. A client's records, CHW recorded danger signs, a referral, and a hospital syphilis result can be traced as a connected chain, even though they are physically held in two separately owned and separately authenticated stores.

8.8 Resolution of the Non Functional Requirements

The 7 non functional requirements mentioned in Chapter 7 (Section 7.2.3) couldn't be resolved on their own, since they concern the infrastructure the mappings run within rather than the mappings themselves. They will be addressed and resolved here.

8.8.1 NFR1 – No Modification to Source Database

The mapping layer operates without modifying either organisation's existing database, as the pipeline reads from the H2 tables through Ontop's OBDA layer and writes only to the RDF output file. Nothing in the pipeline issues a write back to `CLIENT`, `ANC_CLIENT`, or any other source table (Section 8.6).

8.8.2 NFR2 – Rerunnable without Manual Rework

The knowledge graph is rerunnable against updated data without manual rework. The sequence from table validation through to upload runs as a single automated script rather than requiring a manual export from Protégé each time new data is added (Section 8.6).

8.8.3 NFR3 – Confidentiality through Authenticated Access

The confidentiality of the data is maintained by authenticated access to the resulting graph. The AllegroGraph repositories are not openly queryable. Both writing to it

and querying it require valid basic authentication credentials, and the two repositories hold separate credential sets. This separation is what gives each organisation control over its own data rather than shared control over a combined store. KCEI's credentials do not grant access to Pumwani's repository. A federated query spanning both must be authorised at both endpoints. The credentials themselves are read from environment variables at runtime and are not recorded in the pipeline source (Section 8.7).

AllegroGraph's access model supports finer granularity than the single credential set used per repository in the current deployment. The platform allows multiple user accounts to be defined against the same repository, each with its own authentication credentials and its own permission level, read only query access, read write access, or full catalog administration. Rather than every consumer sharing one superuser account, and these permissions are scoped per repository rather than granted platform wide. This deployment currently uses one authenticated service account per repository, held by the pipeline for materialisation and by the small number of stakeholders given direct query access during evaluation (Section 9.4). Adopting per-user accounts, for example a read-only account for a clinician distinct from the pipeline's read-write account, would not require any change to the ontology, the mappings, or the pipeline logic. It is a configuration change within AllegroGraph's existing security model and is identified here as a natural refinement for a production deployment with a larger and more varied set of users.

8.8.4 NFR4 – Extensibility

The semantic model should allow new maternal health concepts to be added without major redesign. This is a distinct claim from the semantic interoperability requirement (Section 7.2.2). That requirement concerns whether the current model correctly represents both systems' concepts. NFR4 concerns whether future concepts could be added without redesigning it. This requirement is a property of RDF and OWL's underlying data model, not something that has to be separately engineered on top of it. A relational database describes a fixed schema. Adding a new concept means adding a column or a table, which can require altering existing tables, updating foreign key constraints, or migrating existing rows. RDF has no such constraints. In the graph, every triple is independent. So, a new class or property is simply a new set of declarations added to the ontology, alongside the existing ones. It never has to touch an already existing triple.

This is how the CHU layer was added to Lina's hospital ontology in this thesis (Section 7.3.1). The 8 CHU classes, 12 object properties, and 58 datatype properties introduced here were declared as `owl:Class`, `owl:ObjectProperty`, and `owl:DatatypeProperty` individuals in the same namespace as Lina's existing terms (`https://w3id.org/dhs/ontology#`). They each carried their own `rdfs:domain`, `rdfs:r`

ange, and rdfs:comment. None of Lina's original 43 classes, 88 object properties, or 162 datatype properties were redefined, renamed, or removed to accommodate them. The combined ontology is a strict superset of what existed before. The two layers connect only through the five cross-linking properties (Section 8.4.2). This is clear evidence that extension requires addition, not modification.

8.8.5 NFR5 – Machine Readability

Healthcare information should be represented in machine-readable formats suitable for automated processing. There are two features of this architecture for this requirement:

First, every concept materialised in the knowledge graph is identified by a URI (a Uniform Resource Identifier) generated from the mapping's subject templates, in the form https://w3id.org/dhs/ontology#HealthSurveillance_HS001. This is not like a row or column number in the H2 table, meaningful only within that database. A URI is globally unique and resolvable. Any system that comes across this identifier, for example AllegroGraph, refers to the exact same entity without ambiguity. This is what makes federation possible without merging. A Patient URI minted by the hospital mapping and referenced by a CHU correspondence triple denotes the same individual in both repositories. So, a query can join across the two endpoints even though the underlying data was never combined into one store.

Second, every class and property in the ontology is formally typed in OWL (the Web Ontology Language). This is the standard used in this ontology. Classes are declared as owl:Class, and properties are declared as either owl:ObjectProperty or owl:DatatypeProperty. They each carry an explicit rdfs:domain and rdfs:range. This formal typing is what separates the graph from a plain text export. A SPARQL engine, a reasoner, or any external application can determine what kind of entity a triple describes and what data type to expect from its value directly from the ontology. It doesn't have to rely on column names, documentation, or human interpretation. Several datatype properties are additionally aligned to LOINC codes through skos:closeMatch. For example, a blood pressure related property is linked to http://loinc.org/rdf/8462-4. This allows external systems that already use LOINC to recognise the same clinical concept, an alignment that free text cannot provide.

These two features are what make the RDF output directly queryable without any transformation step.

8.8.6 NFR6 – Scalability

The architecture should support future expansion to additional healthcare facilities and datasets. This requirement is verifiable by the properties of the mapping layer. The structural mappings that generate `CommunityHealthUnit` and `Facility` have no clause restricting them to a single CHU or facility. `CommunityHealthUnitMapping`'s underlying query is `SELECTCHU_ID,CHU_NAME,LINKED_FACILITY,WARD,SUB_COUNTY,COUNTYFROMCOMMUNITY_HEALTH_UNIT`. And, `FacilityRegisterMapping`'s is `SELECTMFL_CODE,HEALTH_FACILITY,SERVICE_DELIVERY_POINT,COUNTY,SUB_COUNTYFROMFACILITY_REGISTER`. Every other CHU and hospital structural mapping links to its parent through a foreign key join (`HOUSEHOLD_REGISTRATION` to `CHU_ID`, `CLIENT` to `HOUSEHOLD_ID`, and so on) rather than a hardcoded identifier, so the same reasoning extends down each hierarchy. If a second CHU or a second hospital facility were inserted into these tables, along with their associated households, clients, and visits, the next pipeline run would materialise triples for them automatically, with no change to the mapping file itself. The same reasoning extends to the deployment layer. A third producer would be accommodated by adding a third mapping file and a third repository entry to the pipeline's side configuration, rather than by restructuring the existing two.

This is a structural assumption, not a demonstrated result. The current mock dataset has only one CHU (KCEI) and one hospital (Pumwani). The pipeline has never been run against data spanning more than one facility on either side. Whether it works as expected on a larger scale remains untested and is identified here as future work.

8.8.7 NFR7 – Localisation

Each organisation's data must remain on infrastructure it operates and controls, queryable by the other side only through federation. The CHU graph and the hospital graph are held in two separate AllegroGraph repositories, each with its own credentials and neither copied into a store operated by a third party.

What makes this enforceable rather than only physical is authorisation. A query that reaches from one repository into the other through SPARQL's SERVICE keyword (Section 8.9) still requires valid credentials at the endpoint being reached. Being authenticated to query the hospital repository from the CHU side does not transfer ownership of hospital data to CHU. It grants a scoped answer, limited to the fields and correspondence triples that specific query returns and nothing beyond it. NFR3 (Section 8.8.3) is the authentication mechanism that makes this possible. NFR7 is the property that mechanism exists to protect.

8.9 SPARQL based Validation

To demonstrate that an architecture has produced a semantically linked graph that answers our questions, we need to query them for each claim. Because the architecture deploys two separate repositories. Each query below states which repository it is executed against and whether it requires federation. Four of the six run entirely within the Community Health Unit repository, because the correspondence triples they depend on are written into both repositories (Section 8.4.2). The syphilis continuity query and the combined chain must reach the hospital's own clinical structure, which exists only in the hospital repository. These use SPARQL's `SERVICE` keyword to delegate that part of the pattern to the hospital endpoint and join the result. Both were executed with the hospital repository's own credentials supplied at the federated endpoint. The hospital endpoint is shown here in placeholder form. The deployed host and credentials are held in the pipeline's environment configuration and are not reproduced in this document. These results are reported here as raw technical outputs. Their interpretation against the maternal healthcare use cases they demonstrate is taken up in Section 9.3.

8.9.1 Client Matching Query (FR1)

The foundation every other cross linking mapping depends on. This query was run against the Community Health Unit repository. It requires no federation, because `hasCorrespondingPatient` is materialised into both repositories and is therefore answerable from either side.

```
PREFIX : <https://w3id.org/dhs/ontology#>
SELECT DISTINCT ?client ?patient
WHERE {
  ?client :hasCorrespondingPatient ?patient .
}
```

This query returned 17 rows out of the 20 CHU clients that were successfully resolved to a corresponding hospital patient on an exact match of full name and date of birth. The remaining 3 clients have no hospital record in the current dataset.

8.9.2 Syphilis Continuity of Care Query (FR2)

The main use case of this thesis is tracing a maternal syphilis case across the CHU and hospital systems. It was tested via a SPARQL query, run against the AllegroGraph repository.

```
PREFIX : <https://w3id.org/dhs/ontology#>
SELECT DISTINCT ?client ?patient ?vdrlResult ?dangerSigns
```

```
WHERE {
  ?client :hasCorrespondingPatient ?patient .
  SERVICE <https://HOSPITAL_HOST/repositories/Hospital/sparql> {
    ?patient :hasVisit ?visit .
    ?visit :hasHIVCareRecord ?hivRecord .
    ?hivRecord :hasVDRLResult ?vdrlResult .
    FILTER (LCASE(?vdrlResult) = "positive")
  }
  OPTIONAL {
    ?client :hasHealthSurveillance ?surveillance .
    ?surveillance :hasDangerSigns ?dangerSigns .
  }
}
```

This query goes from a CHU `Client`, through the resolved `hasCorrespondingPatient` link, to the hospital `Patient`'s clinical visit records, to the associated syphilis screening result. And where available it comes back to the CHU's own recorded danger signs for that same client.

The query returned 12 rows when run without `DISTINCT`, rather than a clean list of positive cases. This was because when traced against the underlying data, `LABORATORY_RESULTS` in the H2 schema is keyed only by `ANC_NUMBER`. It has no record of which specific ANC visit it was performed at. Therefore, the mapping layer links each pregnancy's single syphilis result to every ANC visit recorded under that pregnancy. For example, a client with 5 ANC visits produces five identical rows for the same underlying positive result. Adding `DISTINCT` collapses this correctly to one row per client. This can also be done because `client`, `patient`, `vdrlResult`, and `dangerSigns` are identical across a given patient's repeated rows. After this addition, the query returns exactly 6 distinct clients, each with a confirmed positive VDRL result. Out of the 6, 4 carry CHU recorded danger signs of sexual violence (Section 6.3.5). Each of these 6 cases is fully traceable from the CHU side, without either system needing to share an identifier at the point of data collection. This query is the clearest demonstration of the two repository design working as intended. The first triple pattern is answered inside the Community Health Unit repository. The `SERVICE` block is answered inside the hospital repository. The join between them happens at query time on the shared `?patient` URI. No hospital clinical record is copied into the CHU repository at any point, and no CHU record is copied into the hospital's. Yet the community observation and the laboratory result are returned together in a single result set.

This result proves that semantic interoperability has been achieved. It also illustrates OBDA based integration.

8.9.3 Pregnancy Linkage Result (FR4)

The corresponding query for `hasCorrespondingPregnancy` was run against the live AllegroGraph repository.

```
PREFIX : <https://w3id.org/dhs/ontology#>
SELECT DISTINCT ?mch ?pregnancy
WHERE {
  ?mch :hasCorrespondingPregnancy ?pregnancy .
}
```

This returned 15 rows of the 17 CHU clients that were successfully linked to a hospital patient. This mapping's subject and object are both at the pregnancy level (`MaternalChildHealth_{MCH_ID}` and `MaternalProfile_{ANC_NUMBER}`) rather than at the visit level, so it does not spread across a client's multiple ANC visits. The 2 unmatched had expected delivery dates apart for more than 14 days.

Every CHU `MaternalChildHealth` record with a matching hospital client now has a single, confirmed, queryable link to its corresponding hospital `MaternalProfile`, directly resolving the pregnancy tracked twice problem.

8.9.4 Referral Linkage Results (FR3)

The equivalent query for `hasCorrespondingReferral` was also run against AllegroGraph.

```
-- Referral linkage query
PREFIX : <https://w3id.org/dhs/ontology#>
SELECT DISTINCT ?chuReferral ?hospitalReferral
WHERE {
  ?chuReferral :hasCorrespondingReferral ?hospitalReferral .
}

-- CHU side referral count (confirms 7)
SELECT (COUNT(DISTINCT ?r) AS ?count) WHERE { ?r a :CHU_Referral . }

-- Hospital side referral count (confirms 6)
SELECT (COUNT(DISTINCT ?r) AS ?count) WHERE { ?r a :Referral . }
```

The query for `hasCorrespondingReferral` returned 6 of the 7 CHU referrals. The remaining one had no corresponding hospital record (Section 8.9.6).

8.9.5 Combined Continuity of Care Chain (FR5)

Each result demonstrated above resolved a single functional requirement. This query demonstrates FR1, FR2, and FR3 operating together, ran against the live Allegro-

Graph repository.

```
PREFIX : <https://w3id.org/dhs/ontology#>
SELECT DISTINCT ?client ?dangerSigns ?chuReferral ?hospitalReferral ?
    vdrlResult
WHERE {
  ?client :hasCorrespondingPatient ?patient .
  ?client :hasHealthSurveillance ?surveillance .
  ?surveillance :hasDangerSigns ?dangerSigns .
  ?client :hasClientReferral ?chuReferral .
  ?chuReferral :hasCorrespondingReferral ?hospitalReferral .
  SERVICE <https://HOSPITAL_HOST/repositories/Hospital/sparql> {
    ?patient :hasVisit ?visit .
    ?visit :hasHIVCareRecord ?hivRecord .
    ?hivRecord :hasVDRLResult ?vdrlResult .
    FILTER (LCASE(?vdrlResult) = "positive")
  }
}
```

This returned 4 rows. The same 4 that were referred for sexual violence and a syphilis test (Section 8.9.2). This confirms the chain. A CHU recorded sexual violence, a CHU referral, that referral's confirmed match to a hospital referral record, and a confirmed positive syphilis result at that same hospital visit.

8.9.6 Referral Traceability Without a Cross System Match

This query is to show that a referral remains fully traceable even with no corresponding hospital record (Section 8.9.4).

```
PREFIX : <https://w3id.org/dhs/ontology#>
SELECT ?referral ?reason ?referralDate
WHERE {
  ?referral a :CHU_Referral .
  ?referral :hasReasonForReferral ?reason .
  ?referral :hasCHUReferralDate ?referralDate .
  FILTER NOT EXISTS { ?referral :hasCorrespondingReferral ?
      hospitalReferral }
}
```

This returned 1 row. Its full record can be queried via `hasClientReferral` despite having no hospital match (Section 8.4.1).

8.10 Operationalising FAIR principles

The implementation described in this chapter and the previous chapter together operationalise the four FAIR principles in a concrete way.

8.10.1 Findable

Findability operates at two levels in this architecture.

At the level of individual records, every entity materialised in the knowledge graph has a unique URI generated from the mapping's subject templates. Therefore, it doesn't exist only as a row meaningful in only one H2 table.

At the level of the datasets themselves, the metadata describing both repositories is registered in a FAIR Data Point and exposed through the Africa Health Data Space (AHDS) (Section 3.8.6). This makes the Community Health Unit and hospital repositories discoverable, and queryable by parties who have not been told they exist. This happens without the underlying patient records ever leaving either institution's own server. The distinction matters for sensitive health data. What is published is the description of the dataset and the means of querying it, not the dataset. Appendix F gives the registration details.

8.10.2 Accessible

Accessibility in this architecture is deliberately not the same as openness. FAIR's Accessible pillar (Section 2.1.6) calls for data to be retrievable through standardised protocols with clear authorisation conditions, not for it to be publicly readable. Both AllegroGraph repositories are queried over the same standard protocol, SPARQL over HTTP. Any authorised client, whether a stakeholder's browser session, another organisation's federated query, or a future integration, retrieves data the identical way rather than through a bespoke interface. What controls accessibility is authentication rather than the query mechanism itself. Neither repository is openly queryable, and both writing to and querying either one requires valid credentials specific to that repository (NFR3, Section 8.8.3). This is what keeps the architecture accessible in the FAIR sense, standardised, protocol based retrieval, while remaining closed in the sense sensitive maternal healthcare data requires. The distinction matters specifically because of Localisation (Section 2.1.6). A repository that is accessible only through authenticated SPARQL, rather than exported as a file, is what allows federated querying (Section 8.9) to answer a cross system question without either organisation's dataset ever having to be transferred or copied to be reached.

8.10.3 Interoperable

The shared ontology is what makes the graph interoperable in practice. A CHU observation and a hospital diagnosis are instances of the shared same classes, rather than in each system's local structure. When the client records resolve to the same person, they are connected by a correspondence triple written into both repositories. Hence the connection is traversable from either side without the two datasets being combined. Also note, because the hospital side of the ontology extends an already existing ontology for hospital interoperability, instead of being built specifically for Pumwani, it makes Pumwani interoperable with other hospitals as well.

8.10.4 Reusable

The pipeline is rerunnable and not a one off export. This makes a new CHU or hospital data entry easy to be validated, checked, materialised, and uploaded again without manual work. The mapping files are also reusable and revisable independent of the source system. The H2 database used throughout this thesis is not a permanent placeholder. As the Ontop mappings operate against any JDBC accessible relational source, replacing the mock H2 database with a live hospital or CHU system would require only a change to the pipeline's connection configuration. It won't need a redesign of the ontology, the mappings, or the pipeline logic described here.

8.10.5 Ownership

Under FAIR-OLR, ownership is federated between the facility that produces the data and the data subject the record describes (Section 2.1.6), rather than belonging exclusively to either party or to a third party infrastructure. This architecture's 2 repository design (Section 8.7) upholds this directly. KCEI holds and authenticates access to the Community Health Unit graph, and Pumwani holds and authenticates access to the hospital graph. Neither the researcher nor any third party infrastructure holds a copy of either dataset. Ownership is expressed architecturally rather than only declared as policy. Each organisation's credentials are the only means of writing to or reading from its own repository (NFR3, Section 8.8.3), and a federated query from the other side returns a scoped answer without transferring ownership of the underlying data (Section 8.7).

8.10.6 Localisation

Localisation requires that data be held under the local control of the facility that produced it, rather than in infrastructure external to that facility (Section 2.1.6). Both AllegroGraph repositories are deployed on the participating institutions' own servers

(Section 3.8.6), so each organisation's data physically resides within the jurisdiction that produced it. This is the specific property NFR7 requires, enforced through the same organisation scoped credentials that satisfy NFR3 (Section 8.8.7). Registering both repositories' metadata in a FAIR Data Point and exposing it through the AHDS (Section 3.8.6) does not weaken this. What is published there is a description of each dataset and how to query it, not the data itself (Section 8.7), so discoverability and Localisation are satisfied simultaneously rather than trading off against each other. Interoperability across the two repositories is achieved by visiting data at query time through SPARQL federation (Section 8.9) rather than by centralising it, the same data visiting pattern VODAN Africa's architecture follows (Section 2.1.6).

8.10.7 Regulatory Compliance

Regulatory Compliance requires that the jurisdiction's own regulatory requirements for the data produced there are respected (Section 2.1.6). This is the least directly demonstrable of the three, since the architecture operates entirely on synthetic data (Section 8.2) and was not required to be certified against Kenya's Data Protection Act or Digital Health Act as a condition of this research. What the architecture does do is remove the specific failure mode raised as a live concern (Section 1.3), that sensitive patient data could be transferred outside the jurisdiction and regulatory framework governing it. Because each organisation's data remains on infrastructure it controls within Kenya (Section 3.8.6), and cross organisation access happens only through authorised, scoped federation rather than data export (Section 8.7), the architecture is structured to operate within a facility's own regulatory boundary rather than assuming compliance can be addressed afterward. Confirming this against the specific requirements of the Data Protection Act (2019) and Digital Health Act (2023) remains outside the scope of this thesis and is identified as future work (Section 10.1.4).

Chapter 9

Evaluation and Findings

9.1 Introduction

This chapter answers sub Research Question D, How do stakeholders evaluate the usability and practical value of the proposed interoperability framework? It evaluates proposed FAIR compliant semantic interoperability framework developed in this research. The objective of the evaluation is to determine whether the framework successfully addresses the interoperability challenges identified within maternal healthcare data exchange between Community Health Units and referral hospitals in Kenya.

The evaluation prioritises the healthcare use cases derived from the functional requirements (Section 7.4). This chapter focuses on interpreting what the resulting knowledge graph (Chapter 8) demonstrates regarding semantic interoperability, continuity of care, referral traceability, and integration of maternal healthcare information across independent healthcare systems.

Together with the feedback from stakeholders, these findings provide an assessment of the practical capabilities of the proposed framework and demonstrate the extent to which the research objectives have been achieved.

9.2 Technical Evaluation

This section evaluates the architecture against the technical criteria (Section 3.9.1). Ontology consistency, mapping correctness, RDF generation success, knowledge graph construction, and query execution. Each criterion is already demonstrated mechanically in Chapter 8. We just assess here if it meets the standard the criterion sets.

Ontology Consistency – The combined ontology is a strict superset of Lina's original hospital model. This property in itself is a form of consistency evidence. It means the extension could not have introduced a conflicting redefinition of an exist-

Table 9.1: Technical Evaluation Summary

Criterion	Evidence	Assessment
Ontology consistency	Section 7.3.2; NFR4	Met
Mapping correctness	Table 8.2; Section 8.4	Met
RDF generation success	Section 8.5	Met
Knowledge graph construction	Section 8.7	Met
Query execution	Section 8.9	Met
Scalability (NFR6)	Section 8.8.6	Untested
Federated query execution	Section 8.9; Section 8.7	Met

ing term. Consistency was also tested against real world data, not only against the ontology's own internal logic. During development one inherited property was found to be declared with a datatype range that did not match how the corresponding field is actually recorded on the MOH 405 register. Thus, the range was corrected accordingly. This is a useful finding on its own. An OWL reasoner can confirm that an ontology is internally coherent. But, it cannot confirm that a declared range matches the real world format of the data it is meant to describe. That only got clear when once real records were mapped and materialised, which is itself an argument for evaluating an ontology against live data rather than design review alone.

Mapping Correctness – Every property whose ontology domain is one of the 8 CHU classes was confirmed to be used somewhere in the mapping file, with no declared but unmapped properties remaining. Table 8.2's 30 TriplesMaps were each traced back to a specific source table. Every row in Table 8.1 is accounted for by either a structural mapping or an entry in Table 8.3's deliberate exclusions. Mapping correctness is therefore assessed as complete for the scope this thesis defined for itself, rather than complete in an absolute sense.

RDF Generation and Knowledge Graph Construction –Materialisation and upload succeeded without manual intervention across the full mock dataset. The pipeline's built in validation step means a malformed or incomplete run would have stopped before reaching AllegroGraph rather than silently uploading partial data. Each side's outcome is additionally tracked and reported independently, so a failure on one side does not silently abort the other. This matters for a deployment in which the two repositories are hosted and owned by different organisations This satisfies NFR1 and NFR2 in practice, not only in design.

Query Execution – All 6 SPARQL queries (Section 8.9) executed successfully. The results from the queries were consistent with how the mock dataset was constructed. The two mismatched pregnancy records and the one non matched referral

were correctly excluded from their respective positive matches rather than incorrectly linked. This indicates that the matching logic is behaving as a genuine filter rather than over matching everything it is given.

Scalability – NFR6 remains the one requirement that the thesis can't demonstrate. The structural mappings place no hardcoded limit on the number of CHUs or facilities. The reasoning for why it should scale has already been discussed (Section 8.8.6). But the architecture has only ever been run against a single CHU and a single hospital. This is a limitation, rather than a failure of this thesis.

9.3 Use Case Evaluation

This section evaluates the 5 FRs (Section 7.2.1). It also shows the two ways in which they were shown to interact, against the practical maternal healthcare use cases. The mechanism and raw results are mentioned in Chapter 8. The sole purpose here is to interpret what that result meant.

9.3.1 Cross System Client Matching

This subsection evaluates FR1 (Section 7.4.1), results demonstrated in Section 8.9.1. The ability to recognise corresponding individuals across independent healthcare systems represents one of the fundamental requirements for achieving semantic interoperability. Without reliable client matching, healthcare information remains fragmented. This prevents healthcare providers from obtaining a complete view of an individual's maternal healthcare journey.

Results of the cross linking system confirms the foundation requirement (Section 8.9.1). Of the 20 CHU clients in the dataset, 17 were successfully resolved to a corresponding hospital patient. The remaining 3 clients returned no match at all. A client with no corresponding hospital record simply has not yet been seen at the hospital in this dataset. The 17 successful demonstrate that name and date of birth matching works. But the framework makes sure that the records that could not be semantically reconciled remained independent. This reflects realistic healthcare situations where community clients may never present at a referral hospital or where corresponding hospital information is unavailable. The framework preserves the integrity of the available data rather than generating artificial relationships between unrelated records.

These findings indicate that the proposed semantic interoperability framework successfully addresses one of the primary challenges identified during the stakeholder analysis by enabling meaningful cross system client identification.

9.3.2 Syphilis Detection

This subsection evaluates FR2 (Section 7.4.2), demonstrated in Section 8.9.2. This is also the use case introduced in Section 5.5 and grounded in the field finding recorded in Section 6.3.5. One of the main objectives of this research was to detect syphilis and improve maternal healthcare by connecting information collected within community healthcare environments with subsequent clinical observations recorded at referral hospitals. In the current workflow these records are maintained in isolation, making it difficult to reconstruct the complete clinical pathway of an individual patient. At the general level, every one of the 6 VDRL-positive clients in the dataset is fully traceable from the CHU side once FR1 has resolved their identity. At the direct level, only 4 of those 6 cases also carry a CHU recorded danger sign referencing sexual violence. This is why the thesis produces the more specific direct link. The remaining 2 positive cases were identified through routine compulsory antenatal screening with no CHU side signal preceding them. They are therefore reachable only through the general route. Not every syphilis case has a community level precursor a CHW could have recorded, and the framework does not need to invent one. What the direct link demonstrates is something the general chain alone could not, i.e. a specific community level disclosure and a specific hospital confirmed result can be connected as the same clinical event, not merely as two facts about the same person. This enabled community observations and subsequent hospital investigations to be interpreted together.

From a clinical perspective this integrated representation provides a more comprehensive understanding of maternal healthcare than either dataset can provide independently. Community information recorded during household visits becomes directly associated with hospital laboratory investigations. This supports improved continuity of care and facilitates more complete clinical interpretation.

9.3.3 Referral Traceability

This subsection evaluates FR3 (Section 7.4.3), demonstrated in Section 8.9.4.

Referral traceability was identified as a significant interoperability challenge within existing maternal healthcare workflows. Referrals play an essential role for transferring patients between community healthcare services and referral hospitals, but the absence of semantic integration frequently limits the ability to determine whether referrals can be followed across organisational boundaries.

From the 7 CHU referrals in the dataset, 6 were successfully matched to a corresponding hospital referral record within the defined time window. This match rate turns what was previously an informal, unverifiable claim on either side into a specific, queryable outcome. 6 of 7 referrals can now be confirmed as having reached the hospital as reported, rather than the continuity of care depending on one party's rec-

ollection over the other's. The remaining unmatched referral is addressed separately in Section 9.3.6.

9.3.4 Pregnancy Record Reconciliation

This subsection evaluates FR4 (Section 7.4.4), demonstrated in Section 8.9.3.

The pregnancy tracked twice problem (Section 6.3.4) is resolved for 15 of the 17 hospital matched clients. The independently maintained CHU and hospital pregnancy records are now linked as the same event rather than existing as two disconnected records. The remaining 2 cases had the expected delivery dates recorded more than 14 days apart between the two systems. This confirms that the matching window would correctly reject a pair of records that describe genuinely different estimates rather than linking them regardless.

By establishing semantic relationships between these independently recorded healthcare events, the framework provided a unified representation of pregnancy information without requiring either healthcare system to alter its existing data model. The ability to reconcile pregnancy records supports a more comprehensive understanding of maternal healthcare by allowing information collected throughout different stages of pregnancy to be interpreted collectively rather than as isolated observations.

These findings indicate that the proposed semantic interoperability framework effectively supports pregnancy record reconciliation and addresses one of the principal fragmentation challenges identified during the analysis of maternal healthcare workflows.

9.3.5 End to End Continuity of Care

This subsection evaluates FR5 (Section 7.4.5) and is demonstrated by the combined chain query in Section 8.9.5. While the previous use cases evaluated individual interoperability functions, this use case assessed whether the proposed framework could represent an entire maternal healthcare pathway spanning multiple healthcare environments. Achieving continuity of care requires more than linking individual records. It requires preserving the relationships between healthcare events as a patient progresses through the healthcare system.

The combined query result returns exactly the same 4 clients already identified as the direct link syphilis cases. This demonstrates that FR1, FR2, and FR3 operate as a single continuous chain on the same real cases, not as three separate capabilities that happen to share a dataset. For each of the 4 clients a CHU recorded disclosure of sexual violence, a CHU referral, that referral's confirmed match to a hospital referral, and a confirmed positive hospital syphilis result are all reachable from a

single starting point in the graph. This is the practical realisation of the continuity of care goal.

The clinical significance of this result is specific rather than general. Section 5.5 established that the majority of adverse maternal syphilis outcomes worldwide do not occur in women who were never screened but in women who were screened and then lost to follow up. The four clients returned by this query are not a demonstration of syphilis detection in the abstract. They are cases in which a community level concern, a referral, and a hospital diagnosis are shown to remain connected as one record rather than becoming three disconnected facts once they cross an organisational boundary.

9.3.6 Traceability Without Cross System Correspondence

This subsection is demonstrated by the query in Section 8.9.6, and returns to the CHW hospital disagreement over referral communication first raised in Section 6.6.

Real world healthcare environments frequently contain incomplete, delayed, or missing information. Consequently, an effective semantic interoperability framework should accurately represent situations in which corresponding healthcare records do not exist, rather than assuming that every community referral must have an associated hospital record.

That one CHU referral without a corresponding hospital record, concerning a client referred for suspected malaria in pregnancy is the case that most directly tests whether the framework preserves referral records that could not be semantically linked to corresponding hospital information. The result confirms that it does. The referral's reason and date remain fully queryable through its direct link to the client, entirely independent of whether a hospital side match was ever found. This distinction matters in practice. A CHU referral for which no hospital record can be located is not evidence that the referral failed or that the client never received care. It may mean that the hospital referral that has not yet been logged, a visit that has not yet occurred, or a hospital record that falls outside the time window used for matching. What this requirement guarantees is that such a referral is never silently lost from the record simply because a cross system match could not be confirmed. This is arguably the more realistic outcome for a production deployment.

9.4 Stakeholder Feedback

Active communication was maintained throughout this thesis with stakeholders from both organisations, Community Health Workers at KCEI and hospital staff at Pumwani Maternity Hospital. Consequently, feedback over the resulting RDFs was also sought from both organisations. The goal was to assess the practical value

and usability of the proposed framework from the perspective of the people who would actually use it.

Feedback was received from the stakeholder representative from the hospital and CHWs. The stakeholders were given access to the AllegroGraph repository for their institution which contained the RDF knowledge graph outputs generated by the framework to be queried, via their authentication access and password.The discussion below proceeds with the hospital side feedback that was obtained.

The hospital stakeholder's feedback directly affirmed the value of a semantic, FAIR compliant approach to referral management.

"The VODAN semantic interoperability framework (using RDF triples) would substantially improve referral management by ensuring that community-generated data are FAIR (Findable, Accessible, Interoperable, and Reusable) and semantically linked across the continuum of care. Knowing what a Community Health Worker (CHW) observed before referral would provide valuable clinical and contextual information during the patient's first facility encounter, enabling more informed triage, continuity of care, and faster decision-making."

This directly addresses FR2, FR3 and FR1 and the continuity of care result (Section 9.3.5). It identifies the pre referral CHU information the framework connects to a hospital record, and frames it in terms of clinical usefulness at the point of care rather than only as a technical linkage.

The stakeholder also raised a condition.

"Rather than creating clutter, this information would be highly beneficial if presented in a concise, standardized format and automatically integrated into the intake workflow. Compared with the current referral process—which is largely paper-based, fragmented, and dependent on manual documentation that is later aggregated into government Health Information Management (HIM) systems—the VODAN approach would require only minimal workflow adjustments while significantly enhancing efficiency."

This is an important distinction between two separate contributions this thesis makes. It acknowledges the value of the semantic integration, connecting CHU and hospital records into a single queryable graph. But it also identifies that this value is conditional on how that information is ultimately surfaced to a clinician at the point of care. The architecture developed in this research produces a queryable RDF knowledge graph (Chapter 8), demonstrated through direct SPARQL queries (Section 9.3). But it does not by itself produce the concise, standardized, intake integrated presentation layer the stakeholder describes. It is a genuine boundary of scope, since building a clinical facing presentation or dashboard layer on top of the semantic architecture is a distinct engineering task from the ontology, mapping, and knowledge graph work this thesis focuses on. The deployment components that make the integrated datasets FAIR in practice have since been put in place. Both

repositories are deployed on the participating institutions' own infrastructure, and their metadata is registered in a FAIR Data Point and exposed through the Africa Health Data Space (Section 3.8.6). What remains outside the scope of this thesis is the clinical presentation layer the stakeholder describes. Building a concise, intake integrated view on top of the semantic architecture is a distinct engineering task from the ontology, mapping, and knowledge graph work undertaken here.

The stakeholder also identified the referral feedback loop as a further area of value.

"It would also strengthen bidirectional communication by allowing facilities to provide referral outcomes back to CHWs, thereby closing the referral feedback loop, which is frequently absent in routine practice."

This is a major point of agreement with the field finding where the CHW and hospital accounts disagreed on whether referral outcomes are actually communicated back in practice (Section 6.6). The stakeholder raises the same gap and describes it as frequently absent in routine practice. This strengthens the case that this is a genuine, recognised weakness in the current referral pathway rather than an artefact specific to the particular disagreement observed.

The stakeholder further described how health facility data is currently used.

"Health facilities routinely generate large volumes of data that are primarily collected to satisfy reporting requirements and are subsequently aggregated into national HMIS platforms, with little opportunity for frontline providers to derive actionable insights from the data they produce. By applying semantic interoperability and FAIR principles, data become immediately reusable at the point of care while remaining available for surveillance, planning, and research."

This observation extends the value of the framework beyond the specific use cases evaluated (Section 9.3). It frames FAIR compliant semantic interoperability not only as a solution to the CHU hospital linkage problem this thesis set out to address, but as a way of making data that hospitals already collect and that is otherwise consumed largely by upward reporting to systems such as DHIS2 (Section 5.6) directly useful to the frontline staff who generate it.

The stakeholder's assessment becomes more critical when addressing the matching mechanism.

"Automatic linkage of community and facility records based on identifiers such as patient name and date of birth is a practical starting point, although additional identifiers (e.g., national ID, phone number, household identifier, or probabilistic matching algorithms) together with human verification for uncertain matches would improve confidence and reduce mismatches."

This is a fair and informed critique, and one this thesis largely agrees with. It is not however an oversight in the design of FR1. The choice to match on full name and date of birth alone was the direct consequence of a gap already established

(Section 6.3.1). It is not an arbitrary simplification. The structural comparison found that while the CHU dataset does record a national ID, the hospital system never captures it. While the hospital system generates its own national patient identifier, the NUPI, this is never recorded on the CHU side. Neither system in its current form shares a common identifier with the other. The phone number is no better. As we established (Section 6.3.2), the CHU records the household head's phone number rather than the individual client's, while the hospital records the patient's own number directly. This means that the two fields frequently do not describe the same person even when both are present. A household identifier is similarly unavailable on the hospital side, since hospitals do not organise clinical records around households at all (Section 7.3.2). Therefore, full name and date of birth were the only descriptive information both systems independently and reliably record about the same person (Section 7.4.1).

The stakeholder's second suggestion of probabilistic matching combined with human verification for uncertain cases is methodologically sound and reflects a scope rather than an unconsidered gap. Introducing probabilistic matching would require calibrating match thresholds and evaluating the resulting trade off between false positive and false negative matches. This task depends on having a realistic volume and variety of real world data to tune against. Introducing a probabilistic layer in this thesis would have added methodological complexity without the real world data needed to validate it meaningfully.

The stakeholder also suggested extending the clinical scope of the framework.

"To maximize value, the system should also integrate laboratory results, diagnoses, treatment plans, referral outcomes, follow-up visits, and patient outcomes, thereby creating a complete longitudinal health record that supports clinical care, community follow-up, and evidence-informed decision-making across all levels of the health system."

This suggestion aligns with the scope boundary already established (Table 8.3). Fields and tables that exist in the H2 database but were intentionally left unmapped, since they fell outside this thesis's syphilis focused scope (Section 5.5). The stakeholder's recommendation to extend coverage toward a complete longitudinal record supports the case that this extension is a natural and low risk next step. NFR4 also establishes that the ontology can accommodate new concepts without redesigning the existing model.

When asked for the stakeholder's opinion on the framework, the response was unambiguous.

"Yes... and it is going to revolutionize data management and shareability/interoperability."

The Community Health Worker representative reviewed the walkthrough guide and the demonstration and confirmed that the proposed functionality and responses

to the evaluation questions were satisfactory. No additional qualitative comments or suggested improvements were provided beyond this confirmation. The more detailed discussion above focuses primarily on the feedback received from the hospital stakeholder.

Taken as a whole this feedback supports the framework on substantially the same terms it set out to be evaluated on. It confirms that the specific interoperability problems this thesis addressed, particularly the referral feedback loop identified (Chapter 6), are recognised as real problems by hospital staff working with the current system. It identifies both a genuine strength of the architecture, the value of connecting CHU and hospital information at the point of care, and two specific areas, matching robustness and clinical scope, where the stakeholder's suggestions align closely with limitations this thesis already identifies in itself.

9.4.1 Response to the Deployed System

The first round above was conducted with the two stakeholders I directly worked with during this research, one from each organisation. Both had been involved from the field visit onwards, and contributed to the workflow accounts the requirements were derived from, and both had a direct interest in the problem the framework addresses. Their assessment is the most detailed evidence this evaluation has.

The second round was a recorded demonstration of the deployed architecture and was shared with the VODAN Africa, KCEI, and Pumwani groups, reaching practitioners and researchers working on FAIR health data who had no involvement in this research. The response confirmed that the deployed architecture, the AHDS registration, and the query results shown were understood and accepted as working as described.

This was an informal indication of approval rather than a structured evaluation. Due to time constraints, no formal meetings or Q&As could be organised for more detailed feedback. It established that the deployed system was seen by an audience outside this research and recognised as functioning as claimed.

9.5 Summary

Taken together, the technical and use case findings indicate that the proposed framework successfully addresses the interoperability challenges identified in Chapter 6, and does so in a way that is honest about where its evidence is strongest and where it remains untested. The stakeholder's feedback adds a third strand of evidence to these findings (Section 9.4). The hospital stakeholder's response confirmed from a clinical and operational perspective rather than a technical one that the specific problems this thesis set out to solve are recognised as real problems in practice.

Chapter 10

Discussion and Conclusion

10.1 Discussion

This chapter draws together the technical and use case findings from Chapter 9, weighs them against the hypothesis and literature gaps set out earlier in this thesis. It closes with the limitations of the work as it currently stands.

10.1.1 Revisiting The Hypothesis

The hypothesis stated that a FAIR compliant semantic interoperability framework could improve the integration and usability of maternal healthcare data between Community Health Workers and referral hospitals in Kenya. Specifically that RDF, linked data, and standardised healthcare vocabularies would reduce fragmentation between the two systems, and that this would translate into better continuity of care (Section 1.6).

The evidence in Chapter 9 supports the integration half of this hypothesis directly. FR1 through FR4 each demonstrated that CHU and hospital records that previously could be fragmented across two systems with no shared identifier automatically resolved into a single connected representation without either system's original structure being altered. The combined continuity of care query (Section 9.3.5) is the clearest evidence of this. A community level disclosure, a referral, and a hospital confirmed diagnosis were shown to be reachable as a single connected record rather than three disconnected facts spread across two systems. This is a direct realisation of the reduced fragmentation the hypothesis anticipated.

The continuity of care half of the hypothesis is also supported. Referral traceability turned an informal, disputed claim between CHW and hospital accounts into a specific, queryable outcome for 6 of 7 referrals. The framework's handling of the one unmatched referral (Section 9.3.6) shows that continuity is preserved even when a cross system match cannot be confirmed. Therefore, the hypothesis is supported on its technical and integration claims.

The usability component of the hypothesis can now be addressed in part, though not conclusively. The hospital stakeholder's response offers direct evidence that a stakeholder working within the system judged the approach favourably. Describing it as something that would "revolutionize data management and shareability/interoperability" and require only minimal adjustment to existing workflow (Section 9.4). This is a meaningful data point in favour of the usability claim. The hypothesis is therefore supported on its technical and integration claims with confidence, and on its usability claim is suggestive rather than conclusive.

10.1.2 Position Against Literature Gaps

Four research gaps were identified (Section 1.3). This section revisits each in turn against what this thesis has actually demonstrated.

The first gap was that the majority of healthcare interoperability research focuses on hospital to hospital integration rather than community to hospital integration. This thesis addresses that gap directly, since its entire architecture is built around resolving CHU and hospital records specifically. Lina's original ontology is a hospital to hospital interoperability model and had no concept of a household, a community health unit, or a CHW observation until this thesis extended it (Section 7.3.2).

The second gap concerned the heavy dependence of existing approaches on FHIR and structural interoperability rather than semantic interoperability. Chukwu et al.'s work in Nigeria (Section 4.2) is the clearest point of comparison here [39]. Their FHIR resource profiles standardised how referral data moves between a primary health care centre and a hospital. But it stopped at the structural level, since no ontology or semantic mapping was introduced to reconcile informal community terminology with hospital coding. This thesis's ontology and OBDA layer is precisely the semantic layer that work did not attempt, applied to a maternal health CHW to hospital context.

The third gap was that limited research combines FAIR principles, RDF modelling, standardised vocabularies, and knowledge graphs specifically in maternal healthcare systems. Zhang et al.'s cancer registry work (Section 4.3) demonstrated that the technique of Ontology Based Data Access using Ontop is a mature and proven method [42]. But it was applied to a single country and to a high resource cancer registry setting. This thesis applies the same technical approach to a low resource domain. Maternal health data spanning a community health unit and a referral hospital in a low resource African context. The comparison is not that this thesis's results are stronger in absolute terms as the two studies operate over entirely different data and domains. But the combination of FAIR principles, RDF, and OBDA identified as underused for maternal healthcare has now been demonstrated to work in that specific context (Section 1.3).

The fourth and most specific gap was the low focus on semantic interoperabil-

ity for maternal syphilis continuity of care between healthcare levels. Amare et al.'s cross border work (Section 4.5) is the closest existing study to this thesis [50]. But it operates at the level of data sovereignty and data visiting across national borders. This thesis instead resolves the problem at the level of individual record linkage within a single country between two specific, independently operated systems. Tegenaw et al.'s referral pathway study in Ethiopia (Section 4.5) is a useful point of comparison for what a manual approach to this same class of problem looks like in practice [48]. Comparing paper referral logbooks against hospital patient files by hand, they found only 13.9 percent of referred patients could be perfectly matched. Because this thesis's FR1 result of 17 of 20 clients automatically resolved is produced on a synthetic dataset rather than real world records, it is not directly comparable as a measure of accuracy. However, this thesis demonstrates that the same class of matching problem Tegenaw et al. solved manually can be automated through ontology based entity resolution, which is the structural contribution this thesis makes to that specific gap.

10.1.3 Positioning Against the Humanitarian Data Space Evaluation

The fourth evaluation component set out in Section 3.9.4 is reported here. The evaluation sets out 5 design decisions every application converting structured input into RDF must resolve,a strengths and weaknesses assessment of each system evaluated, and an 8 step process toward a fully governed, federated data space[33]. Because that assessment describes the state of COMPASS before this thesis extended it, positioning this thesis against it is a measurement of what has changed rather than a comparison against an external benchmark. Positioning this thesis's architecture against all 3 components clarifies what has been achieved and what remains open, beyond this thesis's own stated nonfunctional requirements. It is a comparison against an external framework rather than a measure of change within a single system, since the architecture presented here was built independently of the COMPASS implementation the evaluation assesses (Section 3.3).

Strengths and weaknesses. Among the applications assessed is COMPASS, whose strengths are its use of R2RML and its SNOMED CT/ICD-10 semantics. It's weaknesses are addressed as the absence of provenance records, of a deidentification policy, and of an access policy or audit log.[33]

Both strengths are retained here. The transformation layer remains declarative R2RML executed by Ontop (Section 8.4). The clinical semantics inherited from the hospital ontology are preserved intact, including the LOINC alignments carried on several datatype properties (NFR5, Section 7.2.3). What this thesis adds is a second layer alongside them rather than a replacement for them (Section 7.3.1).

The three weaknesses are not resolved to the same degree. Provenance remains entirely unaddressed and is discussed below. The deidentification weakness is displaced rather than answered. Because this thesis operates exclusively on synthetic data (Section 8.2), no deidentification was required during the research. This means the policy question is deferred to deployment rather than settled by it. The access weakness is partially addressed. NFR3 requires authenticated access to each repository. The two repository split means each producer controls access to its own data rather than to a merged store (Section 8.7). This is access control, not the machine readable access policy the evaluation calls for.

The evaluation also highlights the use of FHIR R5 profiles as a future opportunity, while identifying clinical reidentification and the long term maintenance of the environment as potential threats. The opportunity to adopt FHIR R5 aligns with the approach taken in this thesis rather than replacing it. Following Gebreslassie et al. (2023), this research treats FHIR as the standard for exchanging data within each healthcare facility. The semantic layer just provides the shared meaning that FHIR alone cannot guarantee (Section 4.2) [40]. The risk of reidentification is more significant and should be acknowledged. In FR1 records are linked by matching a patient's full name and date of birth (Section 7.4.1), which is a form of record linkage. Since the system connects records from different data sources that belong to the same individual, there is always a possibility of reidentification. This architecture reduces that risk by keeping each organisation's data in its own repository under its own access controls with only authorised users able to access the data. The protection therefore comes from the deployment model and access controls, not from the matching process itself. Finally, the threat of an unmaintained environment is relevant to the deployment handover discussed in Section 9.4, where responsibility for maintaining the system is transferred to the deploying organisation.

Design decisions. D1 (input structuring) is resolved as early relational structuring via the H2 schema (Section 8.2). D2 (controlled vocabularies) is resolved as a split by layer rather than as a single choice. The hospital layer inherits external clinical semantics from the ontology it extends, including SNOMED CT and LOINC alignment. While the CHU layer's locally defined free text values are deliberately not forced into an external clinical vocabulary (Section 7.3.1). This split is the direct consequence of the semantic interoperability requirement, which holds that neither side's terminology should be forced into the other's (Section 7.2.2). D3 (ontology alignment) is resolved primarily through the introduction of new classes and properties instead of `owl:equivalentClass` alignment to an external ontology. This was a consequence of what exists to align with. The CHU concepts introduced have no counterpart in SNOMED CT, LOINC, or the hospital ontology being extended. So retaining them as constructs declared within the shared data space ontology was the only available route. Although several inherited datatype properties are

aligned to LOINC via `skos:closeMatch` (NFR5, Section 7.2.3). D4 (data space ontology structuring). A single, centrally structured shared ontology within which each side's original vocabulary is preserved as a distinct layer rather than forced into the other's (Section 7.3.1). D5 (RDF specification) is resolved as declarative via R2RML mappings executed by Ontop, rather than imperative code. This holds for the transformation itself. This includes all 5 entity resilution mapping, which are expressed as SQL within R2RML rather than as post hoc code (Section 8.4.2). One step in the pipeline is imperative. The semantic post processing that removes foreign inferred triples before upload is Python rather than a declarative artefact (Section 8.6).

Eight step process. Table 10.1 summarises this thesis's position against each of the eight steps. Each unmet or partially met step is given a reason below rather than only a status.

Table 10.1: Position Against the Eight Step Data Transformation Process

Step	**Status**	**Basis / reason where not met**
1. Specify	Partial	Ontology and mappings are versioned declarative artefacts. No producer published data dictionary or SHACL shape, as neither schema was published by its producer.
2. Validate	Partial	Pipeline validates table presence and record counts pre-transformation. No SHACL shape, as Step 1 has produced no contract to validate against.
3. Transform	Met	All mappings declarative R2RML executed by Ontop; no machine generated mappings requiring approval.
4. Provenance	Not met	No PROV-O instrumentation. Inherited from the predecessor system, where the evaluation names it the central missing element.
5. Identify	Not met	No versioned, checksummed, citation bound release. RDF is regenerated per run rather than published. Persistent `w3id.org` namespace is in place at entity level.
6. Register	Partial	Both repositories are registered in the FDP and discoverable via AHDS. No DCAT distribution or machine readable access policy accompanies the registration, as the data is synthetic and no formal access policy has been negotiated.
7. Federate	Partial	Native SPARQL `SERVICE` across two repositories; correspondence properties used in place of an `owl:sameAs` registry, deliberately.
8. Renegotiate	Not met	Consequence of Step 2: no validation step exists to emit a drift signal for renegotiation to act on.

The evaluation's own conclusion frames future work across three horizons. Short term (governance and fidelity), mid term (provenance and reproducibility), and long term (federation and standards).[33] Read against this structure, this thesis's own position is uneven across the three.

At the short term horizon, this thesis has not published a machine-actionable data dictionary or a SHACL shape for either side of the architecture. It has not externalised a deidentification policy as a versioned artefact separate from the pipeline code. Nor is a persistent identifier assigned to each dataset snapshot, consistent with Step 5 above. Two items at this horizon are already met. Operational credentials are read from environment variables at runtime rather than recorded in the pipeline source or in version control. This satisfies the requirement to keep credentials out of the repository, which this thesis treats as immediate given the sensitivity of the data involved (NFR3, Section 8.8.3). The horizon's first item, replacing non-dereferenceable IRIs with persistent namespaces is also met. Every entity and every ontology term is identified within the `https://w3id.org/dhs/ontology#` persistent namespace rather than a local one.

At the mid term horizon, provenance and reproducibility. This thesis meets one item already rather than deferring it. Mappings are declarative R2RML rather than hand coded. So the migration this horizon calls for is not needed here. Secondly, no mapping in this architecture was machine generated. Hence, there is nothing requiring placement under a human approved registry. PROV-O instrumentation and versioned, checksummed RDF snapshots under continuous integration remain unaddressed, consistent with Provenance's status above.

At the long term horizon, federation and standards. This thesis's two repository architecture already uses native SPARQL `SERVICE` federation. The specific mechanism this horizon identifies as a long term goal, though as noted above, this is achieved through purpose built correspondence properties rather than an `owl:sameAs` reconciliation registry. Both repositories are already registered in the FDP and discoverable through the AHDS (Section 3.8.6), which is itself ahead of where this horizon places registration. What remains open at this horizon is packaging releases with a FAIR Digital Object identifier, which remains future work and was already independently identified as such following stakeholder feedback (Section 9.4).

Taken together, this thesis sits ahead of this evaluation's own long term horizon on federation, roughly level with its mid term horizon on transformation, and behind its short term horizon on governance and fidelity. This uneven profile reflects where this thesis's scope was deliberately concentrated, semantic interoperability and entity resolution, rather than a uniform state of completeness across the framework. The two strengths recorded for the predecessor system are carried forward intact. From its three weaknesses, one is partially addressed, one displaced by the use of synthetic

data, and provenance is left open and identified here as the most consequential of the remaining gaps.

10.1.4 Limitations

There are several limitations.

The dataset used throughout this thesis is synthetic. Both the CHU and hospital tables were populated with mock data constructed to mirror the real structure of KCEI's and Pumwani's systems. The match rates and query results therefore describe how the architecture behaves against data built to resemble the real world, not against the real world itself.

The architecture has only been tested against a single CHU and a single hospital. NFR6 remains untested rather than demonstrated (Section 8.8.6). While the structural mappings place no hardcoded limit on the number of facilities, this remains a design property rather than an observed result.

The clinical scope of the mapping layer is deliberately restricted to syphilis. Table 8.3 lists several fields and two entire tables that exist in the H2 schema but were never mapped.

FR1's client matching relies on an exact match of full name and date of birth. This is the only descriptive information both systems independently record about the same person (Section 7.4.1). But exact matching offers no tolerance for the kind of minor recording differences, such as a spelling variation or a transposed date, that are common in real world data entry and could cause a genuine match to be missed.

The 7 day and 14 day matching windows used in FR3 and FR4 were chosen because a referral being made and a client being seen, or two independently estimated delivery dates are not necessarily recorded on the same day. No specific clinical or operational justification exists for these exact values. They can be best justified as a starting point for this dataset rather than a validated threshold.

Finally, the MOH 100 paper referral form's Section B report back mechanism remains entirely outside the digital framework this thesis proposes (Section 6.5). Whatever coordination it enables between a CHW and a facility leaves no trace in either the CHU or hospital dataset, and cannot currently be queried or reasoned over as part of this architecture.

10.2 Future Work

The Humanitarian Data Space evaluation framework discussed in Section 10.1.3 already identifies a distinct set of infrastructure and governance extensions, schema validation, provenance tracking, versioned data releases, and FAIR Data Point registration which are not repeated here. The extensions below concern the clinical and

organisational scope of the architecture itself, rather than the data governance layer surrounding it.

Although the proposed framework successfully demonstrates semantic interoperability between Community Health Units and referral hospitals, several opportunities remain for future research and development.

The most immediate extension of this work could involve evaluating the framework using real healthcare data within an operational healthcare environment. Such an evaluation would enable assessment of the framework under realistic clinical conditions, including larger datasets, heterogeneous data quality, and continuously changing healthcare information. This would provide additional evidence regarding the scalability, robustness, and practical applicability of the proposed architecture.

Another direction that could be explored is extending clinical scope. The `HIV_SCREENING` and `TREATMENT_AND_PREVENTION` tables already present in the H2 schema but deliberately left unmapped (Table 8.3), are a natural next extension. Since NFR4 establishes that new concepts can be added to the ontology without redesigning the existing model, extending the mapping layer to cover these tables would test that extensibility claim against a second clinical use case beyond syphilis. This direction is also supported by the hospital stakeholder. The stakeholder independently recommended integrating laboratory results, diagnoses, treatment plans, referral outcomes, follow up visits, and patient outcomes to build toward a complete longitudinal health record (Section 9.4).

Future research could also extend the ontology beyond ANC to involve post natal care or child healthcare, chronic disease management, immunisation programmes, or communicable disease surveillance. Expanding the ontology would allow the framework to support broader healthcare interoperability while maintaining the semantic modelling principles developed in this research.

Finally, the proposed framework could be evaluated within larger multi organisational healthcare networks involving additional Community Health Units, referral hospitals, and national health information systems. Such studies would provide further insight into the scalability of ontology based semantic interoperability within national healthcare infrastructures and contribute towards broader implementation of FAIR healthcare ecosystems.

10.3 Conclusion

This research sets out to answer the question of how a FAIR compliant semantic interoperability framework can support the integration and usability of maternal healthcare data between Community Health Workers and referral hospitals in Kenya. It develops a maternal health ontology extending an existing hospital model, an Ontology Based Data Access mapping layer connecting both organisation's relational

data to it without modifying either source system, and a knowledge graph capable of answering cross system queries that neither the CHU nor the hospital's own data could answer independently.

The evaluation showed that this architecture successfully resolves the specific interoperability gaps identified, the absence of a shared identifier between CHU and hospital records, the lack of a structured way to trace a syphilis case across the two systems, the disagreement between CHW and hospital accounts of referral communication, and the independent, disconnected tracking of the same pregnancy by both systems (Chapter 6). The combined continuity of care result in particular demonstrated that these are not four isolated capabilities but a single connected chain, capable of tracing a community level disclosure through to a hospital confirmed diagnosis as one continuous record (Section 9.3.5).

The thesis also contributes towards the specific gaps identified in the literature (Section 1.3). Where existing FHIR based work in comparable low resource settings stops at structural interoperability. This framework adds the semantic layer that work did not attempt. Where Ontology Based Data Access has been proven effective in high resource, single country domains. This thesis demonstrates the same technique applied to a low resource, multi tier maternal health context. And where the closest existing work on FAIR compliant maternal syphilis interoperability operates at the level of cross border data sovereignty. It resolves the problem one level down, at the level of individual record linkage between two independently operated systems within a single country.

The stakeholder feedback obtained during evaluation (Section 9.4) supports these technical findings from a clinical and operational perspective. A hospital stakeholder confirmed that the specific problems this framework addresses are recognised gaps in current practice. The proposed approach was judged on the capability of meaningfully improving how referral data is managed and shared. The same stakeholder also independently identified two of the limitations this thesis had already recognised in itself. The need for stronger entity resolution and broader clinical scope. Feedaback was obtained form both stakeholders, but the depth of the feedback differed. The CHW confirmed the functionality of the existing workflows, and also indicated the successful operation of the walkthrough. However, the CHW feedback consisted primarily of confirmation rather than detailed qualitative comments.

Overall, this research demonstrates that FAIR compliant semantic interoperability, built through ontology extension, Ontology Based Data Access, and knowledge graph representation, can be applied concretely to the specific problem of connecting community and hospital level maternal healthcare data in Kenya. The architecture is technically sound, addresses the use cases it was designed for, and is judged useful by the two stakeholder groups it was built to serve. What remains is testing this architecture against real data, at greater scale, and with a more concrete balanced

stakeholder evaluation. s

List of Abbreviations

AHDS	Africa Health Data Space
AI	Artificial Intelligence
ANC	Antenatal Care
API	Application Programming Interface
AUN-FOS	Africa University Network on FAIR Open Science
CHU	Community Health Unit
CHW	Community Health Worker
CHP	Community Health Promoter
COMPASS	Contextual Operations Multi-Point Access to Secure SPARQL
DSR	Design Science Research
EHR	Electronic Health Record
FAIR	Findable, Accessible, Interoperable and Reusable
FHIR	Fast Healthcare Interoperability Resources
FDP	FAIR Data Point
HIS	Health Information System
HL7	Health Level Seven
HIV	Human Immunodeficiency Virus
ICD	International Classification of Diseases
KCEI	Kamukunji Community Empowerment Initiative
KG	Knowledge Graph
LIACS	Leiden Institute of Advanced Computer Science
LOINC	Logical Observation Identifiers Names and Codes
OBDA	Ontology-Based Data Access
OMOP	Observational Medical Outcomes Partnership
OWL	Web Ontology Language
RDF	Resource Description Framework
R2RML	RDB to RDF Mapping Language
SPARQL	SPARQL Protocol and RDF Query Language
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
URI	Uniform Resource Identifier

WHO	World Health Organization

Bibliography

[1] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. doi: 10.1038/sdata.2016.18.

[2] Celia Alvarez-Romero, Alicia Martínez-García, A. Anil Sinaci, Mert Gencturk, Eva Méndez, Tony Hernández-Pérez, et al. FAIR4Health: Findable, accessible, interoperable and reusable data to foster health research. *Open Research Europe*, 2:34, 2022. doi: 10.12688/openreseurope.14349.2.

[3] A. Anil Sinaci, Francisco J. Núñez-Benjumea, Mert Gencturk, Malte-Levin Jauer, Thomas Deserno, Catherine Chronaki, et al. From raw data to FAIR data: The FAIRification workflow for health research. *Methods of Information in Medicine*, 59(S 01):e21–e32, 2020. doi: 10.1055/s-0040-1713684.

[4] Nigel Shadbolt, Wendy Hall, and Tim Berners-Lee. The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, 2006. doi: 10.1109/MIS.2006.62.

[5] Peter Munyao Kithuka, Beatrice Amy Nesidai Kithuka, Erick Kioko Mekala, Lydia Muthamba Munuve, Geoffrey Juma Malachi, and Daisy Khayeri Omutanyi. Barriers and enablers of health information systems interoperability in kenya: A scoping review. *East African Journal of Health and Science*, 8(2): 488–505, 2025. doi: 10.37284/eajhs.8.2.3775.

[6] Philomena Ngugi, Ankica Babic, and Martin C. Were. A multivariate statistical evaluation of actual use of electronic health record systems implementations in kenya. *PLOS ONE*, 16(9):e0256799, 2021. doi: 10.1371/journal.pone.0256799.

[7] Abdullahi Abubakar Kawu, Rens Kievit, Adamu Abubakar, Mirjam van Reisen, Dympna O'Sullivan, and Lucy Hederman. Exploring the integration of patient generated health data in a FAIR digital health system in low-resourced settings: A user-centered approach. In *Proceedings of the 4th African Human Computer Interaction Conference (AfriCHI 2023)*, pages 215–220, New York, NY, USA, 2024. ACM. doi: 10.1145/3628096.3629059.

[8] Health Policy Watch. Kenya's high court suspends us health deal as civil society urges african leaders to ensure 'fair terms'. https://healthpolicy-watch.news/kenyas-high-court-suspends-us-health-deal-as-civil-society-urges-african-leade December 2025. Accessed August 2026.

[9] JURIST. Kenya's health data deal with the US: What the agreement gets right — and what it misses in the age of AI. https://www.jurist.org/commentary/2026/01/kenyas-health-data-deal-with-the-us-what-the-agreement-gets-right-and-what-it- January 2026. Accessed August 2026.

[10] AllAfrica. Kenya: High court suspends $1.6bn kenya-US health cooperation framework over data privacy concerns. https://allafrica.com/stories/202512110268.html, December 2025. Accessed August 2026.

[11] Capital FM Kenya. Court of appeal temporarily lifts high court orders blocking kenya–US health deal. https://www.capitalfm.co.ke/news/2026/05/court-of-appeal-lifts-high-court-orders-kenya-us-health-deal-2026/, May 2026. Accessed August 2026.

[12] Mirjam van Reisen, Samson Yohannes Amare, Reginald Nalugala, Getu Tadele Taye, Tesfit Gebremeskil Gebreselassie, Araya Abrha Medhanyie, Erik Schultes, and Mohamed Mpezamihigo. Federated FAIR principles: Ownership, localisation and regulatory compliance (OLR). *FAIR Connect*, 1(1):63–69, 2023. doi: 10.3233/FC-230506.

[13] Reginald Nalugala, Putu Hadi Purnama Jati, Samson Yohannes Amare, Maxwell Omare, Jacinta Wairimu, Charles Kahiro, William Nandwa, Seth Okeyo, Dennis Kinoti, Aliya Aktau, Albert Mulingwa, and Mirjam van Reisen. Introducing data sovereignty over patient data: Patient data ownership in residence of health facilities in Kenya. In Mirjam van Reisen, Samson Yohannes Amare, Lauren Maxwell, and Munyaradzi Mawere, editors, *FAIR Data, FAIR Africa, FAIR World: Internationalisation of the Health Data Space*, pages 65–96. Langaa RPCIG, Bamenda, 2025. doi: 10.5281/zenodo.15382826.

[14] Mirjam van Reisen, Samson Yohannes Amare, Reginald Nalugala, Getu Tadele Taye, Tesfit Gebremeskil Gebreselassie, Araya Abrha Medhanyie, Erik Schultes, and Mohamed Mpezamihigo. Federated FAIR principles: Ownership, localisation and regulatory compliance (OLR) of sensitive data and personal data. *FAIR Connect*, 1(1):63–69, 2023. doi: 10.3233/FC-230506.

[15] Putu Hadi Purnama Jati, Markus Sintong Tambah Lasroha, and Mirjam van Reisen. Data privacy and interoperability in Indonesia's SATUSEHAT platform.

In Mirjam van Reisen, Samson Yohannes Amare, Lauren Maxwell, and Munyaradzi Mawere, editors, *FAIR Data, FAIR Africa, FAIR World: Internationalisation of the Health Data Space*, pages 361–396. Langaa RPCIG, Bamenda, 2025. doi: 10.5281/zenodo.15382947.

[16] Healthcare Information and Management Systems Society (HIMSS). Interoperability in healthcare. https://www.himss.org/resources/interoperability-healthcare, 2020.

[17] Fouzia Amar, Alain April, and Alain Abran. Electronic health record and semantic issues using fast healthcare interoperability resources: Systematic mapping review. *Journal of Medical Internet Research*, 26:e45209, 2024. doi: 10.2196/45209.

[18] Republic of Kenya. The digital health act, 2023. Kenya Gazette Supplement No. 192 (Acts No. 15); https://new.kenyalaw.org/akn/ke/act/2023/15/eng@2023-11-24, 2023.

[19] Republic of Kenya. The data protection (general) regulations, 2021. Legal Notice 263 of 2021, Kenya Gazette Vol. CXXIV–No. 6, 14 January 2022; https://new.kenyalaw.org/akn/ke/act/ln/2021/263/eng@2022-01-14, 2021.

[20] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001. doi: 10.1038/scientificamerican0501-34.

[21] Richard Cyganiak, David Wood, and Markus Lanthaler. RDF 1.1 concepts and abstract syntax. W3C Recommendation, https://www.w3.org/TR/rdf11-concepts/, 2014.

[22] Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. Linked data on the web (LDOW2008). In *Proceedings of the 17th International Conference on World Wide Web*, 2008.

[23] Boris Motik, Peter F. Patel-Schneider, and Bijan Parsia. OWL 2 web ontology language: Structural specification and functional-style syntax (second edition). W3C Recommendation, https://www.w3.org/TR/owl2-syntax/, 2012.

[24] Steve Harris and Andy Seaborne. SPARQL 1.1 query language. W3C Recommendation, https://www.w3.org/TR/sparql11-query/, 2013.

[25] Rudi Studer, V. Richard Benjamins, and Dieter Fensel. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1-2):161–197, 1998. doi: 10.1016/S0169-023X(97)00056-6.

[26] SNOMED International. What is SNOMED CT? https://www.snomed.org, 2025.

[27] World Health Organization. ICD-11 for mortality and morbidity statistics. https://icd.who.int, 2026.

[28] Regenstrief Institute. LOINC: Logical observation identifiers names and codes. https://loinc.org, 2026.

[29] Guohui Xiao, Davide Lanti, Roman Kontchakov, Benjamin Cogrel, Diego Calvanese, et al. The virtual knowledge graph system Ontop. In *The Semantic Web – ISWC 2020*, volume 12507 of *Lecture Notes in Computer Science*, pages 259–277. Springer, 2020. doi: 10.1007/978-3-030-62466-8_17.

[30] Franz Inc. AllegroGraph: Enterprise knowledge graph solutions. https://allegrograph.com, 2026.

[31] George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, et al. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Studies in Health Technology and Informatics*, 216:574–578, 2015.

[32] Lina Dadić. Enabling cross-institutional semantic interoperability between healthcare facilities through fairification of antenatal care emr systems. Master's thesis, Leiden University, 2026. Unpublished Master's thesis.

[33] Tesfit Gebremeskel Gebreslassie, J. Stocker, Rens Kievit, and Mirjam van Reisen. Data transformation process in the humanitarian data space. 2026. Forthcoming.

[34] VODAN Africa. 1st phase of VODAN africa project (april–september 2020). https://www.vodan-totafrica.info/vodan-africa.php?i=10&a=1st-phase-of-vodan-africa-project-april-%E2%80%93-september-2020, 2020.

[35] Celia Alvarez-Romero, Alicia Martínez-García, A. Anil Sinaci, Mert Gencturk, Eva Méndez, Tony Hernández-Pérez, et al. FAIR4Health: Findable, accessible, interoperable and reusable data to foster health research. *Open Research Europe*, 2:34, 2022. doi: 10.12688/openreseurope.14349.2.

[36] Elena Lazarova, Sara Mora, Norbert Maggi, Carmelina Ruggiero, Alessandro Cosolito Vitale, Paolo Rubartelli, and Mauro Giacomini. An interoperable electronic health record system for clinical cardiology. *Informatics*, 9(2):47, 2022. doi: 10.3390/informatics9020047.

[37] Joshua C. Mandel, David A. Kreda, Kenneth D. Mandl, Isaac S. Kohane, and Rachel B. Ramoni. SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association*, 23(5):899–908, 2016. doi: 10.1093/jamia/ocv189.

[38] Muhammad Ayaz, Muhammad F. Pasha, Mohammed Y. Alzahrani, Rahmat Budiarto, and Deris Stiawan. The fast health interoperability resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities. *JMIR Medical Informatics*, 9(7):e21929, 2021. doi: 10.2196/21929.

[39] Emeka Chukwu, Lalit Garg, Nkiruka Obande-Ogbuinya, and Vijay Kumar Chattu. Standardizing primary health care referral data sets in nigeria: Practitioners' survey, form reviews, and profiling of fast healthcare interoperability resources (FHIR). *JMIR Formative Research*, 6(7):e28510, 2022. doi: 10.2196/28510.

[40] Tesfit Gebremeskel Gebreslassie, Mirjam van Reisen, Samson Yohannes Amare, Getu Tadele Taye, and Ruduan Plug. FHIR4FAIR: Leveraging FHIR in health data FAIRfication process: In the case of VODAN-A. *FAIR Connect*, 1(1): 49–54, 2023. doi: 10.3233/FC-230504.

[41] Sandra Heiler. Semantic interoperability. *ACM Computing Surveys*, 27(2):271–273, 1995. doi: 10.1145/210376.210392.

[42] Hansi Zhang, Yi Guo, Qian Li, Thomas J. George, Elizabeth Shenkman, François Modave, and Jiang Bian. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. *BMC Medical Informatics and Decision Making*, 18(S2):129–147, 2018. doi: 10.1186/s12911-018-0636-4.

[43] SNOMED International. What is SNOMED CT? https://www.snomed.org, 2025.

[44] Samson Yohannes Amare, Liya Mamo Weldu, and Tesfit Gebremeskel. Harmonising antenatal care records across four african countries: A comparative analysis and development of a standardised data model. In Mirjam van Reisen and Munyaradzi Mawere, editors, *FAIR Data, FAIR Africa, FAIR World – Internationalisation of the Health Data Space*, pages 301–. Langaa RPCIG, 2025.

[45] Mercy Chepkirui, Stephanie Dellicour, Rosemary Musuva, Isdorah Odero, Benson Omondi, Benard Omondi, Eric Onyango, Hellen Barsosio, Lilian Otiso, Gordon Okomo, Maina Waweru, Maia Lesosky, Tara Tancred, Yussif Alhassan, Simon Kariuki, Feiko terKuile, and Miriam Taegtmeyer. Missed opportunities

for digital health data use in healthcare decision-making: A cross-sectional digital health landscape assessment in homa bay county, kenya. *PLOS Digital Health*, 2025. doi: 10.1371/journal.pdig.0000870. Published 13 June 2025.

[46] Philip Mwago. Tracking patients across points of care: Interoperability between OpenMRS and the community health toolkit. Medic, https://medic.org/stories/tracking-patients-across-points-of-care-interoperability-between-openmrs-and-t 2024. 21 June 2024.

[47] Government of Kenya, Ministry of Health. Kenya health information systems interoperability framework. https://www.data4sdgs.org/sites/default/files/services_files/Kenya%20Health%20Information%20Systems%20Interoperability%20Framework.pdf.

[48] Geletaw Sahle Tegenaw, Demisew Amenu, Girum Ketema, Frank Verbeke, Jan Cornelis, Bart Jansen, Natasha McDonald, and Jackline Oluoch-Aridi. Analysis of low resource setting referral pathways to improve coordination and evidence-based services for maternal and child health in ethiopia. *PLOS ONE*, 17(8): e0273436, 2022. doi: 10.1371/journal.pone.0273436.

[49] Hamufare Dumisani Mugauri, Memory Chimsimbe, Gerald Shambira, Shepherd Shamhu, Juliet Nyamasve, Manes Munyanyi, Robert Gongora, Simukai Zizhou, Meggie Gabida, and Richard Makurumidze. A decade of designing and implementing electronic health records in sub-Saharan africa: A scoping review. *Global Health Action*, 18(1):2492913, 2025. doi: 10.1080/16549716.2025.2492913.

[50] Samson Yohannes Amare, Liya Mamo Weldu, Araya Abrha Medhanyie, and Mirjam van Reisen. Testing the cross-border africa health data space: Monitoring, tracking and prevention of mother to child syphilis transmission. In Mirjam van Reisen and Munyaradzi Mawere, editors, *FAIR Data, FAIR Africa, FAIR World – Internationalisation of the Health Data Space*, pages 517–. Langaa RPCIG, 2025.

[51] Ministry of Health, Kenya. Kenya community health policy 2020–2030. https://chu4uhc.org/wp-content/uploads/2023/08/Kenya-Community-Health-Policy-Signed-1.pdf, 2020.

[52] Republic of Kenya. The primary health care act, 2023. Kenya Gazette Supplement No. 190 (Acts No. 13); https://new.kenyalaw.org/akn/ke/act/2023/13/eng@2023-11-24, 2023.

[53] Kamukunji Community Empowerment Initiative (KCEI). Home / about us. https://kceicbo.org/, 2026.

[54] Nairobi City County. Pumwani maternity hospital. https://nairobi.go.ke/pumwani-maternity-hospital, 2026.

[55] N. Saman Wijesooriya, Roger W. Rochat, Mary L. Kamb, Prasad Turlapati, Marleen Temmerman, Nathalie Broutet, and Lori M. Newman. Global burden of maternal and congenital syphilis in 2008 and 2012: A health systems modelling study. *The Lancet Global Health*, 4(8):e525–e533, 2016. doi: 10.1016/S2214-109X(16)30135-8.

[56] Jeremiah Laktabai, Victoria L. Mobley, Wendy Prudhomme-O'Meara, and Steve M. Taylor. Associations between antenatal syphilis test results and adverse pregnancy outcomes in western kenya. *The American Journal of Tropical Medicine and Hygiene*, 107(2):401–406, 2022. doi: 10.4269/ajtmh.22-0083.

[57] Dorcas Sarone Omondi Aduda and Nomvuyo Mkhize. Ethical issues evolving from patients' perspectives on compulsory screening for syphilis and voluntary screening for cervical cancer in kenya. *BMC Medical Ethics*, 15:27, 2014. doi: 10.1186/1472-6939-15-27.

[58] Completeness of malaria indicator data reporting via the district health information software 2 in kenya, 2011–2015. *Malaria Journal*, 16:308, 2017. doi: 10.1186/s12936-017-1973-y.

[59] Open Health News. District health information system 2 (DHIS2). https://www.openhealthnews.com/resources/district-health-information-system-2-dhis2, 2026. Reports Kenya as first Sub-Saharan African country with a fully online DHIS2 deployment, completed September 2011.

[60] OpenMRS. KenyaEMR distribution. https://openmrs.atlassian.net/wiki/spaces/docs/pages/25475417/KenyaEMR+Distribution, 2026.

[61] The Star (Kenya). Health ministry rolls out an electronic community information system. https://www.the-star.co.ke/news/realtime/2023-10-17-health-ministry-rolls-out-an-electronic-community-information-syste 2023.

Appendix A

Field Visit to Kenya

As part of this research, a field visit was conducted to Nairobi, Kenya, in January 2026 under the supervision of Prof. Dr. Mirjam van Reisen. The visit provided an opportunity to engage directly with healthcare stakeholders and gain a practical understanding of maternal healthcare workflows and interoperability challenges.

A.1 Sensitive Data Sovereignty Conference

During the visit, we attended the Sensitive Data Sovereignty Conference, held at Tangaza University, Nairobi, on 27 January 2026. The conference brought together researchers, policymakers, healthcare professionals, humanitarian organisations, and technology experts to discuss topics including data sovereignty, FAIR data principles, ethical Artificial Intelligence, health data governance, and the development of African Health and Humanitarian Data Spaces.

The conference provided valuable background for this research by highlighting the importance of FAIR-compliant infrastructures and semantic technologies for responsible healthcare data sharing.



Figure A.1: At the Sensitive Data Sovereignty Conference, Tangaza University, Nairobi, Kenya (January 2026) with Lina and Prof. Dr. Mirjam van Reisen.

A.2 Visit to Pumwani Maternity Hospital and KCEI

Following the conference, we visited Pumwani Maternity Hospital and met with representatives from Kamukunji Community Empowerment Initiative (KCEI), including Community Health Workers, as well as staff from Pumwani Maternity Hospital.

These meetings provided first-hand insight into maternal healthcare workflows, data collection practices, referral processes, antenatal care, and the interoperability challenges between Community Health Workers and referral hospitals. These discussions informed the requirements and design of the semantic interoperability framework presented in this thesis.

Special thanks are extended to Whitney from KCEI and William Nandwa from Pumwani Maternity Hospital for their valuable contributions.



Figure A.2: Group photograph with representatives from KCEI and Pumwani Maternity Hospital during the visit to Pumwani Maternity Hospital, Nairobi, Kenya (January 2026).

Appendix B

Stakeholder Information

The proposed semantic interoperability framework was informed through discussions with stakeholders from both the community and hospital healthcare settings during the field visit to Kenya.

Permission was obtained from the stakeholders to reference their contribution within this thesis.

Name	**Organisation**	**Role**	**Contribution**
Whitney	Kamukunji Community Empowerment Initiative (KCEI)	Community Health Worker	Explained community maternal healthcare workflows, referral procedures and Community Health Unit data collection processes.
Contact Information	———		
William Nandwa	Pumwani Maternity Hospital	Hospital Representative	Explained hospital maternal healthcare workflows, antenatal care processes, referral management and hospital data practices.
Contact Information	`wmnandwa@gmail.com`		

Table B.1: Stakeholders consulted during the field visit.

Appendix C

Ontology Development and Knowledge Graph

This appendix contains selected screenshots illustrating the ontology development and implementation of the proposed semantic interoperability framework.

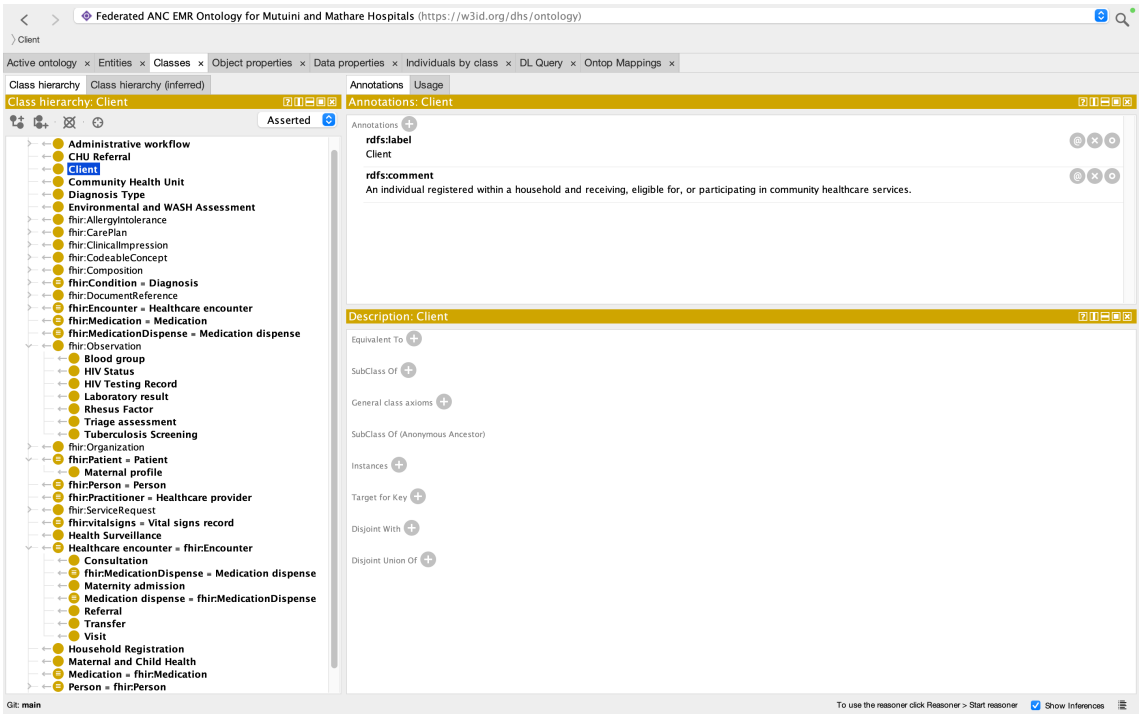


Figure C.1: Ontology hierarchy in Protégé.(Classes)

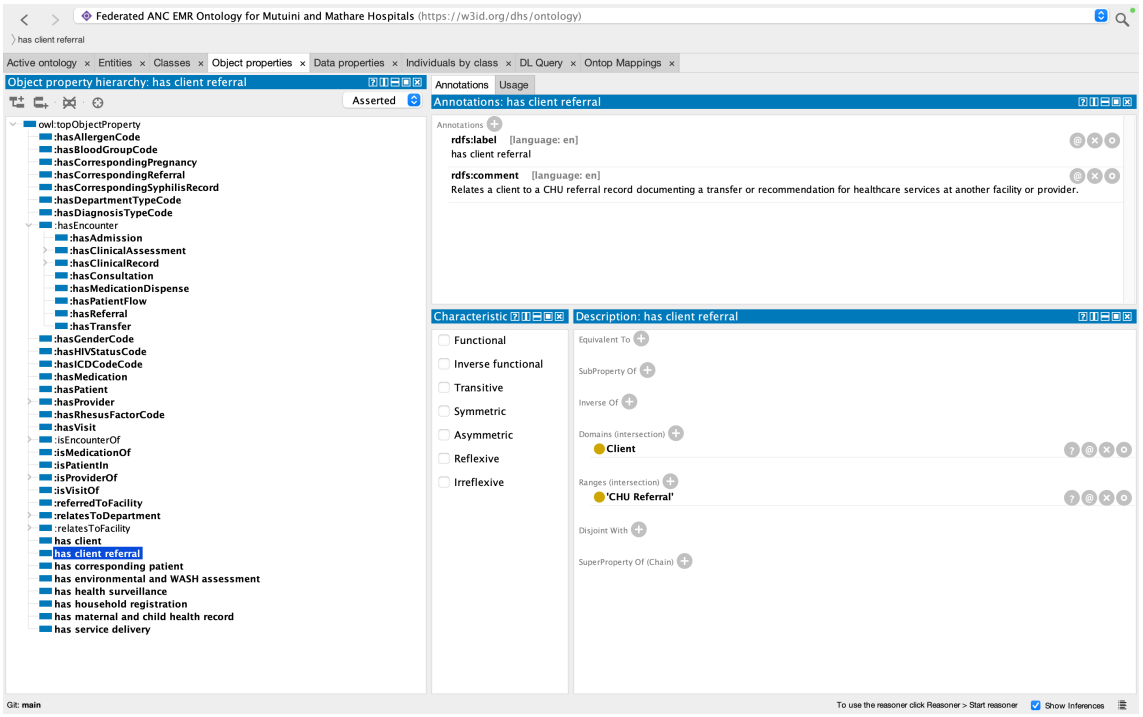


Figure C.2: Ontology hierarchy in Protégé (Object Properties)

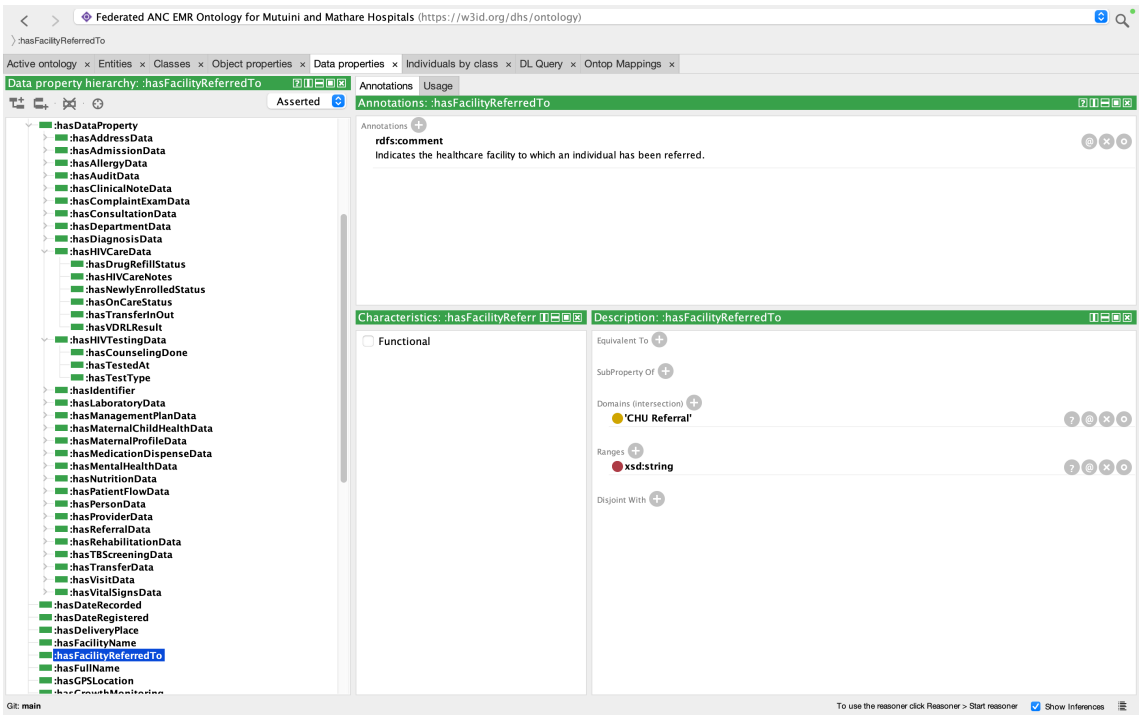


Figure C.3: Ontology hierarchy in Protégé (Data Properties)

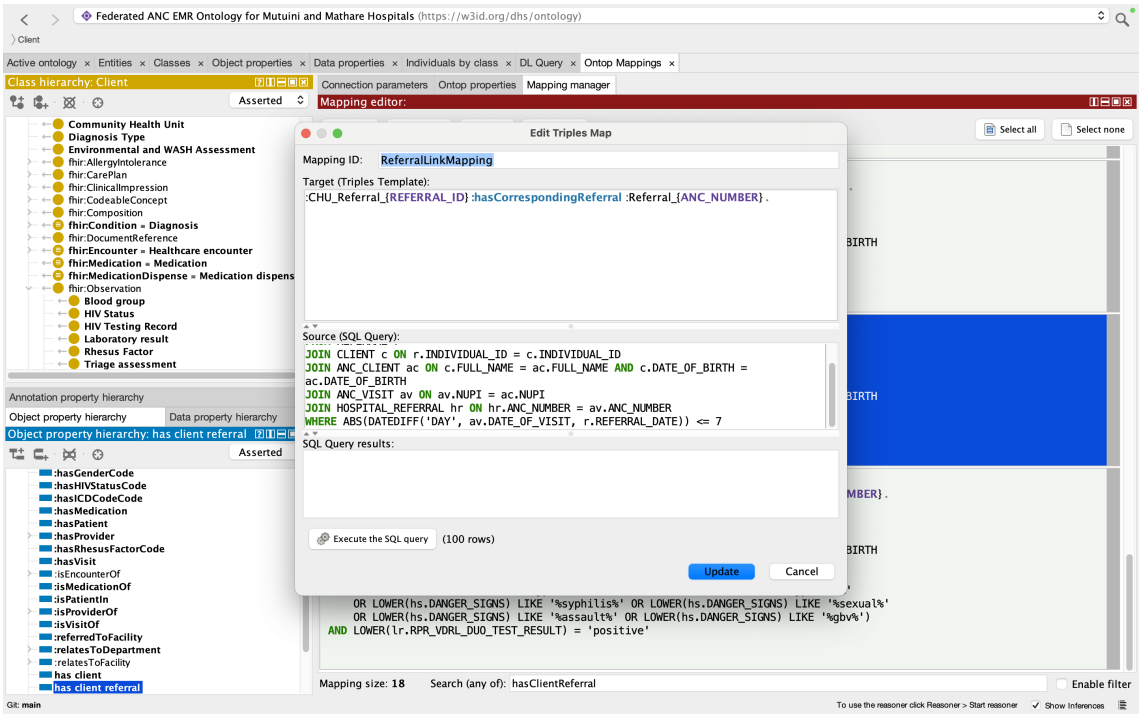


Figure C.4: Example Ontop mappings.

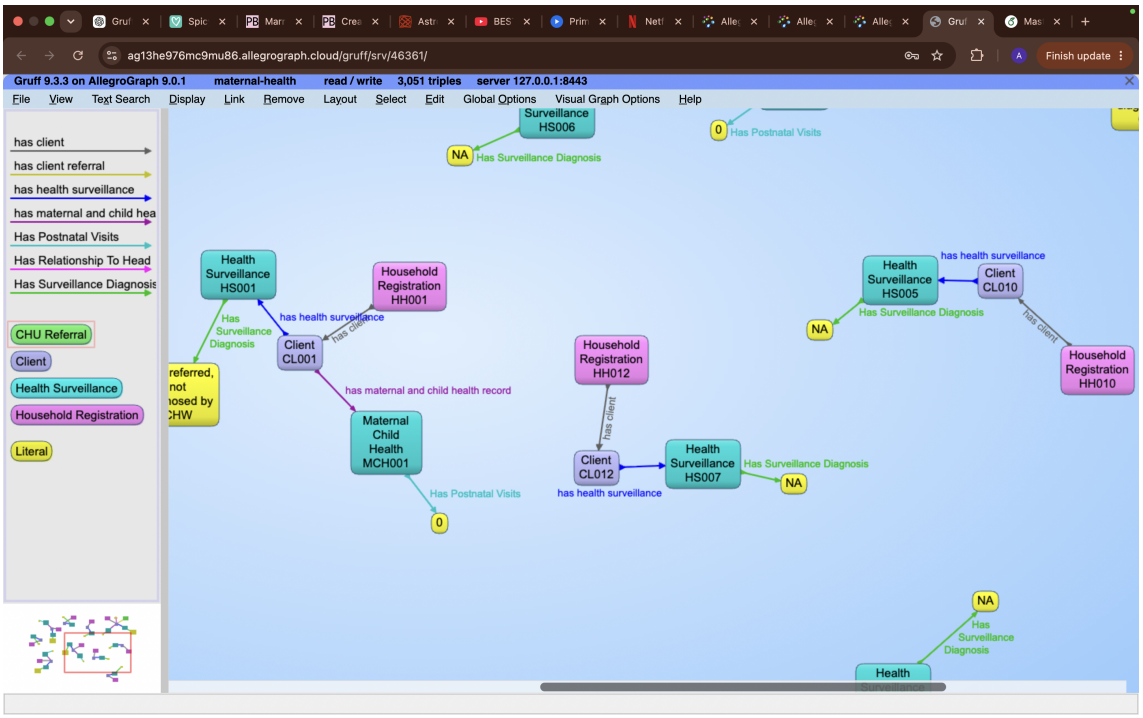


Figure C.5: Knowledge graph visualisation in AllegroGraph.

Appendix D

Implementation Repository

The complete implementation developed for this research is publicly available through GitHub.

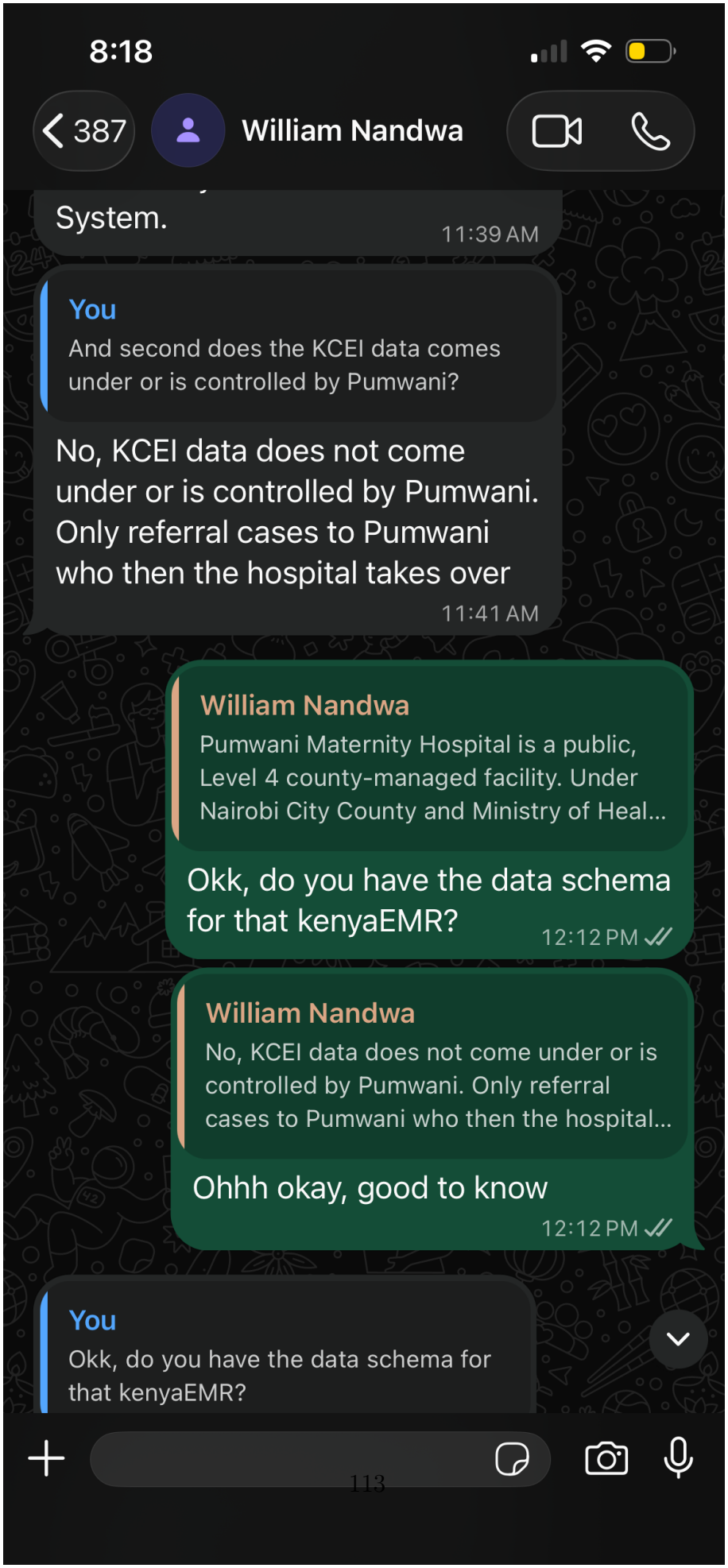
Repository Name
ANC-healthcare-data-FAIRifaction-

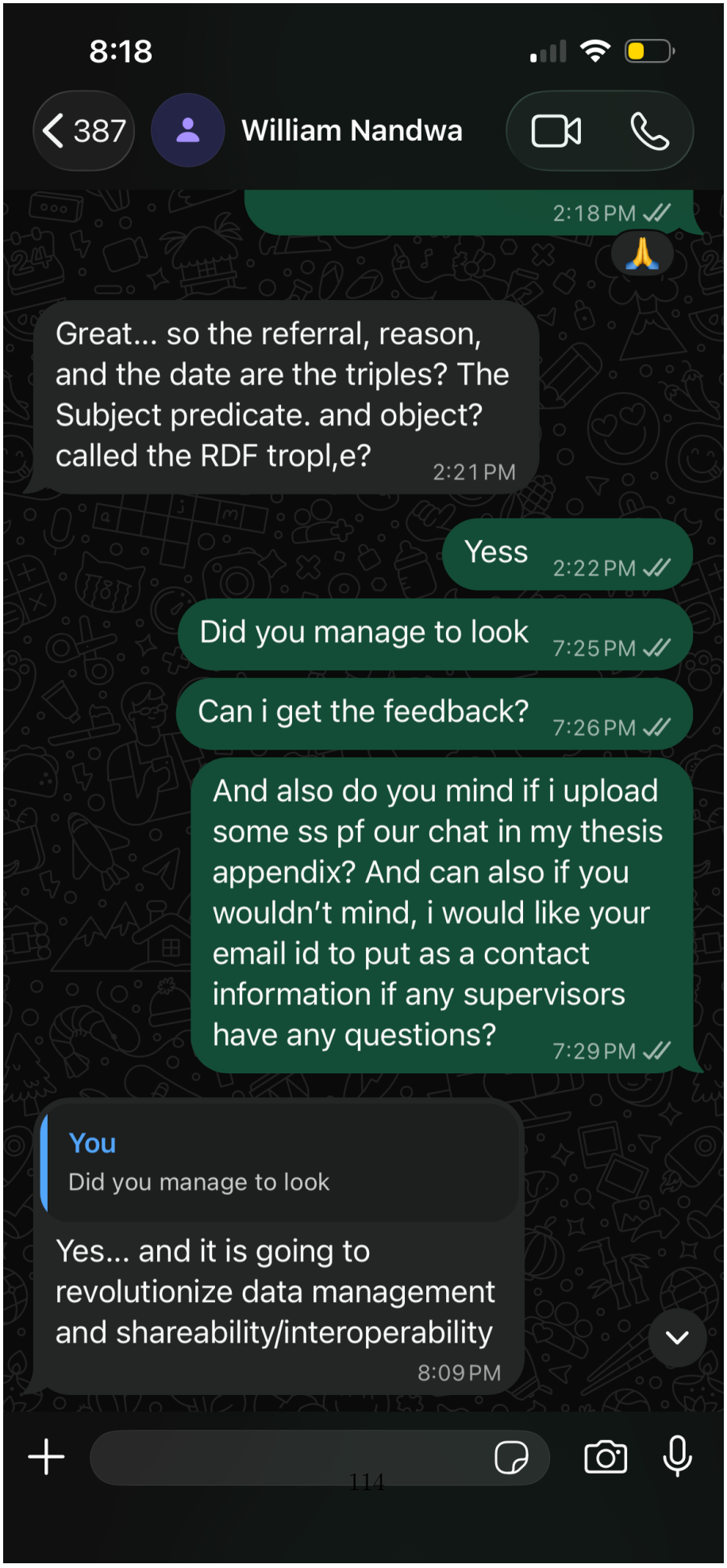
Repository Link
https://github.com/aryastark1614/ANC-healthcare-data-FAIRifaction-

Appendix E

Research Ethics and Consent

This appendix contains supporting documentation relating to stakeholder engagement and ethical considerations undertaken during this research.


8:18
387
William Nandwa
System.
11:39 AM
You
And second does the KCEI data comes under or is controlled by Pumwani?
No, KCEI data does not come under or is controlled by Pumwani. Only referral cases to Pumwani who then the hospital takes over
11:41 AM
William Nandwa
Pumwani Maternity Hospital is a public, Level 4 county-managed facility. Under Nairobi City County and Ministry of Heal...
Okk, do you have the data schema for that kenyaEMR?
12:12 PM
William Nandwa
No, KCEI data does not come under or is controlled by Pumwani. Only referral cases to Pumwani who then the hospital...
Ohhh okay, good to know
12:12 PM
You
Okk, do you have the data schema for that kenyaEMR?


8:18
387
William Nandwa
2:18 PM
Great... so the referral, reason, and the date are the triples? The Subject predicate. and object? called the RDF tropl,e?
2:21 PM
Yess
2:22 PM
Did you manage to look
7:25 PM
Can i get the feedback?
7:26 PM
And also do you mind if i upload some ss pf our chat in my thesis appendix? And can also if you wouldn't mind, i would like your email id to put as a contact information if any supervisors have any questions?
7:29 PM
You
Did you manage to look
Yes... and it is going to revolutionize data management and shareability/interoperability
8:09 PM

Appendix F

AllegroGraph Repository Access

To facilitate verification of the implementation presented in this thesis, access to the AllegroGraph repository used during the development and evaluation of the proposed semantic interoperability framework is provided below. The repository contains the RDF triples generated from the ontology-based mappings, the materialised knowledge graph, and the SPARQL endpoint used throughout the evaluation.

Server URL `https://pumwani.tail2fb347.ts.net:10000`

Repository Name `maternal-health`

Username `Stakeholders`

Password `Stakeholder1234`

Repository Description The repository contains the RDF knowledge graph generated from the Community Health Unit and hospital mock datasets after ontology-based mapping and RDF materialisation. Examiners may use the provided credentials to inspect the stored triples, graph visualisation, and execute SPARQL queries presented in this thesis.

Note The repository is provided solely for the evaluation of this Master's thesis. It contains only synthetic (mock) data and does not include any real patient information. Although the credentials above are valid and reflect the authentication mechanism described in Section 8.8.3, the repository has been deliberately left open for the duration of the examination period so that examiners can inspect it without needing to enter them.

F.1 FAIR Data Point Registration

The metadata for both the Community Health Unit and hospital repositories is registered in a FAIR Data Point, making the datasets findable and discoverable

independently of direct AllegroGraph access.

FAIR Data Point https://pumwani.tail2fb347.ts.net

F.2 Africa Health Data Space

Both repositories are additionally exposed through the Africa Health Data Space (AHDS), the federated ecosystem described in Section 3.8.6.

AHDS Listing https://africahealthdataspace.com/datasets/

Dataset Names `CHU Maternal Healthcare Dataset and Hospital Maternal Healthcare Dataset`

Username `Stakeholders`

Password `Stakeholder1234`

F.3 Demonstration Video

A recorded demonstration walks through the localisation of the data in the two locally deployed repositories, the FAIR Data Point, and the AHDS, and explains how these components together enable interoperability and discoverability without revealing the underlying datasets (Section 9.4).

Video Link https://drive.google.com/file/d/1KiswrZYcaQWHkLlXE5V1LZ5qEpQb2bZB/view?usp=sharing