

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



LSHTM Research Online

Kanjala, C; (2020) Provenance of "after the fact" harmonised community-based demographic and HIV surveillance data from ALPHA cohorts. PhD thesis, London School of Hygiene & Tropical Medicine. DOI: <https://doi.org/10.17037/PUBS.04655994>

Downloaded from: <https://researchonline.lshtm.ac.uk/id/eprint/4655994/>

DOI: <https://doi.org/10.17037/PUBS.04655994>

Usage Guidelines:

Please refer to usage guidelines at <https://researchonline.lshtm.ac.uk/policies.html> or alternatively contact researchonline@lshtm.ac.uk.

Available under license. To note, 3rd party material is not necessarily covered under this license: <http://creativecommons.org/licenses/by-nc-nd/3.0/>

<https://researchonline.lshtm.ac.uk>

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



**Provenance of “after the fact” harmonised community-based
demographic and HIV surveillance data from ALPHA cohorts**

Chifundo Kanjala

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of London

September 2019

Department of Population Health

Faculty of Epidemiology and Population Health

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Research group affiliation: ALPHA network and Population Studies Group

Supervisors

Professor Basia Zaba
Medical Demography,
Department of Population Health,
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine

Professor Jim Todd
Applied Biostatistics
Department of Population Health,
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine

Dr Emma Slaymaker
Associate Professor
Department of Population Health,
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine

Advisory Committee members:

Dr Jay Greenfield
Health Informatics consultant and DDI Developer
United State of America

Dr Tito Castillo
Application Architect
University College London Hospitals NHS Foundation Trust

Mr Gareth Knight
Project Manager (Research Data Management)
Library & Archives Services
London School of Hygiene and Tropical Medicine

Dr David Beckles
Independent Computer Software Professional
United Kingdom

Professor Mia Crampin
Clinical Epidemiology
Department of Population Health,
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine

DECLARATION OF OWN WORK

All students are required to complete the following declaration when submitting their thesis.

Please note: Assessment misconduct includes any activity that compromises the integrity of your research or assessment of it will be considered under the Assessment Irregularity Policy. This includes plagiarism, cheating and failure to follow correct progression and examination procedures.

Please see the following documents for further guidance:

- [Research Degrees Handbook](#)
- [Assessment Irregularities Policy](#)

Supervisors should be consulted if there are any doubts about what is permissible.

1. STUDENT DETAILS

Student ID Number	364327	Title	Mr
First Name(s)	Chifundo		
Surname/Family Name	Kanjala		
Programme of Study	PhD		
LSHTM Email (if this is no longer active, please provide an alternative)	chifundo.kanjala@lshtm.ac.uk		

2. TITLE OF THESIS


Title of Thesis	Provenance of "after the fact" harmonised community-based demographic and HIV surveillance data from ALPHA cohorts
------------------------	--------------------------------------------------------------------------------------------------------------------

3. DECLARATION

I have read and understood the LSHTM's definition of plagiarism and cheating. I declare that this thesis is my own work, and that I have acknowledged all results and quotations from the published or unpublished work of other people.

I have read and understood the LSHTM's definition and policy on the use of third parties (either paid or unpaid) who have contributed to the preparation of this thesis by providing copy editing and, or, proof reading services. I declare that no changes to the intellectual content or substance of this thesis were made as a result of this advice, and, that I have fully acknowledged all such contributions.

I have exercised reasonable care to ensure that the work is original and does not to the best of my knowledge break any UK law or infringe any third party's copyright or other intellectual property right.

Student Signature	
Date	30/08/2019

ABSTRACT

Background: Data about data, metadata, for describing Health and Demographic Surveillance System (HDSS) data have often received insufficient attention. This thesis studied how to develop provenance metadata within the context of HDSS data harmonisation - the network for Analysing Longitudinal Population-based HIV/ AIDS data on Africa (ALPHA). Technologies from the data documentation community were customised, among them: A process model - Generic Longitudinal Business Process Model (GLBPM), two metadata standards - Data Documentation Initiative (DDI) and Standard for Data and Metadata eXchange (SDMX) and a data transformations description language - Structured Data Transform Language (SDTL).

Methods: A framework with three complementary facets was used:

- Creating a recipe for annotating primary HDSS data using the GLBPM and DDI,
- Approaches for documenting data transformations. At a business level, prospective and retrospective documentation using GLBPM and DDI and retrospectively recovering the more granular details using SDMX and SDTL.
- Requirements analysis for a user-friendly provenance metadata browser.

Results: A recipe for the annotation of HDSS data was created outlining considerations to guide HDSS on metadata entry, staff training and software costs. Regarding data transformations, at a business level, a specialised process model for the HDSS domain was created. It has algorithm steps for each data transformation sub-process and data inputs and outputs. At a lower level, the SDMX and SDTL captured about 80% (17/21) of the variable level transformations. The requirements elicitation study yielded requirements for a provenance metadata browser to guide developers.

Conclusions: This is a first attempt ever at creating detailed metadata for this resource or any other similar resources in this field. HDSS can implement these recipes to document their data. This will increase transparency and facilitate reuse thus potentially bringing down costs of data management. It will arguably promote the longevity and wide and accurate use of these data.

ACKNOWLEDGEMENTS

Heart felt gratitude goes to my supervisors – Basia Zaba (1949–2018), Jim Todd and Emma Slaymaker. Though she is not here to see my completed work, I am so thankful to Basia. She foresaw the importance of the FAIR guiding principles for scientific data management well before they were popular in population health. She was fully committed to seeing me through my studies up to the time of her passing. I am also very thankful to Jim and Emma who took over to guide me to completion.

I was privileged to have a brilliant advisory team - Jay Greenfield, Tito Castillo, David Beckles, Gareth Knight and Mia Crampin. Jay selflessly volunteered to guide me through a crucial part of my thesis, and we had endless Skype calls with him training me to think like a health informatician. To Tito, thanks for the searching questions and pointers on next steps at various points during the thesis journey. Many thanks to all for the guidance and for reading through several versions of my work, giving me the much needed feedback.

Thanks to the HDSS leaders (Mia Crampin and Mark Urassa) and colleagues at the MEIRU and TAZAMA projects for being fantastic hosts. You provided a conducive environment for me to do both my work and studies. The Nairobi and Kisesa HDSS leaders for allowing me to use their data transformation routines and structured metadata in the project.

Thanks to all who participated in my Skype survey for your time and insights.

Thanks to Arofan Gregory, Joel Francis, Cho Kabudula, Benjamin Clark, Baltazar Mtenga, Clara Calvert, Rachel Scott, Keith Branson, Estelle McLean, Francesca Cavallaro, Susie Schaffnit and Jackie Saul for the ideas, encouragement and fun times during my studies. To the ALPHA network colleagues, thank you for the friendly and engaging atmosphere during workshops and my visits. To my UK-based colleagues Milly Marston, Paul Mee, Georges Reniers, David Beckles, Frankie Liew, Christina Albertsen, Christina Breach and all in the department of Population Health for the warm welcome each time I visited London.

To my family, I say thank you. My wife Belinda aka Sunshine and son Ngaavongwe (NDK) – my home supervisors who believed in me, permitted me to work after hours and endured periods of my absence – you are superstars!! My parents and siblings and your spouses for supporting me in laying an academic foundation upon which I built this PhD. The Katsoka family, for being lovely hosts during my South Africa visits. Godlove, Matilda, Pardon, Carol, the Jaisi and Odilo and Mlambo families and my spiritual family - FIF church Lilongwe, thanks for the fellowship and encouragement. To others, too numerous to mention by name, thank you for playing a part helping me to realise this PhD.

Above all, I am grateful to God, the Almighty, for giving me the ability to work through my thesis. In Him I live, I move and have my being ... (Acts 17:28).

DEDICATION

I would like to dedicate this thesis to my elder brother, Samson Kanjala and sister-in-law Erifa. Samson and Erifa have been fantastic life coaches throughout my life. I thank them for teaching me the value of hard work and for encouraging me to be persistent in the face of adversities. I wish I had a thousand arms to hug them with...

TABLE OF CONTENTS

ABSTRACT	4
LIST OF FIGURES	11
LIST OF TABLES	13
DEFINITIONS AND ABBREVIATIONS	14
1. INTRODUCTION	19
1.1 Background	19
1.2 Research Problem	20
1.3 Study Contributions	21
1.4 Related work	22
1.4.1 Use of metadata standards in public health research and epidemiology	22
1.4.2 Data pooling, harmonisation and sharing: benefits versus documentation demands	22
1.4.3 ALPHA: Brief overview and Use of CiB technology	23
1.4.4 Structured documentation of data transformations	23
1.4.5 Generic process models	23
1.5 Research questions	24
1.6 Aim	24
1.7 Objectives	24
1.8 Delimitations of scope and key assumptions	24
1.9 Thesis outline	25
2. LITERATURE REVIEW	28
2.1 Introduction	28
2.1.1 Scholarly databases, keywords and literature search strategy	28
2.1.2 General picture of data documentation and sharing in public health research	30
2.2 Standardisation, Metadata, and metadata standards	31
2.2.1 Standardisation in Public Health research and epidemiology	31
2.2.2 Metadata definition, types and uses	32
2.2.3 Metadata and information technologies	33
2.3 Metadata standards and process models applicable to demographic and epidemiological surveillance data	36
2.3.1 Dublin core	36
2.3.2 Data Documentation Initiative (DDI)	36
2.3.3 Statistical Data and Metadata eXchange (SDMX)	41
2.3.4 Generic Process models	44
2.3.5 Generic Statistical Information Model (GSIM)	48
2.3.6 Metadata support for describing data transformations	50
2.4 Software tools for implementing metadata standards	53
2.4.1 Generic DDI Codebook tools	53
2.4.2 DDI Lifecycle Tools	55
2.5 Data documentation practices among data harmonisation projects	57

2.6	Metadata standards implementation in HDSS studies	58
2.7	Discussion.....	58
3.	STUDY SETTINGS – ALPHA Network.....	60
3.1	Introduction	60
3.2	Health and Demographic Surveillance System	60
3.2.1	Rationale for establishing HDSS in LMIC - defective civil registration and dearth of reliable vital statistics	60
3.2.2	HDSS core concepts.....	61
3.2.3	HDSS reference data model	62
3.3	ALPHA network overview	64
3.4	ALPHA Data Production environment and processes	67
3.4.1	Centre in a Box major components.....	70
3.4.2	Pentaho ETL processes in ALPHA	70
3.4.3	CiB Metadata management overview	73
3.5	Chapter summary	76
4.	OPEN-ACCESS FOR EXISTING LMIC DEMOGRAPHIC SURVEILLANCE DATA USING DDI	78
4.1	Abstract.....	78
4.2	Keywords.....	78
4.3	Introduction	78
4.4	Study settings and methods.....	80
4.4.1	Study settings	80
4.4.2	Study methods	81
4.5	Results	83
4.5.1	Mapping Kisesa study HDSS data production to GLBPM	83
4.5.2	Implementation of DDI in Nesstar Publisher and Colectica Designer	84
4.6	Discussion.....	88
4.7	Summary	91
5.	HIGH LEVEL, STRUCTURED METADATA FOR ALPHA DATA TRANSFORMATIONS:.....	92
5.1	Introduction	92
5.1.1	Aims and objectives	93
5.1.2	Related work	94
5.1.3	Chapter overview	95
5.2	Connecting Chapter 5 to Chapters 3 and 4	95
5.3	Methods	98
5.3.1	Determining metadata content and structure for ALPHA ETLs	98
5.3.2	ALPHA ETL through GLBPM lenses: Mapping and specialising GLBPM	99
5.3.3	Metadata infusion file - template for domain metadata capturing.....	101
5.4	Results	102
5.4.1	Results of literature review on African demographic and epidemiological surveillance systems.....	102

5.4.2	Mapping ALPHA ETL sub-jobs to GLBPM	102
5.4.3	Specialising the mapped GLBPM steps	104
5.4.4	Input (pre-condition) and output (post-condition) data records	106
5.4.5	Metadata infusion file	107
5.5	Discussion.....	109
6.	LOWER LEVEL, STRUCTURED METADATA FOR ALPHA DATA TRANSFORMATIONS:	112
6.1	Introduction	112
6.1.1	Aim.....	114
6.1.2	The bigger picture	114
6.1.3	Chapter overview	116
6.2	Methods	117
6.2.1	Data quality assessment sub-job analysis.....	117
6.2.2	Pentaho to SDDL Mapping	119
6.2.3	Structured documentation for indicators.....	119
6.3	Results	120
6.3.1	Mapping Pentaho steps to SDDL	120
6.3.2	Development of SDDL code from the mapping.....	121
6.3.3	Structured documentation of data quality indicators	122
6.4	Discussion.....	124
7.	PROVIDING END USERS WITH ACCESS TO ALPHA PROVENANCE METADATA 126	
7.1	Introduction	126
7.1.1	Objective	126
7.2	Methods	126
7.2.1	Development of mock-up diagrams for use in elicitation study	126
7.2.2	Mock-up diagrams of the proposed features: The details	127
7.2.3	Recruitment of study participants.....	131
7.2.4	Data collection.....	131
7.2.5	Data analysis.....	132
7.3	Results	132
7.3.1	Organisational diversity, education, work experience and roles of interviewees 132	
7.3.2	Scores for proposed features	134
7.3.3	Rationale for or against having proposed features and suggested improvements.....	134
7.3.4	Overarching aspects.....	137
7.3.5	Feedback requiring structural changes to metadata	139
7.4	Discussion.....	140
7.5	Data availability statement.....	142
8.	CONCLUSION	143
8.1	Introduction	143
8.2	Documentation of primary data from the network partners	144
8.2.1	Objective	144
8.2.2	Summary of findings.....	144

8.2.3	Research contributions	145
8.2.4	Recommendations.....	145
8.2.5	Future work.....	145
8.3	Documentation of data harmonisation processes	146
8.3.1	Objective	146
8.3.2	Summary of findings.....	146
8.3.3	Research contributions.....	146
8.3.4	Recommendations.....	147
8.3.5	Future work.....	148
8.4	Provenance metadata browser software requirements	148
8.4.1	Objective	149
8.4.2	Summary of findings.....	149
8.4.3	Research contributions.....	149
8.4.4	Recommendations.....	149
8.4.5	Future work.....	150
8.5	Study limitations	150
9.	REFERENCES	152
APPENDIX A	METADATA INFUSION FILE SCHEMA	164
APPENDIX B	INFUSION FILE	171
APPENDIX C	Background information.....	200
1.	What is ALPHA?	200
2.	ALPHA data and “modus operandi”	200
3.	Data harmonisation in ALPHA	201
4.	Mock-ups.....	201
5.	Terms used in mock-ups	202
6.	Why ALPHA network interviewees?.....	202
7.	Why CLOSER project interviewees?	202
8.	Requirements overview.....	202
9.	Types of Requirements.....	202
10.	References.....	203
APPENDIX D	Information sheet.....	204
APPENDIX E	Consent form.....	206
APPENDIX F	Questionnaire guide.....	207
	<i>Question 2.</i>	210
	<i>Question 5.</i>	215
	<i>Question 6.</i>	216

LIST OF FIGURES

Figure 1: Broad literature review topics and search strategies	29
Figure 2: Literature search results	30
Figure 3: Data lifecycle model spanning from study design to data analysis.....	38
Figure 4: Examples of dimensions, observation values and attributes for data on mortality rates in Masaka HDSS in Rural Uganda before and after Antiretroviral drugs roll out.	42
Figure 5: DDI and SDMX complementary nature	43
Figure 6: Generic Statistical Business Process Model.....	44
Figure 7: Generic Longitudinal Business Process Model.....	47
Figure 8: GLBPM "Tornado" view showing two rounds of data collection.....	48
Figure 9: Generic Statistical Information Model information objects	49
Figure 10: Relationship between GSBPM and GSIM	49
Figure 11: Metadata loss during processing performed after data collection.....	52
Figure 12: The dearth of vital statistics in LMIC especially sub-Saharan Africa	61
Figure 13: Primary entities: Residential unit, individual and social group	61
Figure 14: HDSS dynamic cohort representation	62
Figure 15: Demographic surveillance reference data model.....	63
Figure 16: Locations of ALPHA network member sites in eastern and southern Africa	65
Figure 17: ALPHA data management overview.....	68
Figure 18: ALPHA Specification 6.1 ETL process in Pentaho.....	72
Figure 19: Structured metadata catered for in the current CiB	73
Figure 20: Various stages of importing a file into Nesstar Publisher	75
Figure 21: Variable level metadata in Nesstar Publisher	75
Figure 22: Steps in the Kisesa study HDSS data collection round	81
Figure 23: Kisesa study HDSS Documentation in Nesstar Publisher.....	83
Figure 24: Exemplar Pentaho data transformation	92
Figure 25: Input objects, processes and output objects in the ALPHA ETL and their structured documentation: High level.....	96
Figure 26: Mapping of ALPHA ETL to the GLBPM.....	104
Figure 27 : Input and output data records for sub-job 002 CORE Data Quality Metrics .	106
Figure 28: Snippet of an algorithm overview within the infusion metadata file	108
Figure 29: Input objects, processes and output objects in the ALPHA ETL and their structured documentation	115
Figure 30: ALPHA Specification 6.1 ETL process in Pentaho.....	116

Figure 31: Transformations in the sub-job 002 CORE Data Quality Metrics.....	118
Figure 32: Transformation CORE Illegal start events.....	118
Figure 33: Exemplar SDTL for the recode step translated to its namesake in SDTL.....	121
Figure 34: Reshape wide and save commands in SDTL	122
Figure 35: Proposed DDI 4 augmented Process model.....	127
Figure 36: Data pipeline, its constituent tasks and their details.....	128
Figure 37: Metadata elements for describing steps in a task.....	129
Figure 38: Provision for definition of concepts related to a step in a task.....	129
Figure 39: Task-centric view showing a task and its input and output data.....	130
Figure 40: Dataset centric view – showing a dataset and tasks creating and using it.....	130
Figure 41: Dataset-centric view: Dataset structure (variable name and type)	131
Figure 42: A priori Coding Scheme based on DDI 4 Process model and proposed features	133
Figure 43: Scores given by interviewees for features displayed in the mock-ups	135
Figure 44: Metadata browser potential user groups	137

LIST OF TABLES

Table 1: Metadata schemes and their fields of application	35
Table 2: Correspondence between Codebook sections and Lifecycle modules	39
Table 3: Selected Characteristics of ALPHA network member research centres.....	66
Table 4: ALPHA datasets.....	69
Table 5: Sub-job Pentaho names and their labels.....	73
Table 6: Counts of items involved in the documentation of Kisesa HDSS	84
Table 7: Nesstar Publisher (NP) and Colectica Designer (Colectica) documentation.....	85
Table 8: Training materials and software costs.....	87
Table 9: Description of the objects and processes involved in ALPHA ETL – High level view	97
Table 10: Numbers of job entries and transformation steps in the ALPHA 6.1 ETL.....	99
Table 11: Mapping ALPHA ETL sub-jobs to GLBPM.....	103
Table 12: Specialisation of GLBPM steps for four exemplar ALPHA ETL sub-jobs	105
Table 13: Numbers of pre and post conditions for each sub-job in specification 6.1 ETL	107
Table 14: Algorithm overview for 002 CORE Data Quality Metrics.....	117
Table 15: Pentaho steps and their SDTL equivalents.....	120
Table 16: SDMX concepts for quality metrics.....	122
Table 17: Dimensions for the quality metrics	123
Table 18: Keys and data points for the quality metrics	123
Table 19: Work experience of the interviewees	134
Table 20: Rationale for having and against proposed features and suggested improvements	136
Table 21:	209
Table 22	211
Table 23	214

DEFINITIONS AND ABBREVIATIONS

ADESBPM - African Demographic and Epidemiological Surveillance Business Process Model. A specialisation of the Generic Longitudinal Business Process Model (GLBPM) designed as a reference framework defining and describing the activities and information objects involved in the management of event data from HDSS studies.

ALPHA - Analysing Longitudinal Population-based HIV/ AIDS data on Africa
A network of autonomous, longitudinal demographic and HIV/ AIDS surveillance research studies in Africa. The network runs data analysis training workshops on demographic correlates and consequences of HIV (ALPHA 2013).

Business Process – A set of coordinated activities or steps to perform one or more functions with the goal of delivering a service or product to a client (Weske 2007)

Class – a blueprint for the information objects found in a system being designed (Weisfeld 2008)

Data life cycle – the entire course of existence of a dataset, from study conceptualisation to analysis and archiving and feeding back to earlier stages (DDI Alliance 2018a).

Data Model

A mapping of the contents of an information model (defined later in this list) into a form that is specific to a particular type of data store or repository (Schoenwaelder and Pras 2003)

DC – Dublin Core

A general purpose metadata standard comprising of a set of fifteen elements used for resource descriptions (DCMI 2013).

DDI – Data Documentation Initiative

DDI is an effort to create an XML-based standard for documenting individual level social and behavioural science data (DDIAlliance.org 2013). This effort is overseen the DDI Alliance. The DDI Alliance is a self-funding organization whose members vote on the development of the specification (DDIAlliance.org 2013)

DDI 4

A version of DDI currently under development that is model based. It comprises of two main parts. An information model comprising of a library of objects and functional views of the model constructed from subsets of the library. Each views supports a specific application of the specification. The information model can be implemented in various technologies including XML and RDF (both XML and RDF are defined later in this list) (DDI Alliance 2014b)

GLBPM – Generic Longitudinal Business Process Model

A reference model that defines and describes processes involved in the production of longitudinal individual level research data (I. Barkow, Block, Greenfield, Gregory, Hebing, Hoyle, and Zenk-Möltgen 2013). Derived from the GSBPM with the purpose of describing processes related to the production of human science research data. Figure 7 in Chapter 2 shows the GLBPM.

GSBPM – Generic Statistical Business Process Model

A reference model that defines and describes processes involved in the production of statistics. It comprises of four levels (UNECE Secretariat 2009). The Statistical Business Process Model itself, the nine phases of the business process, the sub-processes in each of the phases and the description of these sub-processes. The GSBPM is shown in Figure 6, Chapter 2.

HDSS – Health and Demographic Surveillance System

A community-based information system for monitoring vital events (births, deaths and migrations) and key health indicators over time (INDEPTH Network 2002)

HTML – Hyper Text Markup Language

HTML is a markup language for describing web pages (W3schools.com 2013). HTML documents are text documents with tags which are texts in brackets as follows <html> embedded in the document. HTML has a pre-defined set of tags and syntax rules that are used to define how web page content displays in a web browser.

IHSN - International Household Survey Network

An informal collaboration of international agencies aiming to improve availability, accessibility and quality of developing countries' survey data (Ihsn.org. 2013)

Information Model

A representation of concepts, relationships, constraints, rules and operations which together form a model which provides an explicit interpretation criteria for a chosen domain of discourse (Eurostat, Directorate B: Statistical Methodologies and Tools and Unit B-5: Statistical Information Technologies 2010). An information model provides formalism to the description of a domain of discourse without constraining the mapping of the model to an implementation system.

Interoperability – The capability to communicate, execute programs, or transfer data across contexts in a manner that is highly automated and resulting in minimal information loss (Duval 2001).

Machine actionability – The ability for a digital object to provide information that is relayed in a consistently structured manner to a computer agent thus facilitating autonomous exploration by the computer agent (Wilkinson et al. 2016).

Object - A person, place, concept, thing represented by a system under consideration. A class is a blueprint for these objects (Weisfeld 2008).

Ontology – A formal representation of a set of concepts (classes of objects) within a domain and their relationships (Ontotext 2019)

OWL – Web Ontology Language

A language aimed to be the standardised ontology language for the semantic web (Antoniou and Van Harmelen 2004)

RDF - Resource Description Framework

A framework developed by the World Wide Web Consortium (W3C) for encoding, reusing and exchanging structured metadata. RDF makes possible the automated semantic processing of information by imposing the needed structural constraints on the metadata (Eric Miller 1998)

SDMX – Statistical Data and Metadata eXchange

SDMX is an International Standard Organisation (ISO) metadata standard to describe statistical data and metadata to facilitate exchange, processing and sharing among statistical and other organisations (SDMX Technical Working Group 2018a). It is an initiative birthed in the Official Statistics community with particular strengths in describing aggregated statistical data.

SDTL - Structured Data Transform Language

A model for describing data transformations developed as part of the Continuous Capture of Metadata (C2Metadata) project (C2Metadata 2017)

Semantic Web

A broad range of ideas and technologies concerned with bringing meaning to the vast amount of information on the web (Ontotext 2019). It is a vision to express web content in a form that is more easily processible by computer agents and to use software agents in a way that takes advantage of this representation (Antonioni and Van Harmelen 2004)

Software Requirements Specification – A document that lays out the requirements for a software system to be built. It provides a basis for agreement between the client and developer on what the software will need to accomplish (Pressman 2010)

UML – Unified Modeling Language

A general purpose modeling language for specifying, visualising, constructing and documenting the artefacts of a software system. It is also used for business modeling and other non-software systems (Visual Paradigm n.d.)

VTL - Validation and Transformation Language

A standard language for defining a set of operators, their syntax and semantics pertaining to the validation and transformation of any kind of statistical data (SDMX 2019).

XML – eXtensible Markup Language

XML is a markup language designed to transport and store data (W3schools.com 2013). Unlike HTML, XML tags are not pre-defined, the XML document author defines the tags.

While HTML tags give information on how to display the content of the HTML document, XML tags gives an idea on the meaning of the contents of the XML document.

XML Schema – An XML Schema describes the structure of an XML document (W3schools.com 2013). They define what elements can go into an XML document.

1. INTRODUCTION

1.1 Background

A Health and Demographic Surveillance System (HDSS) is a community-based information system for monitoring vital events (births, deaths and migrations) and key health indicators over time (Sankoh and Byass 2012; INDEPTH Network 2002). Data from HDSS serve key roles in the monitoring of the vital and health status of largely undocumented populations (Sankoh and Byass 2012). HDSS are a medium-term solution to deficiencies in civil registration and population-based health data across many Low and Middle Income Countries (LMIC) (Sankoh and Byass 2012; Setel et al. 2007; Mikkelsen et al. 2015). They also work as platforms for clinical trials or more general population-based research (Di Pasquale 2018). In addition, they contribute to multi-study data harmonisation collaborations. Examples of such collaborations include the network for Analysing Longitudinal Population based HIV/ AIDS data on Africa (ALPHA) on HIV epidemiology (Reniers et al. 2016; Slaymaker et al. 2017) and the African NCD Longitudinal Data Alliance (ANDLA) (ANDLA 2019). Unlike the cross-sectional population surveys (The DHS Program 2019; UNICEF MICS 2019) or decennial population censuses performed in the LMIC, HDSS are in a unique position to capture cause and effect relationships among health determinants and outcomes. They are therefore a resource of significant population health importance.

The development and maintenance of all the required details of the HDSS populations is a complex affair (Di Pasquale 2018; INDEPTH Network 2002; Benzler, Herbst, and MacLeod 1998). This is because HDSS are open cohorts permitting the members to leave or to be added to the study over time. Thus, they present special database management challenges in monitoring the population dynamics. Long term management and sharing of these data hinges on the availability of sufficient data documentation capturing their nuances.

Metadata, often defined as data about data, refer to the information that describes data or other resources, usually on the internet, helping to locate, retrieve and manage them (NISO 2004). They act as the bridge between data and their use, without them, data are just a meaningless collection of numbers (Ryssevik 1999). Metadata are most useful when they are packaged in standards (Duval 2001; Blank and Rasmussen 2004). A metadata standard represents a common view of how metadata within a domain of interest can be described (I. Barkow 2016; Gregory, Pascal, and Ryssevik 2009). Data described in a standardised manner

are discoverable (Wellcome Trust 2014), that is, they can be found on the web. They also can be better accessed, exchanged and manipulated in highly automated ways across contexts, otherwise known as interoperability (I. Barkow 2016; Duval 2001). Over the years, the international community has developed a number of metadata standards. The Digital Curation Centre (DCC) (DCC 2019) has a comprehensive cross discipline list of the standards. In public health research and epidemiology, the standard gaining popularity is the Data Documentation Initiative (DDI) (Wellcome Trust 2014). It is a standard for documentation of individual-level data in the social and behavioural sciences (Miller and Vardigan 2005; DDI Alliance 2018d). Currently, DDI has two strands which are DDI 2 also called DDI Codebook and DDI 3 or DDI Lifecycle. A new version, DDI 4, is also under development (DDI Alliance 2019), DDI 4 seeks compatibility with earlier versions while offering enhanced flexibility via an information model (William Block et al. 2012). Another metadata standard, the Statistical Data and Metadata eXchange (SDMX) standard, originally designed for statistical data exchange (SDMX Technical Working Group 2018a), has also been proposed for the description of public health data, for instance, by the French Sentinelles network (Turbelin and Boëlle 2013).

1.2 Research Problem

There is evidence on the use of metadata standards among African HDSS for the documentation of their primary data. A number of studies have metadata standards-based data catalogues (Africa Health Research Institute 2018; African Population and Health Research Center 2015; Ifakara Health Institute 2019). However, the literature is silent on what steps, considerations and choices these HDSS have taken in implementing the standards. This silence is a cause for concern because metadata standards are, of necessity, designed for an audience much wider than the HDSS community. Their target audience makes them too generic for an HDSS to directly apply off the shelf. They need contextualisation to suit the specific HDSS. Not knowing how to implement these standards, implies not fully understanding the requirements for creating sufficient metadata for use within the HDSS and to accompany shared datasets.

The Centre in a Box (CiB) infrastructure has been developed to automate data harmonisation across different HDSS (Herbst et al. 2015). Initially used by INDEPTH for iSHARE (INDEPTH Network 2013) which brings together demographic data, it is now also used by ALPHA for health data, specifically HIV data. Although CiB represents the best data harmonisation and curation solution for HDSS to date, it has limitations from the

perspective of data documentation best practices. The CiB has two limitations relating to metadata.

1. It does not cater for documentation of input data from the HDSS being harmonised
2. It only provides tool-specific documentation of data harmonisation processes. Tool specific-metadata tend to lock data into proprietary systems (Corti and Gregory 2011). This hampers data exchange across platforms (I. Barkow 2016; Duval 2001).

Therefore, as it stands, the existing literature has not clearly established what is required to sufficiently document HDSS data and the secondary datasets derived from them. This means that we do not fully understand what it takes to provide metadata support for long term management and sharing of these data. In addition, it is also widely recognised among the producers and users of these data that there is a considerable gap between the information required to effectively use the data and what is available from the pre-existing documentation. This position puts HDSS studies at risk of creating sub-optimal metadata and thus erroneous data interpretation and use by both the producers and external investigators. Ultimately, this hampers the potential improvements of the health and survival of the populations of interest.

1.3 Study Contributions

To contribute towards addressing the foregoing problems, this thesis investigates the implementation of metadata standards within the context of HDSS data harmonisation and pooling, the ALPHA network. It proposes an end to end provision of accessible and structured metadata for “after the fact” harmonised ALPHA datasets.

The three main contributions of this work are the following:

- It proposes steps, considerations and choices to make when implementing existing versions of the DDI standard within a typical HDSS setting using the Kisesa HDSS also known as the Magu HDSS (Kishamawe et al. 2015) in north-western Tanzania as a prototype (Chapter 4).
- It explores software-agnostic and structured documentation of the ALPHA data harmonisation processes performed using the Pentaho Data Integration tools (Pentaho Corporation 2018) within the CiB environment (Chapters 5 and 6).
- It gathers requirements for a user-friendly platform for presenting provenance metadata for ALPHA datasets to guide software developers (chapter 7).

1.4 Related work

1.4.1 Use of metadata standards in public health research and epidemiology

The public health research community has been slower in implementing the metadata standards when compared to other disciplines such as economics, genetics or geography (Wellcome Trust 2014). Consequently, the lack of accessible and structured metadata has often been cited as one of the major barriers to the full utilisation of existing public health research data (Bergeron et al. 2018; Wellcome Trust 2014; Van Panhuis et al. 2014; Pisani and AbouZahr 2010). The slow pace of adopting metadata standards is also reflected in HDSS studies (Chandramohan et al. 2008).

1.4.2 Data pooling, harmonisation and sharing: benefits versus documentation demands

Pooled data give the opportunity to analyse temporal and spatial variations in epidemic patterns (Bosch-Capblanch 2011) and provides the statistical power otherwise not available using a single study (Fortier et al. 2010). On the other hand, data sharing facilitates maximisation of knowledge and potential health benefit (Walport and Brest 2011), reduces duplication of data collection efforts, enables producers to get credit for the shared data, affords wider quality assessments from the user community and increases prospects for access to data from different sources for comparative analyses (Pisani et al. 2016). Without metadata, these opportunities cannot be fully realised (Bergeron et al. 2018).

When studies are planned with harmonisation in mind, such as in the case of the Demographic and Health Surveys (DHS) (The DHS Program 2019), harmonisation and pooling is relatively straightforward. However, in cases where data are pooled after having been collected independently using different instruments, and with no pre-planned harmonisation - otherwise known as retrospective or “ex – post” harmonisation (Granda and Blasczyk 2016), the harmonisation needed prior to pooling the data will usually involve complex transformations of the primary datasets. Care has to be taken to handle this complexity in a valid manner addressing cultural, legal and scientific challenges associated with the harmonisation (Fortier et al. 2010). Besides the scientifically sound development of the harmonisation process, the pooled data also require comprehensive documentation to communicate the provenance of the data (Fortier et al. 2017).

1.4.3 ALPHA: Brief overview and Use of CiB technology

ALPHA is a data pooling, sharing and joint analysis effort among HDSS belonging to the network. It brings together ten autonomous research partners running HDSS sites in Eastern and Southern Africa with interest in HIV epidemiology. Since its inception in 2005, the network has regularly derived harmonised datasets from the partners' operational databases and has performed cross-site data analyses answering a number of important research questions listed on the network's webpage (<http://alpha.lshtm.ac.uk/publications/>).

ALPHA is now seeking to standardise the production of its datasets across the partners and to provide accompanying provenance documentation for efficient data management and sharing. Inspired by the success of the CiB technology, ALPHA is adopting and extending the scope of the CiB to cater for the network's various datasets.

The CiB provides a self-contained, secure and robust environment for creating and curating harmonised datasets. At the centre of the CiB functionality is the Pentaho data integration software (Pentaho Corporation 2018). Pentaho provides a graphical extract-transform-load (ETL) designer to simplify the creation of data transformations. It has a rich library of pre-built components to access, prepare, and blend data from various sources. Its graphical interface serves as a form of documentation of the transformations.

1.4.4 Structured documentation of data transformations

Relating to data transformations metadata, two promising approaches are being developed in the official statistics and the social science domains. These are the Validation and Transformation Language (VTL) (SDMX Technical Working Group 2018a) and the Structured Data Transform Language (SDTL) (C2Metadata 2017) respectively. These two approaches have not yet been widely applied outside the settings within which they are being developed.

1.4.5 Generic process models

Also relevant to HDSS and ALPHA data documentation are two reference, standards-based models for data production processes. The first one is called the Generic Statistical Business Process Model (GSBPM) (UNECE 2018b). The GSBPM was developed in the official statistics domain. The second model is the GLBPM - Generic Longitudinal Business Process Model (B. I. Barkow et al. 2013) developed within the DDI community. The GSBPM describes the set of business processes needed to produce official statistics. It is a standard framework and provides common terminology to help national statistical offices to streamline production processes and to share methods. It also serves as a template for

process documentation (UNECE 2018b). The GLBPM is a specialisation of the GSBPM focussing on modelling of longitudinal survey data production. Mapping the HDSS and ALPHA data production to these process models would foster standard description of the various activities involved and a common understanding of those activities.

1.5 Research questions

In considering the research problem, a number of key research questions emerge, these are:

1. What steps and considerations does a typical HDSS need to make in adopting and adapting the DDI metadata standard for documentation of its primary data?
2. How can standards for describing data and statistics production be customised and extended to create software-agnostic documentation of ALPHA data harmonisation processes?
3. What features are required by users in an ALPHA data provenance documentation browsing and searching platform?

1.6 Aim

The aim of this research is to facilitate standardised documentation of the ALPHA data provenance within the CiB environment through the adaptation and extension of metadata standards and the associated data production generic process models.

1.7 Objectives

- (1) To adopt and adapt the DDI metadata standard for the annotation of HDSS primary data using Kisesa HDSS, north western Tanzania as prototype (Chapter 4)
- (2) To develop a standards-based framework for the documentation of retrospective data harmonisation routines performed in ALPHA and similar networks (Chapters 5 and 6)
- (3) To gather and analyse the requirements for a user-friendly provenance metadata platform for ALPHA datasets (Chapter 7)

1.8 Delimitations of scope and key assumptions

The domain of interest in this thesis is the African longitudinal health and demographic surveillance. Particular focus is placed on the high HIV prevalence regions of Eastern and

Southern African HDSS involved in ALPHA network. It therefore does not consider HDSS in Africa but outside ALPHA or those in Asia or the Oceania.

This thesis does not consider forms of data other than population-based health and demographic surveillance data. Health facility data are only considered in the context of their use for creating ALPHA datasets. National health surveys or national population census data are out of the scope. The data harmonisation technology considered in this study are the Pentaho data transformations performed within the CiB environment. Standardisation is a much broader topic than metadata standardisation, there are many standards applicable to public health research. These include:

1. The Medical Subject Headings (MeSH) which categorises biomedical concepts to facilitate indexing in biomedical journals.
2. The International Classification of Diseases (ICD) which is a system for classifying morbidity entities according to an established criteria (World Health Organization 2016)
3. The Systemized Nomenclature of Medicine Clinical Terms (SNOMED CT) (McMahon 2017)
4. The Logical Observation Identifiers Names and Codes (LOINC) (LOINC 2019) is a universal standard for identifying laboratory test results and other clinical observations for use by clinical information systems (Huff et al. 1998).
5. The Health Level 7 (HL7) (HL 7 2019, 7)
6. The Health Level 7 Fast Healthcare Interoperability Resources (HL7 FHIR) (HL7 2019, 7)
7. The Clinical Data Interchange Standards Consortium (CDISC) (McMahon 2017).

However, this thesis only considers the metadata standards used in the public health arena. Further, there is no software developed or evaluated as part of this thesis. Rather, two major products are aimed at: (1) Provenance metadata entailing documentation of the input data from ALPHA partners and the data harmonisation processes done to create the ALPHA datasets and (2) requirements for a metadata browsing software that provides user-friendly access to the developed provenance metadata. These synthesised requirements will be used by developers to create a requirements specification document which will then guide their work.

1.9 Thesis outline

Here is a quick scheme through of the thesis summarising the contents of each of the chapters.

Chapter 2: LITERATURE REVIEW

Reviews the literature on metadata and metadata standards, generic process models, data documentation practices in harmonisation projects and within the HDSS. It gives an overview on the state of the art on the standards and associated tools for public health research data documentation. It also points out the weaknesses of these standards and the tools related to the documentation of provenance of the ALPHA harmonised datasets.

Chapter 3: STUDY SETTINGS – ALPHA Network

This chapter will look at the rationale for setting up HDSS, their field and office data operations, the formation of the ALPHA network and its data management practices and the network's studies. In addition, it also describes the CiB data documentation capabilities.

Chapter 4: OPEN-ACCESS FOR EXISTING LMIC DEMOGRAPHIC SURVEILLANCE DATA USING DDI

This chapter investigates a recipe for the implementation of metadata standards within an HDSS. It analyses the choices, steps and considerations to be made by a typical HDSS at the beginning of a DDI data documentation endeavour. This includes what version of DDI to use, what tools to use, personnel training considerations and software costs involved.

Chapter 5: HIGH LEVEL, STRUCTURED METADATA FOR ALPHA DATA TRANSFORMATIONS:

This chapter focusses on the development of high-level metadata for ALPHA data harmonisation. It brings together the GLBPM, its specialisation, and the DDI 4 and Pentaho information models to develop structured high-level metadata to describe ALPHA data provenance. It seeks to develop business level provenance metadata.

Chapter 6: LOWER LEVEL, STRUCTURED METADATA FOR ALPHA DATA TRANSFORMATIONS:

In chapter 6, a more granular description of the harmonisation routines is developed by mapping the Pentaho data transformation details to the SDTL and documenting data aggregates, in form of data quality metrics, using the SDMX standard. These metadata complement the high level metadata developed in Chapter 5.

Chapter 7: PROVIDING END USERS WITH ACCESS TO ALPHA PROVENANCE METADATA

This chapter elicits and analyses the perspectives of data management and research experts working within ALPHA and other data harmonisation projects. It seeks to define a list of requirements for a user-friendly provenance metadata browsing software for ALPHA.

Chapter 8: CONCLUSION

This chapter will summarise the main findings from the thesis, discuss how the various pieces of the project fit together, conclusions to be drawn from this work and recommendations for future research building on to the presented work.

2. LITERATURE REVIEW

2.1 Introduction

The literature on data documentation is very broad. Most substantive disciplines of research have their own metadata standards and best practices (I. Barkow 2016; Murtha Baca 2008). This chapter focuses on data documentation literature applicable to public health research and epidemiology. The existing literature is reviewed through addressing the following questions and topics on the demand and provision of metadata useable by both humans and computer agents in this domain:

- Data documentation and sharing in public health research
 - Standardisation, metadata standards, generic process models and attendant technologies and tools
 - Use of metadata standards and associated tools in HDSS studies
 - Data documentation practices among multi-site data harmonisation collaborations
 - How far do current metadata standards support documentation of data transformations?
- Discussion of the main message from the literature review

This chapter seeks to establish that though there are standards and tools in the data documentation community available for HDSS studies to use, none of the existing work readily fits ALPHA's documentation needs. Thus, contextualisation and extension are needed. The discussion section at the end reiterates the need for further work to customise and extend existing standards for use in HDSS data harmonisation and similar projects.

2.1.1 Scholarly databases, keywords and literature search strategy

Three scholarly databases were searched for publications to use in this review. These are PubMed, Web of Science and Library, Information Science & Technology Abstracts. The combinations of keywords and concepts shown in Figure 1, complemented by Boolean operators provided the search strategy for relevant papers to include. Each of the strategies under the four headings in Figure 1 was used individually and then in various combinations using the "AND" operator to search for literature on two or more topics.

Figure 2 shows the results of the searches conducted in the scholarly databases PubMed, Library, Information Science and Technology Abstracts and Web of Science. 491 articles initially retrieved were excluded from the literature review due to two main reasons. (1) They

were from domains such as GIS, neuroscience, genetics and astronomy - these have their own standards which are unsuitable for datasets produced in population-based longitudinal studies. (2) They did not refer to data documentation, metadata, data curation or data sharing.

Full text screening resulted in the exclusion of 53 more articles. Though these referred to data sharing or data management, they did not give any details regarding data documentation. Scholarly databases searches resulted in 20 articles being selected for literature review.

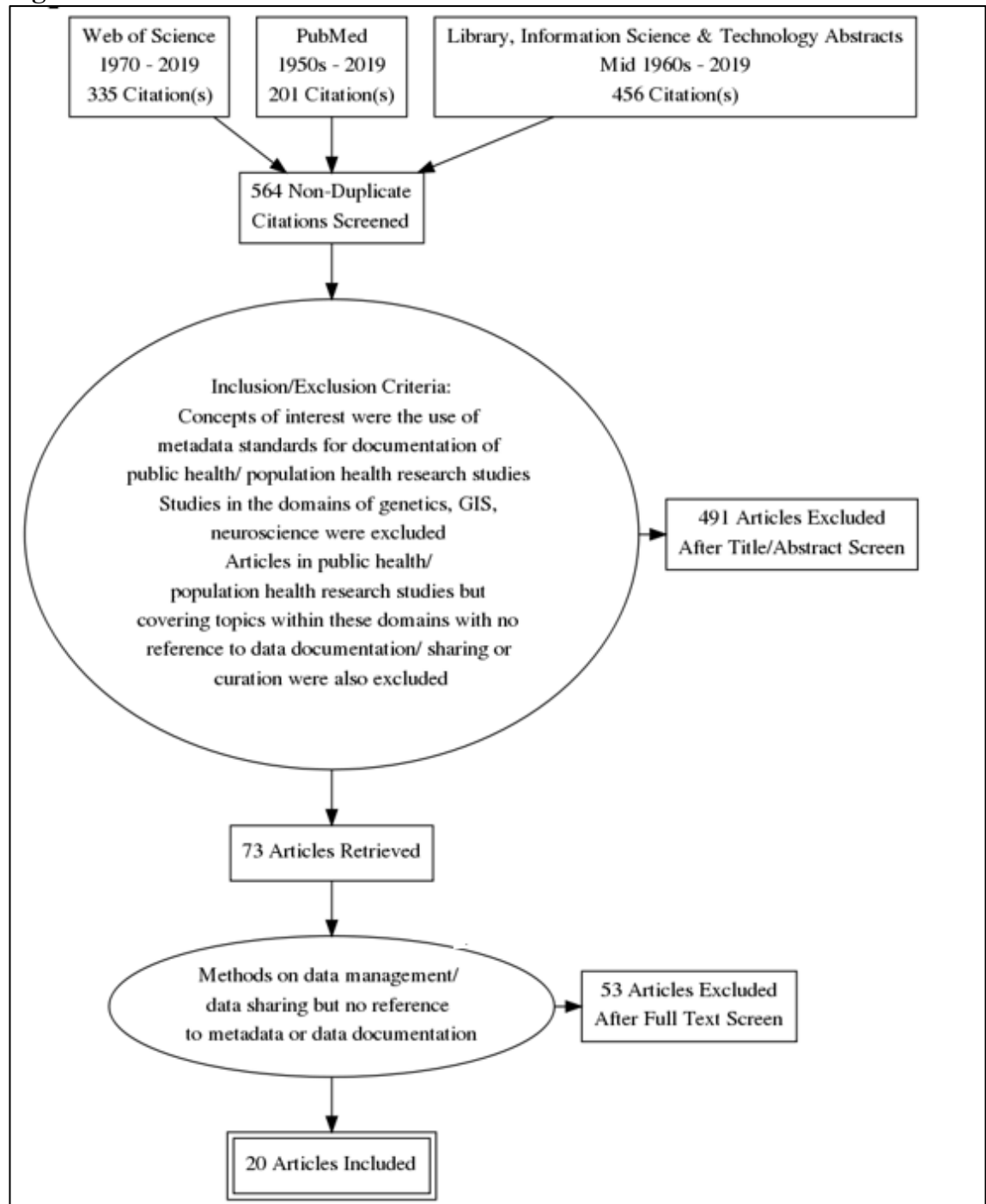
Figure 1: Broad literature review topics and search strategies

<p>Data documentation ("Data documentation" OR metadata OR "Documentation standard" OR SDMX OR DDI OR "Statistical Data and Metadata Exchange" OR "Data Documentation Initiative" OR "data annotation" OR "data curation" OR "data archiv*" OR "data repository" OR "data catalog*" OR "data sharing" OR "data discovery" OR "open data" OR "research repository" OR "data resource" OR "data standard*" OR "research data management" OR "Findable Accessible Interoperable and Reusable" OR "FAIR data" OR "FAIR principle*") NOT "drug-drug interaction*"</p> <p>Health and demographic surveillance and other population-based studies ("Health and demographic Surveillance system" OR "Demographic surveillance" OR "Demographic surveillance system" OR "Health and demographic surveillance systems" OR "cohort stud*" OR "population-based stud*" OR "longitudinal surveillance" OR "longitudinal observation" OR "longitudinal stud*" OR "Population-based cohort" OR "public health surveillance" OR "Longitudinal data" or "public health data" OR "individual participant data") NOT (genom* OR Clinica* OR Trial)</p> <p>Data harmonisation and or pooling studies Multinational OR "multi-national" OR "multi-countr*" OR "multi-site" OR "data harmoni*" OR "data transformation" OR "Extract Transform and Load" OR ETL OR "data integration" OR "data processing" OR multicent* OR harmon* OR "retrospective data harmonization" OR meta-analysis</p> <p>Low and Middle Income Countries "Low and Middle income countri*" OR LMIC OR "developing countr*" OR "sub-sahar*" OR Africa OR "sub sahar*"</p>

Backward and forward citation tracking, using the aforementioned databases and Google scholar, supplemented the articles retrieved using the search strategies. In addition, targeted searches on institutional websites for relevant content, conference presentations and working papers were also conducted.

This review is not systematic and cannot claim to be exhaustive. It however helps to shed light on the state of art in terms of data documentation in public health research and HDSS studies including in studies where data harmonisation is involved.

Figure 2: Literature search results



2.1.2 General picture of data documentation and sharing in public health research

The public health research domain has generally been described as sluggish in terms of adopting metadata standards and data sharing (Bergeron et al. 2018; Chandramohan et al. 2008; Pisani and AbouZahr 2010; Walport and Brest 2011; Wellcome Trust 2014). Concerns to do with confidentiality of the study participants and the need to meet primary research

grant aims before making the data available have often been given as reasons for restricted data sharing (Pisani et al. 2016). Meanwhile, the funders have sent a clear message regarding their demands for deriving maximum health benefits from publically funded research via data sharing (Walport and Brest 2011; Pisani et al. 2016). Journals are singing the same chorus too. A number of publishers including Plos, Springer Nature, Science and Elsevier have policies that promote public access to data used in a publication submitted to them (Federer et al. 2018). Support for publishing data profiles in the International Journal of Epidemiology (Oxford University Press 2019) or in dedicated data journals (Pauline Ward 2016; Candela et al. 2015) is paving ways for this to happen but the landscape is still very much varied (Wellcome Trust 2014).

Calls for data sharing in the literature are tantamount to calls for data documentation usable by both humans and computer agents. This is so because effective data sharing needs to comply to the widely cited FAIR guiding principles for data management and stewardship (Wilkinson et al. 2016). These principles guide those wishing to enhance the discovery and reuse of their data holdings on how to make their data Findable, Accessible, Interoperable and Reusable - FAIR (Wilkinson et al. 2016). FAIR principles acknowledge the need to accommodate both human and software agents as key users of the data resources. While humans have better intuitive sense of reading cues to work out the meaning and intent of digital objects than computers, they need the assistance of computers to cope with the scale and complexity of present day data. Computer agents need structured documentation to discover and manipulate digital data in automated ways (Wilkinson et al. 2016; Miller and Vardigan 2005).

2.2 Standardisation, Metadata, and metadata standards

2.2.1 Standardisation in Public Health research and epidemiology

Standardisation is not a foreign concept in public health research. McMahon (2017) classified health information standards in two broad categories in her thesis – encoding standards and data exchange standards. Encoding standards are controlled terminologies/ vocabularies used to systematically organise information to support knowledge management. These include vocabularies such as the MeSH, ICD, SNOMED CT and LOINC (McMahon 2017). On the other hand, exchange standards are frameworks for supporting clinical information exchange. They incorporate terminologies in their broader scope of representing information exchanged in clinical management systems. They serve as domain models for health

information constructs and processes. The HL7 FHIR and CDISC are examples in this category (McMahon 2017). Most public health researchers working with HDSS data would be familiar with the MeSH and ICD 10 encoding standards. The MeSH is used for literature and biomedical concepts searching while the ICD 10 is used in the analysis of morbidity and causes of death. Probably fewer of them have exposure to HL7 and CDISC. These two are more likely to be used by data managers who are responsible for data processing and exchange.

The encoding standards snugly fit as part of the metadata standards described in the next section. The DDI metadata standard uses controlled vocabularies for a number of metadata fields where only specific standardised content is expected. On the other hand, the metadata standards serve different purposes from those served by HL7 and CDISC. However, the need for interoperability between the foregoing exchange standards and DDI, for example, is acknowledged in the data documentation community (William Block et al. 2012). This has in turn led the data documentation community to grow the data and record types that, using FAIR principles, it is capable of describing.

2.2.2 Metadata definition, types and uses

Metadata are usually defined as data about data (NISO 2004). They refer to the information that describes the location, context and significance of a resource, usually on the internet, helping to retrieve and manage it (NISO 2004).

Metadata differ by discipline or professional community (Murtha Baca 2008). In the past it was mainly the concern of information professionals but with the advent of the internet, users who are not necessarily information professionals are creating metadata (Murtha Baca 2008). Relating to digital resources, metadata serve the following purposes:

Resource discovery, access, organisation, interoperability, long term preservation, sharing and digital identification among other uses (Murtha Baca 2008).

If the resource of interest are data – the users of the data need metadata in order to understand and use in an effective and responsible manner, the data they are provided with (Gregory, Pascal, and Ryssevik 2009). For research results to be meaningful, the producer of the estimates needs to understand the input data. While primary users of data have a wealth of informal and “undocumented” knowledge, a secondary user only relies on the documentation accompanying the data (Gregory, Pascal, and Ryssevik 2009).

In cases where data from various sources and time periods are combined, metadata will facilitate accurate processing and aggregation of those data. The longer the distance between the production of the data and the use of those data, the more important the metadata

become for accurate use of the data (Gregory, Pascal, and Ryssevik 2009). This also relates to the use of the data for purposes other than those foreseen by the producers.

Metadata are also responsible for providing enough detail for the resource to facilitate access and to “future proof” electronic resources (Corti and Gregory 2011; Van den Eynden et al. 2011). For microdata, this goes beyond a simple data dictionary. It entails information of study design, data collection and processing methods, details of deviations from initial collection and processing plans and so on. The aim is to provide enough details to help potential users to decide on the quality of the data and their suitability for the users’ needs. This level of detail has traditionally been provided by data codebooks (Inter University Consortium of Political and Social Research 2019) – mainly in form of free text paper, word processing or PDF documents. While these traditional codebooks provide the metadata required by human data users, they pose particular challenges for the other equally important users, the computer agents (Wilkinson et al. 2016; IHSN 2012). These latter users need structured information in order to exchange and process the data and metadata in automated ways.

The advent of the internet and the subsequent need to exchange data and metadata among computers via this platform has been one of the main drivers increased demand for machine actionable metadata (Gregory, Pascal, and Ryssevik 2009; IHSN 2012). Technologies have been created to store, transport and aid the processing of these metadata.

2.2.3 Metadata and information technologies

XML

There is a shared understanding among modern information technology practitioners that data exchange among computer systems requires standardised information models and a common language (IHSN 2012). This understanding has resulted in the establishment of the eXtensible Markup Language (XML). XML is both a language and a suite of technologies (IHSN 2012). As a language, it has syntactical and grammatical rules to ensure that XML documents are properly written. It comprises of a suite of technologies which include XML Schema for describing the structure of the XML document, XPath and XQuery for searching and querying the XML and many others (IHSN 2012). XML provides the required structure for automated data exchange and processing. In addition, its conversion into forms that are human user-friendly such as web pages and PDFs can be automated. This makes XML a viable metadata storage format catering for both the needs of humans and software agents.

The semantic web

The semantic web is a vision to express web content in a form that is more easily processible by computer agents and to use software agents in a way that takes advantage of this representation (Antoniou and Van Harmelen 2004). The current state of the information on the web is organised in such a way that the meaning of the contents – semantics – is accessible to humans – mainly in the form of hyperlinked HTML documents. While documents can be located by computers through provided links, software agents are not able to understand the meaning (Ontotext 2019). The difficult task of finding and mining meaning of these documents is left to human beings. When structure is added to the information, software agents can process it and derive meaning in automated ways (Ontotext 2019).

Despite its critical role in automated data exchange, XML has two limitations relating to the ideas propagated in the semantic web movement. The first limitation is that it does not sufficiently capture the semantics of the data being exchanged (Antoniou and Van Harmelen 2004). The second one is to do with the use of the XML Schemas. XML schemas exist to ensure that a given XML document is valid – compliant with the structure defined in the schema. Ironically, this functionality which is important to ensure validity of XML documents, tends to restrict the XML documents from augmenting themselves with relevant data/ metadata which are not included in what the schema caters for. This makes interoperability across metadata standards expressed in XML only difficult. Semantic web technologies in the form of Resource Description Framework (RDF) and Web Ontology Language address this limitation in XML.

Resource Description Framework (RDF) and Web Ontology Language (OWL)

A framework for structured description of the web content called Resource Description Framework (RDF) developed by the World Wide Web Consortium (W3C) makes it possible to process the information in automated ways (Ontotext 2019). It structures the information into statements comprising of a subject, a predicate and an object. The statements are described as triples. Uniform Resource Identifiers (URI's) are used to identify the objects, their properties and values. An extension of RDF, a vocabulary called RDF Schema (RDFS), adds the properties and classes of RDF resources and the hierarchies of those classes and properties. The properties and classes of the resources are further enriched in their expressiveness and their semantics formalised through the Web Ontology Language (OWL). OWL is an ontology language for the semantic web (Antoniou and Van Harmelen 2004).

Metadata standards

Metadata are usually organized in metadata standards. Metadata standards represent a common view of how metadata within a domain of interest can be described and exchanged (Gregory, Pascal, and Ryssevik 2009). Metadata standards range from the multi-purpose to the more specialized standards that are used to describe resources in different fields (Greenberg 2005).

Table 1 provides a summary of the standards and their areas of application. This list is not exhaustive. A more comprehensive list of metadata standards is provided on the DCC website (Digital Curation Centre 2019)

The current study does not go into detail about the various standards. Rather, it analyses the use of the Data Documentation Initiative (DDI) specification which is suitable for the documentation of social science and population health individual level datasets (Rasmussen and Blank 2007a; Wellcome Trust 2014) similar to those collected in the ALPHA network. In addition, it also applies the SDMX standard for documentation of aggregates generated in the course of assessing the quality of the created ALPHA datasets.

Table 1: Metadata schemes and their fields of application

Metadata standard	Purpose/ Area of use
Dublin Core (DC)	Describes resources on the web. It has multidisciplinary application (NISO 2004)
Text Encoding Initiative (TEI)	Marking up electronic texts for example; novels and plays, This standard primarily supports research in the humanities (NISO 2004)
DataCite Metadata Schema	Complete and consistent identification of a resource for citation and retrieval (DataCite 2019)
EML - Ecological Metadata Language	A metadata specification for the ecology discipline (DCC 2019)
FGDC/CSDGM - Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata	Standard defining metadata for describing digital geospatial data required by the USA Federal Government (DCC 2019)
ISO 19115	A metadata standard for geographic information and services (DCC 2019)
PREMIS (Preservation Metadata: Implementation Strategies)	Metadata standard for long term preservation of digital objects (DCC 2019)
PROV	Standard for enabling interoperable interchange of provenance information in heterogeneous environments (DCC 2019)
Data Documentation Initiative (DDI)	Metadata specification for description of microdata produced by surveys or other observational methods in the social, behavioural, economic, and health sciences (DDI Alliance 2018d)
Statistical Data and Metadata eXchange (SDMX)	Metadata standard for the description of statistical datasets, their exchange and sharing (SDMX Technical Working Group 2018a)

2.3 Metadata standards and process models applicable to demographic and epidemiological surveillance data

2.3.1 Dublin core

Dublin core is a domain agnostic, generic resource description metadata standard for discovery on the web (DCC 2019). It is implemented at two levels which are simplified and qualified (NISO 2004). It was formed at the metadata workshop done in 1995 in Dublin Ohio (NISO 2004). It thus derived its name from the place where the workshop was held. The Dublin Core Metadata Initiative is responsible for maintaining this standard (DCC 2019). It was endorsed as an International Standards Organisation standard (ISO Standard 15836:2009 as of February 2009). The simplified Dublin Core comprises of fifteen elements which are title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage and rights (NISO 2004). There are more detailed versions of the Dublin Core, the qualified level, which allow the user to provide more granular descriptions of some elements through the use of qualifiers (DCC 2019). For example, the date element can be qualified by distinguishing between date of creation of a resource and date of modification. Due to its simple nature, Dublin core is used as part of many metadata standards mainly providing bibliographic metadata. In this study, it is implemented as part of the DDI standard. It is easy to learn but, by itself, it is not up to the task of expressing the complex relationships or concepts such as the ones required for ALPHA data.

2.3.2 Data Documentation Initiative (DDI)

Introduction

The Data Documentation Initiative is an XML-based specification for documenting microdata obtained from surveys or other observational methods in the social, behavioural, economic, and health sciences (DDI Alliance 2018d). It was first conceived in 1995 (Vardigan, Heus, and Thomas 2008) mainly by social science archives who sought to standardise the descriptions of the data that they were receiving from depositors (William Block et al. 2012). Its original aim was to cover the archival aspects of social science data (Data Documentation Initiative 2009). DDI is currently being developed and maintained by the DDI Alliance (<https://ddialliance.org/>). The original model, which has matured into DDI Codebook, was not suitable for documenting longitudinal data. It was also not able to support changes in the study instruments, or study concepts as the study progressed and had limited machine actionability (Data Documentation Initiative 2009), these limitations are still

present with the current Codebook version. However, the particular strengths of Codebook are that it is simple and easy to understand (William Block et al. 2012). These together with the availability of tools have led to the rapid and worldwide implementation of Codebook.

The introduction of DDI 3, also known as DDI Lifecycle, has made it possible to document each stage of the data lifecycle from the study design to the data analysis and repurposing (using data for purposes that were not foreseen during the study design phase) stages. DDI Lifecycle does all that DDI Codebook can do. In addition, it supports the documentation of two or more rounds of longitudinal data or related studies through the use of its groups, resource package and comparison components (Hansen et al. 2011).

Currently, there are two DDI strands used in production systems – DDI Codebook and DDI Lifecycle. Further, DDI 4 is also under development.

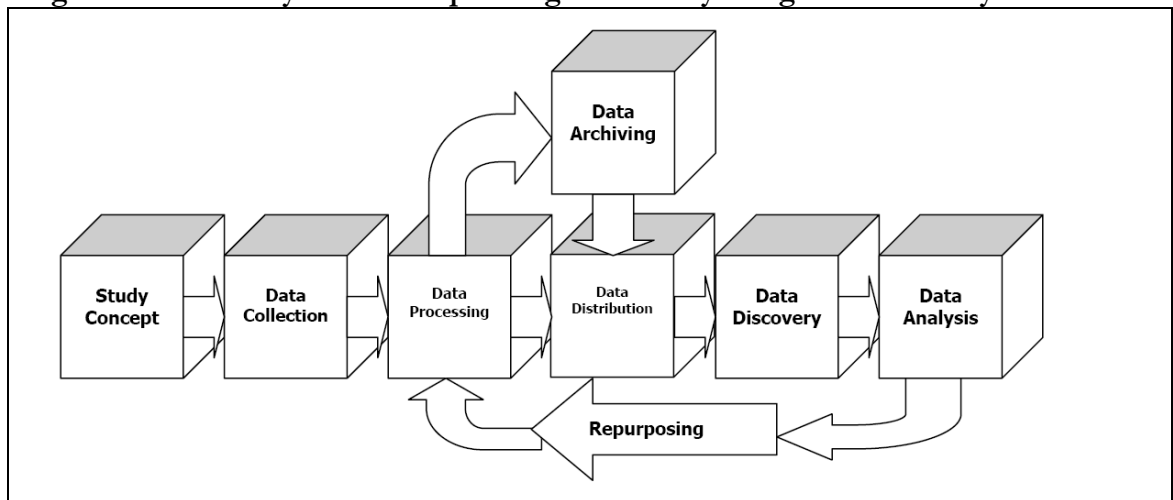
DDI Codebook

Codebook comprises 5 sections which are document description, study description, data file description, variable description, and other materials (Dupriez and Greenwell 2007).

The document description section captures bibliographic metadata about the DDI file including who prepared it, the identifier for the DDI file, its version and when it was prepared – most of the elements in this section are from the Dublin Core standard. The study description section captures high-level, metadata about the study as a whole. The data file description section describes the study data file. The variable description section describes each of the variables in the data file. Most of the structured metadata are around variable documentation. Starting with variable label, associated question(s) from which the variable was created, response options, data type, notes and many more. Very much like the content one would find in a data dictionary or codebook. The fifth section is the other study materials section. This section describes the questionnaires, technical reports, and other resources used in the study.

DDI Lifecycle

DDI Lifecycle, based on the data lifecycle model (Figure 3), is more comprehensive than Codebook. It is made up of two parts, the conceptual model and the XML Schemas derived from the conceptual model.

Figure 3: Data lifecycle model spanning from study design to data analysis

(Data Documentation Initiative 2009)

In addition, two packaging structures worth mentioning are the modular structure and the metadata Schemes.

Modules bring together sets of information related to specific activities in the data lifecycle. There are about 16 modules in total in DDI Lifecycle. Among them are the study conception module which entails metadata elements to do with study conception and design, the data collection module capturing metadata relating to questionnaires or other instruments used for data collection, the logical data structure module and the physical data structure module contain sets of metadata related to data processing, the archiving module which contains metadata elements related to archival aspects and so on (Data Documentation Initiative 2009; DDI Alliance 2014a). The five sections of documentation in Codebook are still present in Lifecycle and are spread throughout the Lifecycle modular structure as shown below in Table 2.

Lifecycle's expanded scope includes support for reuse of metadata throughout the data lifecycle by providing identifiers for metadata elements and publishing them. For example, a set of response categories used in a questionnaire during data collection can be entered in Lifecycle and given a unique identifier. These responses can then be referred to on the basis of that identifier in the later stages of the data lifecycle. For instance, a variable corresponding to the question of interest would not need a new entry of response options, the ones entered as part of the data collection metadata will be reused. Similarly, concepts, questions and response options and many other metadata elements can be reused through this referencing system across several waves of a study.

Table 2: Correspondence between Codebook sections and Lifecycle modules

DDI Codebook	DDI Lifecycle
Document Description	Archiving module
Study Description	Study conception and Data collection modules
File description	Physical structure data structure
Variable description	Data Collection Process, Physical Data Structure and Logical Data Structure
Other study materials	Other material class of the relevant module

Metadata schemes are lists of reusable items of a specific type (DDI Alliance 2014a). For example, response categories, question items, concepts and so on could be placed in schemes. This provides a grouping of enumerated representations of concepts.

Lifecycle also possesses structures for metadata exchange and long term metadata management (DDI Alliance 2014a).

Compared to Codebook, Lifecycle represents a major advance from primarily focussing on data archiving and human readable metadata. It provides support for every stage of the data lifecycle and increased machine actionability.

DDI 4 – Moving forward

DDI 4 is a new version of DDI currently under development. It may be published as its own product or incorporated over time into the DDI 3 series (DDI Developer personal communication). This version is an advancement of the capabilities of the standard and will be based on an information model (William Block et al. 2012). The model, a representation of important artefacts involved in the entire data lifecycle and their relationships will give DDI flexibility in terms of technical expressions, unlike the current versions which are primarily expressed in XML. In addition, it will improve communication with other disciplines and standards (William Block et al. 2012). It also seeks to cater for data resulting from a wider variety of collection methods including administrative registers, electronic health records, measurements from medical equipment and instruments (William Block et al. 2012). Thus, it expands from the survey questionnaires method primarily catered for in the current versions used in production systems. It is still focused on supporting the research data lifecycle (Figure 3).

DDI 4 comprises of two parts, a library of classes and a set of functional views (DDI Alliance 2014b). The library of classes includes primitives, extended primitives and classes which makes use of the primitives. The classes and primitives within the library comprise the information model presented in the Unified Modeling Language (UML) (DDI Alliance 2014b). On the other hand, a subset of classes can be combined together within a functional view to support a particular use case (DDI Alliance 2014b). This use case may be related to

data collection, dissemination, and management among other functions. In this study, a data management use case is considered. The classes are organised in packages depending on the part of the data lifecycle that they pertain to (DDI Alliance 2014b).

Being based on an information model makes the new version more flexible than the current ones. Codebook and Lifecycle are presented in XML with no explicit information model. An information model is at a higher level of abstraction than a data model (Schoenwaelder and Pras 2003), it does not impose an implementation format, thus for implementation, DDI 4 has the potential to be expressed in a variety of technical formats that include XML, relational database schema, semantic web technologies (revisited later when the relationship between DDI and the semantic web is discussed) (William Block et al. 2012) and so on.

The proposed expansions to DDI include better communication and interoperability with other standards, documentation of process and workflows, sampling and qualitative data (Dan Gillman and Arofan Gregory 2015). While all these expansions are important and deserve attention, it is the structured annotation of process and workflow that is of particular interest to this thesis. This ability has potential to aid the description of data transformations performed to create ALPHA datasets. Though DDI 4 introduced process and workflow documentation right from the initial iterations of the model, these aspects have mainly focussed on data collection with insufficient coverage of data management activities and information objects (DDI Alliance 2015a). They require augmentation in order to adequately capture the data transformation processes in ALPHA.

Incompatibility between DDI versions

One of the challenges currently facing the DDI standard is the incompatibility between versions. DDI has gone through a series of changes since its inception. A number of versions of the DDI Codebook strand have been developed, among them, DDI 1.x, DDI 2.0, DDI 2.1 and DDI 2.5 (DDI Alliance 2018c). Same goes for Lifecycle which has DDI 3.0, DDI 3.1, DDI 3.2 (DDI Alliance 2018c) and DDI 3.3 currently under development. Compatibility is very limited between Codebook and Lifecycle and also between versions of each strand. It requires tools to convert between these versions and across the strands. The Colectica tools (discussed later in section 2.4.2) have conversion capabilities and tools such as Sledgehammer (DDI Alliance 2018c) also offer this capability. With Colectica, this would involve costs of purchasing the services tools. Sledgehammer has a community version and a commercial version. Its free version is limited to a certain number of variables and observations.

In light of these incompatibilities and provisions made within DDI 4, also still under development, HDSS studies need to concentrate on the identification and provision of

optimal metadata for the data under their custody using the DDI versions catered for in the tools readily available/ affordable to them. In this case, Nesstar Publisher which is provided for free and discussed in section 2.4 of this thesis. In addition, the HDSS community will need to keep up with ongoing work on development of alternative metadata editors and conversion tools to identify and adopt the ones most suitable for their circumstances when change of tools is needed.

DDI RDF vocabularies

One of the main interests the DDI community in the semantic web technologies is ability of these technologies to facilitate interoperability between DDI and other standards. Because Codebook and Lifecycle are in XML, some effort have gone into creating RDF/ OWL vocabularies for sections of these DDI versions (Bosch et al. 2013; Joachim Wackerow, Larry Hoyle, and Thomas Bosch 2014; Cotton et al. 2013). With the DDI 4 version, RDF and XML are the currently available bindings (formal language representations) of the information model.

2.3.3 Statistical Data and Metadata eXchange (SDMX)

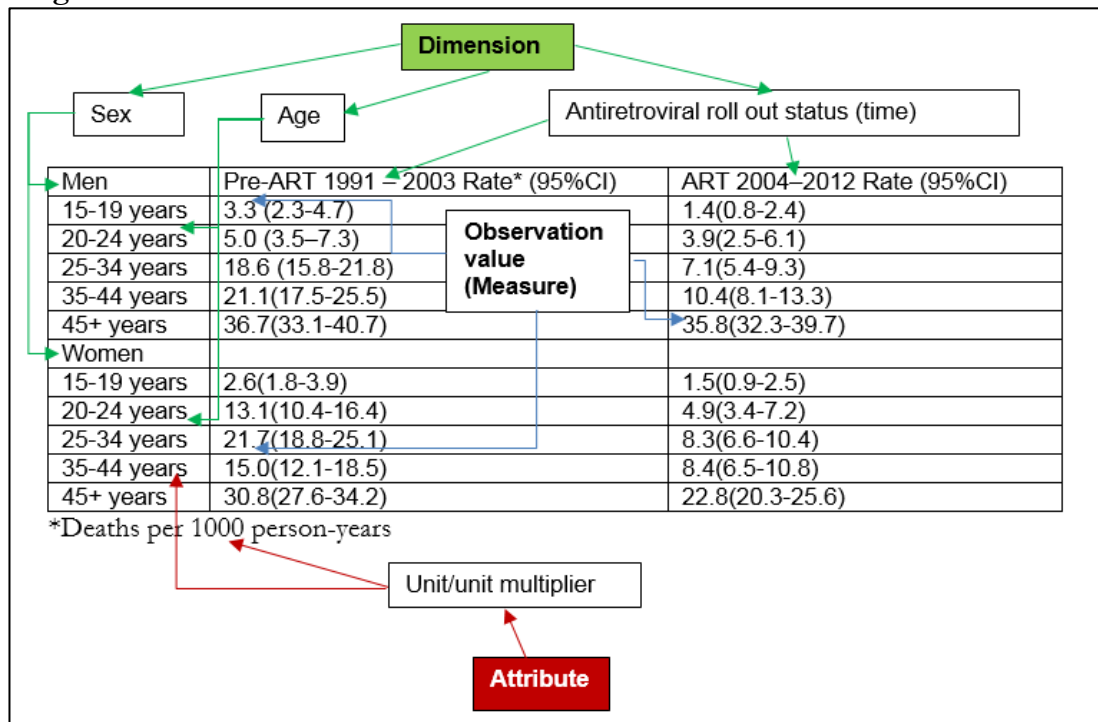
SDMX is an International Organisation for Standardization (ISO) standard for describing statistical data, normalising their exchange and sharing standard for statistical information exchange (SDMX Technical Working Group 2018a). It was formed in 2001 by seven institutions concerned with the production and exchange of official statistics, financial and economic data. These are the Bank of International Settlements, the European Central Bank, Eurostat, the International Monetary Fund, the Organisation for Economic Co-operation, the United Nations Statistics Division and World Bank (Stahl and Staab 2018). SDMX consists of technical standards, statistical guidelines and IT architecture and tools (SDMX Technical Working Group 2018b). It provides a way of modelling statistical data, metadata and data exchange processes. While it has microdata describing capabilities, its forte are aggregated statistics presented in multidimensional tables (Gregory and Heus 2007)

Technical standards

SDMX has an information model that captures the data and metadata structures and data exchange related characteristics of a dataset and metadata of interest. Relating to data structure, a data structure definition (DSD) defines the characteristics of the data by identifying and defining concepts and sub classifying them into dimensions, attributes and observation values (Stahl and Staab 2018). Concepts represent the basic building blocks elementary for the understanding of the data. Dimensions are the uniquely identifying or

classifying properties of data points of a dataset and measures are the actual observation values. Attributes do not have identifier characteristics; they are only descriptive. A combination of dimensions that uniquely identifies a value within a multi-dimensional table for aggregated data is referred to as a key. The description provided by the DSD is called structural metadata. In addition, concepts have a representation. This is either in form of coded values or textual. Dimensions are always coded values while attributes can either be coded values or textual. Figure 4 shows examples of the various elements of a DSD. It draws from published mortality rates for an HDSS in the south western part of Uganda, the Masaka general population cohort (Asiki et al. 2016).

Figure 4: Examples of dimensions, observation values and attributes for data on mortality rates in Masaka HDSS in Rural Uganda before and after Antiretroviral drugs roll out.



A combination of sex, age, antiretroviral drugs roll out period uniquely identifies a mortality rate value in the table so that combination is a key. The age is measured in *years* and the mortality rates are expressed per *1000 person-years*. These are attributes, giving additional information about the data.

A metadata structure definition (MSD) provides additional description of the concepts through what are referred to as “reference” metadata in SDMX speak. These are metadata to do description of the content of concepts, methods used to create the dataset of interest and the quality frameworks.

Statistical guidelines

These come in form of content oriented guidelines concerned with the harmonisation and interoperability of terminology, Codelists, classification of statistical subject matter domains among other things and the proper implementation of the SDMX standard (Eurostat, Directorate B: Statistical Methodologies and Tools and Unit B-5: Statistical Information Technologies 2010)

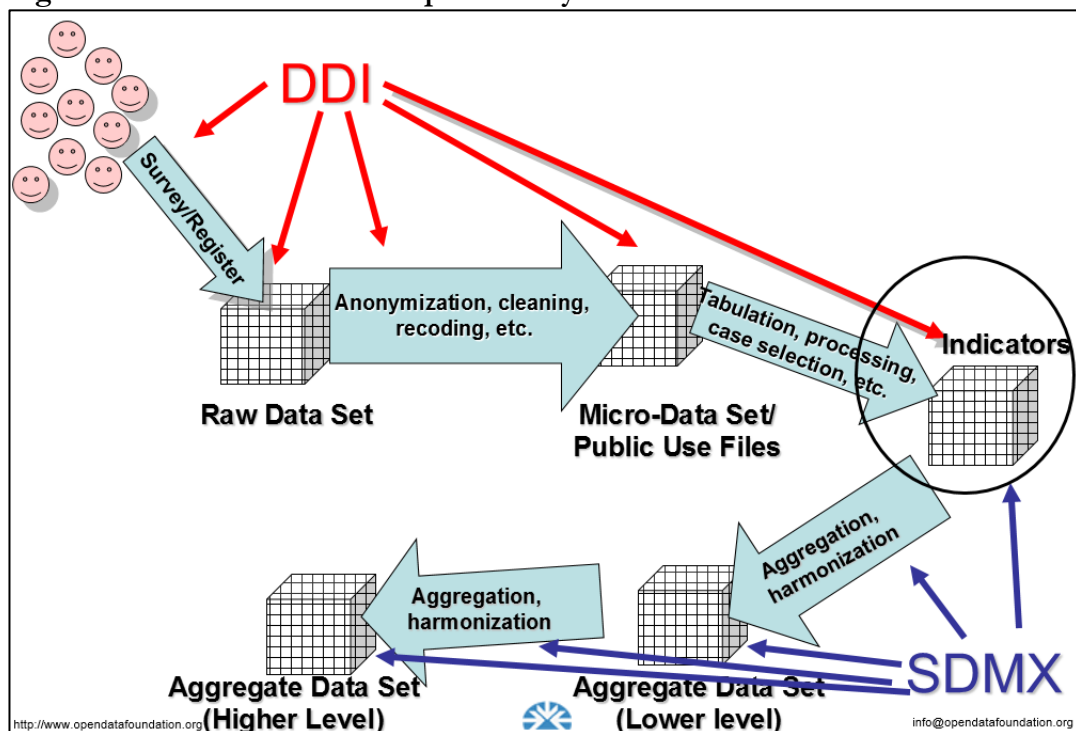
IT Architecture and tools

Besides the information model and the statistical guidelines, SDMX also comprises of formats for the metadata, IT architecture which considers basic process patterns and tools for implementing the standard. These are described in detail in the SDMX literature (Eurostat, Directorate B: Statistical Methodologies and Tools and Unit B-5: Statistical Information Technologies 2010).

SDMX and DDI

SDMX and DDI have been considered to be generally complementary with DDI having strengths in the microdata description area while SDMX catering for aggregates. The diagram in Figure 5 depicts this relationship between the two standards. However, this diagram is old, having been published in 2008. Both standards have since gone through various changes which may have resulted in a relationship that is different from what is shown in the figure.

Figure 5: DDI and SDMX complementary nature



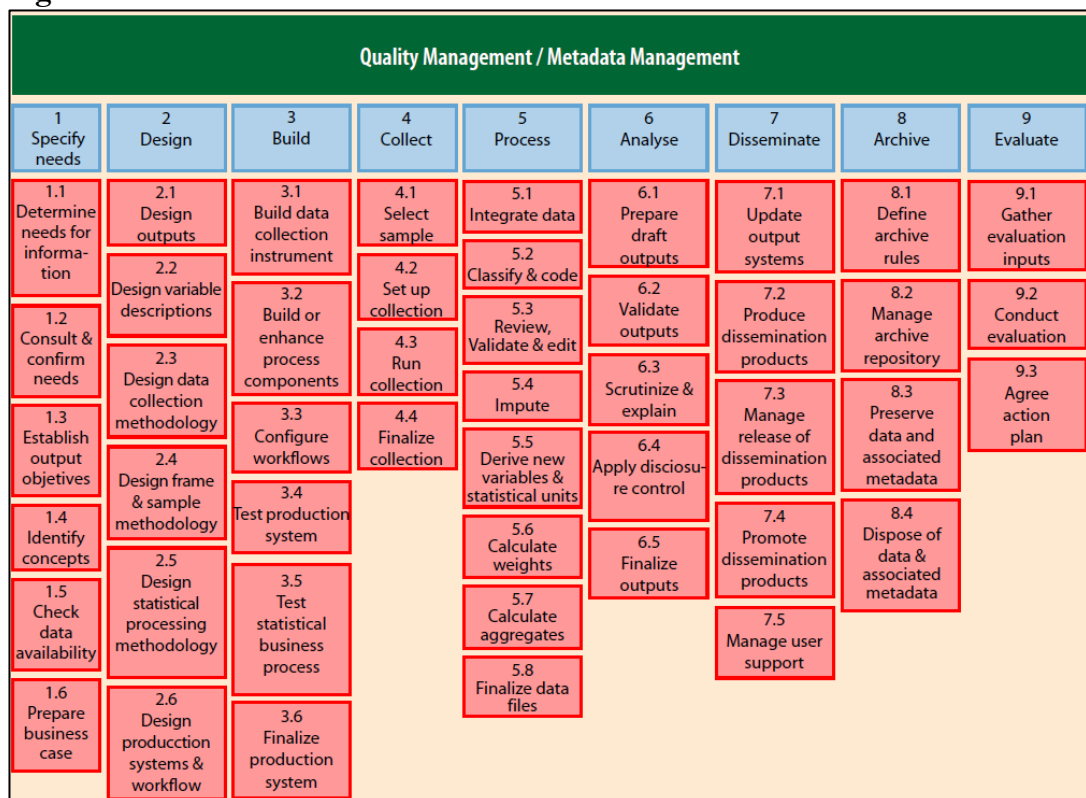
Source (Arofan Gregory and Pascal Heus 2008)

SDMX has been widely applied in official statistics. In the area of health, this model has been specialised to create SDMX – Health Domain (SDMX-HD) (Turbelin and Boëlle 2013). SDMX-HD however did not go beyond prototyping as it was considered too complex to implement (Braa and Sahay 2017).

2.3.4 Generic Process models

The aforementioned implementation standards give us the metadata content either produced or consumed at the various stages of the data life cycle depicted in Figure 3. However, they do not guide us regarding the processes/ activities involved in producing or consuming the data. Process modelling can fill this gap giving us a holistic view of the processes and facilitating the capture of metadata at their source thus minimising information loss (Ausborn, Rotondo, and Mulcahy 2014; I. Barkow 2016). Two models are considered: the Generic Statistical Business Process Model (UNECE 2018b; UNECE Secretariat 2009) and its specialisation for use in the longitudinal surveys realm, the Generic Longitudinal Business Process Model - GLBPM (B. I. Barkow et al. 2013).

Figure 6: Generic Statistical Business Process Model



Source (Thérèse Lalor and Steven Vale 2013)

Generic Statistical Business Process Model (GSBPM).

The GSBPM is a reference model for the process of producing official statistics and was developed by the official statistics community (UNECE Secretariat 2009) drawing inspiration from the statistics New Zealand's generic business process model (Dunnet 2007). Figure 6 shows the GSBPM.

The GSBPM aims to provide a standard way to describe procedures within and between national and international statistical organisations. It is part of the “Modernisation of Official Statistics” effort by the High Level Group on Modernisation of Official Statistics (UNECE 2019b). Modernisation of statistics production was embarked on to address the challenges facing statistical organisations by streamlining and standardising the data production processes and services. The challenges include increasing demand for data products, the advent of big data, increased competition for skilled labour and budget cuts (Thérèse Lalor and Steven Vale 2013). GSBPM is one model among many covering various aspects of the modernisation of official statistics production. GSBPM facilitates process definition and description in a coherent way, provides common terminology and a framework for quality assessment. It comprises of nine main phases which are “Specify Needs”, “Design”, up to “Evaluate”. Each of the phases has sub-processes under them giving further details on what the phase entails.

Generic Longitudinal Business Process Model GLBPM

The GLBPM describes the process of longitudinal research survey data production.

The GLBPM is a specialisation of the GSBPM for purposes of modelling longitudinal survey data production. The GLBPM can be mapped to DDI lifecycle model to identify which metadata content is associated with which phase of the data production process (B. I. Barkow et al. 2013). Figure 7 shows the GLBPM. It has nine high level phases of which each has sub-processes under it providing details of the various activities belonging to each of the phases. The GLBPM is a flexible and non-sequential model; it does not require that each data production process involve all the activities in Figure 7. A given production process does not have to follow a linear path from the first phase to the ninth.

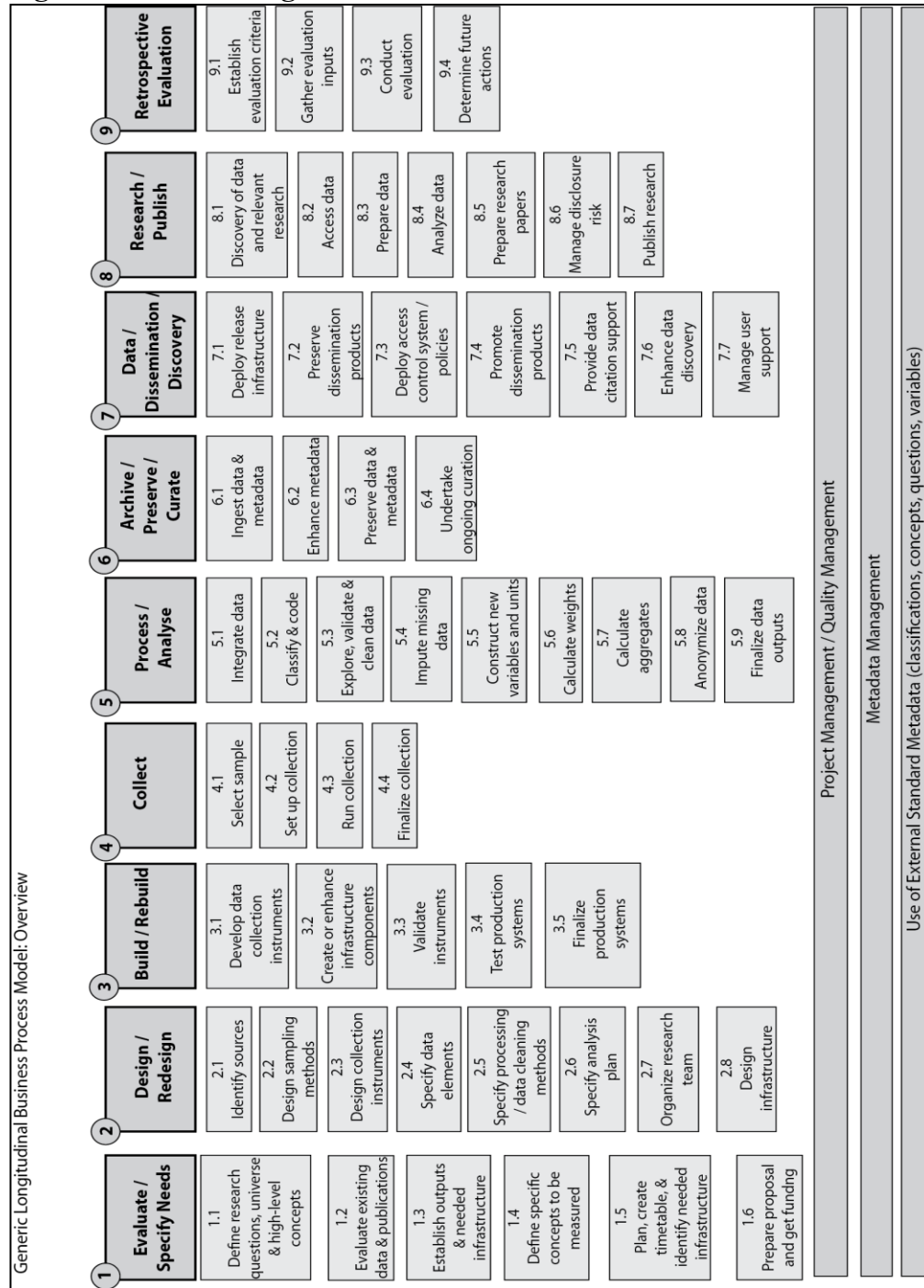
Another view of the GLBPM is the so called “tornado” view which considers multiple rounds of data collection for a longitudinal study (B. I. Barkow et al. 2013; Hoyle et al. 2011). This view is depicted in Figure 8. It shows how metadata and data flow from one wave of data collection to the next. The stages and processes passed through in the first round serve

to inform the second round leading to reconceptualisation of the study and/ or other changes as the study progresses over time (Hoyle et al. 2011).

Relating to primary HDSS data, changes in concepts, target populations and data collection instruments across data collection rounds can be captured through the metadata structure provided by the tornado view. For harmonised datasets such as those resulting from networks such as ALPHA, changes in universes, concepts, time periods, granularity of response categories, assumptions on loss to follow-up cut-off points and other missing data can be also be documented within this framework.

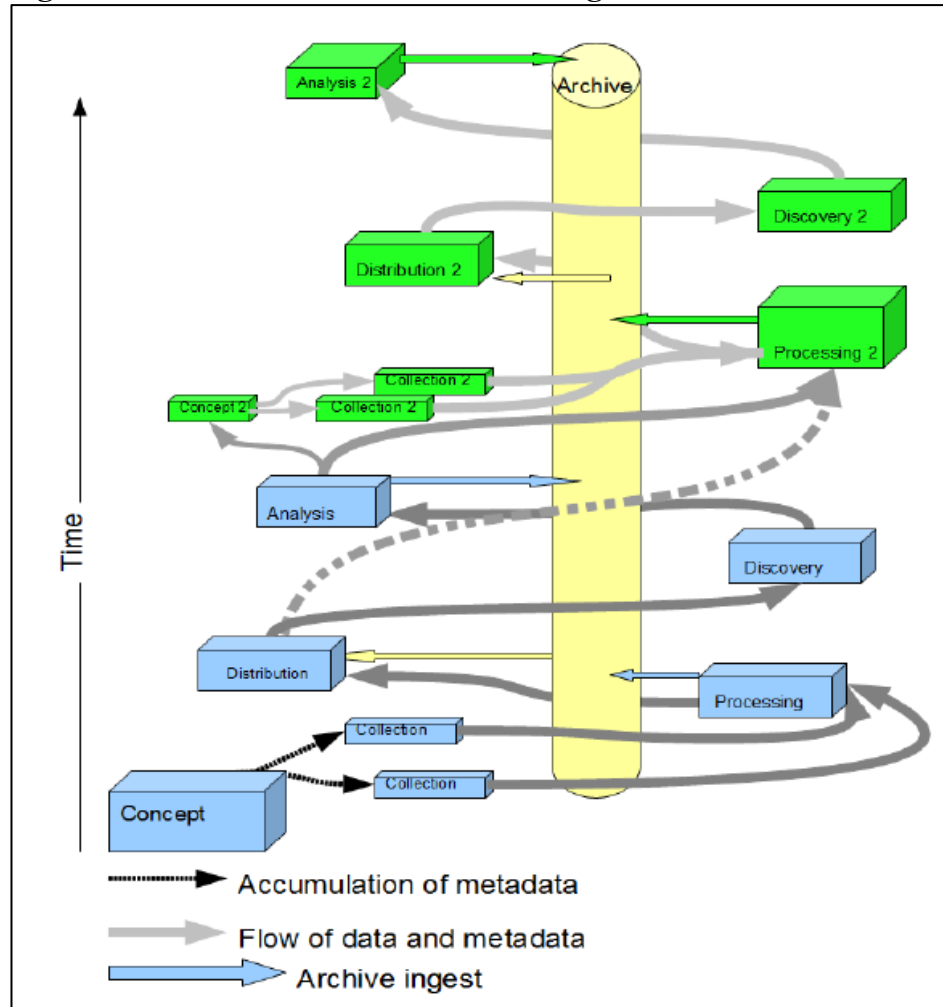
However, this thesis is mainly concerned with pinning down optimal metadata for a single traversal/ or data pipeline involved in harmonising HDSS data to a specification provided by a network such as ALPHA. Metadata showing links between two traversals can only be determined after those for one traversal have been thoroughly investigated and determined.

Figure 7: Generic Longitudinal Business Process Model



(B. I. Barkow et al. 2013)

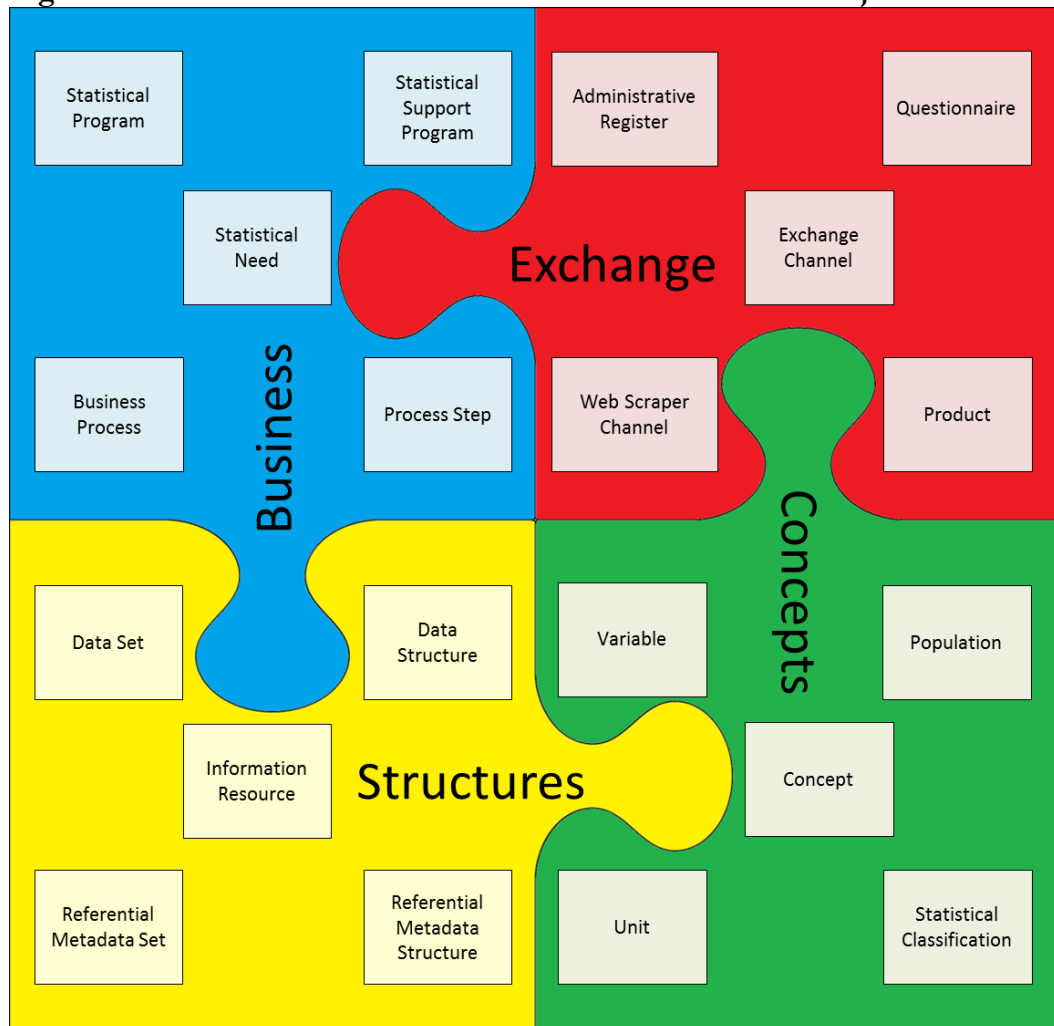
Figure 8: GLBPM "Tornado" view showing two rounds of data collection



(Hoyle et al. 2011)

2.3.5 Generic Statistical Information Model (GSIM)

GSIM is an information model, designed within the official statistics domain for capturing, at a conceptual level, the pieces of information (information objects) flowing between the activities involved in the production of statistics (UNECE 2018a). The information objects involved include data, metadata, rules and parameters among others. It is designed to complement the GSBPM capturing the information objects used in or produced from the sub-processes of the GSBPM. It comprises of five groups of information objects which are base, business, exchange, concepts and structure. Each of these groups comprise of classes of objects relevant to them. Figure 9 shows four of the five groups and the information objects within each of those groups.

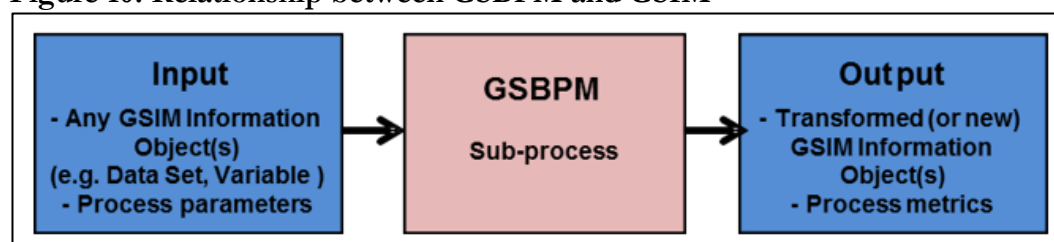
Figure 9: Generic Statistical Information Model information objects

Source (UNECE 2018a)

The full detailed model is huge and cannot be presented on page but it is available on the UNECE managed online platform (UNECE 2019a) showing all objects and their interrelationships.

GSIM and GSBPM

GSIM supports the GSBPM by describing the information objects flowing between the GSBPM sub-processes as illustrated in Figure 10.

Figure 10: Relationship between GSBPM and GSIM

Source (UNECE 2018a)

GSIM, DDI and SDMX

DDI and SDMX, as implementation standards, are aligned as much as possible with GSIM (UNECE 2018a). As a conceptual model, GSIM does not compete with DDI and SDMX, rather it add a conceptual layer between the GSBPM and the implementation standards to help reduce the gap between the process model and its implementation (I. Barkow 2016). Using GSIM can help to distinguish between information differences that are conceptual and those that are purely technical (UNECE 2018a). This distinction is important in transforming and harmonising information objects coming from different implementation formats. For instance, if some information is captured in a DDI standard compliant format and other related information is captured within a database management system, mapping these information objects to GSIM can indicate which objects are actually conceptually different. This could be of use in creating structured metadata from legacy systems in HDSS settings where metadata are currently in diverse formats.

GSIM is primarily focussed on official statistics so it does not necessarily map to all the objects relating to HDSS research.

2.3.6 Metadata support for describing data transformations

Existing literature on structured metadata for data transformations can be classified into three approaches. These are (i) the approach taken in the current versions of DDI (DDI Alliance 2015; Marker et al. 2009), (ii) the Validation and Transformation Language (VTL) and (iii) the Structured Data Transform Language (SDTL) (C2Metadata 2017).

DDI Codebook and Lifecycle

Lifecycle can be used to support implicit and explicit data comparison through the “Group” and “Comparison” modules respectively (Hansen et al. 2011). These modules enable researchers to document similarities and differences in the data. While this comparison can be applied between the source data used in a data harmonisation exercise and the resulting data, it does not tell what was done to create the harmonised data.

Both Codebook and Lifecycle allow for a mixture of textual process descriptions and the inclusion of the source code in the documentation. Under DDI Codebook this is limited to the “recoding and derivation” element of the standard (Nesstar 2011; Dupriez and Greenwell 2007) which allows for free text description of data transformations relating to a variable. It also allows the addition of the source code.

Lifecycle has much more structure for process descriptions compared to Codebook, it has facilities such as “processing events”, “processing instructions” “lifecycle events” (DDI

Alliance 2018b). Processing events is descriptive and distinguishes between cleaning operations, control operations, weighting operation or data appraisal. Processing instructions include general and generation instructions. These contain a description and the command code (DDI Alliance 2018b).

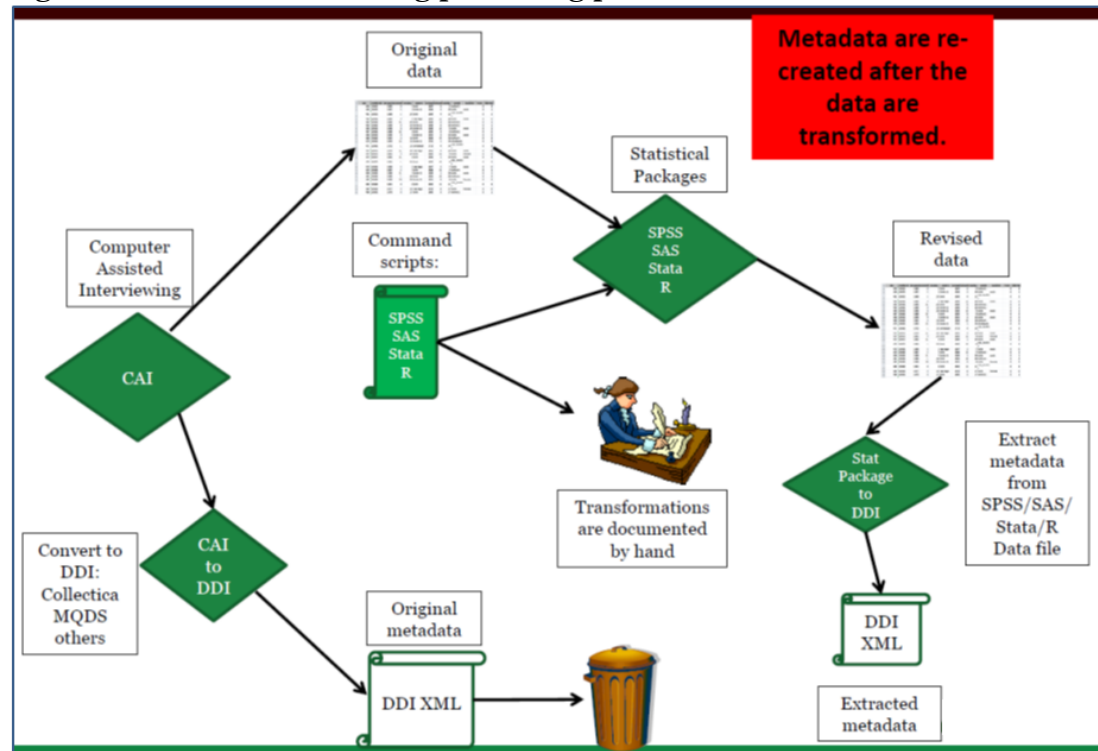
In their essence, the current provisions in both Codebook and Lifecycle give textual descriptions of data processing and the command code used. They leave users with proprietary software code as the description of the lower level data transformation detail. Thus, the disadvantages of using proprietary code for documentation are not dealt with in this solution. In addition to being proprietary, most of that code is heavily context dependent and it is difficult to track all the relevant sections of code that are spread out across files, this can undermine the utility of the information. For example if recoding is done first, and then new variables constructed later, it might all be relevant but the temptation might be only to include the recode step. So it is very dependent on the way the data manager thinks through the data harmonisation and organises their code.

Besides these DDI capabilities, literature has advanced our understanding on this topic in two fronts: the Validation and Transformation Language (VTL) and the Structured Data Transform Language (SDTL).

VTL and SDTL

VTL was developed as a data transformations extension to the SDMX standard (SDMX Technical Working Group 2018a). It captures the details of data validation and transformations carried out at variable and dataset levels. It also captures the transformations used for creating statistics from microdata. VTL aims to be executable. It is a standard syntax for expressing validation and editing rules (a set of operators, their syntax and semantics. It aims to first express the validation and transformation rules then convert them into specific programming languages for execution. It provides a technology neutral expression at business level of the processing taking place.

On the other hand, starting in 2014, a project called Continuous Capture of Metadata (C2Metadata) was proposed and funded by the US National Science Foundation. C2Metadata aims to address the metadata loss that happens during processing after electronic data collection as shown in the diagram in Figure 11.

Figure 11: Metadata loss during processing performed after data collection

Source (Alter 2018)

C2Metadata was formed to develop tools for documentation of data transformations carried out in the statistical packages SAS, SPSS, Stata and R (Alter et al. 2017).

The C2Metadata project is developing SDTL which documents transformations performed in statistical packages. The developers of SDTL are participants within the DDI community. The scope of VTL and SDTL is not the same and they have potential to be used in a complementary way.

In the context of ALPHA data integration use case, there is need for an intermediate language because Pentaho is being used to execute both Pentaho transformation steps as well as scripts written in Stata, R, Ruby, JavaScript, among many. We need a way of looking into all Pentaho steps, those done using inbuilt/ native steps and those running scripts from other software. Otherwise the latter are black boxes. SDTL provides that window.

The intention with VTL on the other hand is

“...to provide a language usable by statisticians to express logical validation rules and transformations on data, described as either dimensional tables or unit-record data. The assumption is that this logical formalization of validation and transformation rules could be converted into specific programming languages for execution (SAS, R, Java, SQL, etc.)...” (SDMX Technical Working Group 2018c, 2).

So using VTL we can produce scripts in various languages as output. In our use case we already have these scripts, mainly in Stata format. ALPHA rather needs a translation of Stata

code into a software agnostic form. SDTL offers that facility since the C2Metadata project has developed a Stata parser for converting Stata to SDTL. Moreover, the project is also working on expressing the SDTL in natural language (Ionescu 2018).

What both VTL and SDTL lack is a high level description of the data transformations that can be used for business communication. They provide a means for documenting the details, thus, for a generalist wanting an idea of what is going on in the transformations without being bogged down by the details, neither of these languages are ideal, they make it hard to see the forest for the trees.

Generic process models

The GSBPM has been widely adopted by national statistics offices across the high income nations (Brancato and Simeoni 2012; UNECE Secretariat 2009; Ausborn, Rotondo, and Mulcahy 2014). Often, the national statistics offices have mapped their own business processes to the model. In some cases, attempts have been made to specialise the GSBPM to meet local contexts. In the majority of these cases, no attempts has been made to add structure to the specialisation. The Australian Bureau of Statistics (ABS) is one of the few examples where the specialisation has been done in a structured fashion (UNECE 2018c). The ABS example is the most detailed mapping to date in terms of contextualising the GSBPM in a structured format. ABS adapted GSBPM and GSIM to create their own models which they have been using and continuously improving upon. The use of these models have been reported to have improved communication, reduced costs of data production through reuse and increased potential for automation of tasks (Alistair Hamilton, Eden Brinkley, and Therese Lalor 2012). The ABS application covered a scope broader than what is aimed for in this thesis, it considered the entire data life cycle and is done within the official statistics domain. This thesis is particularly looking at the provenance of harmonised datasets.

2.4 Software tools for implementing metadata standards

2.4.1 Generic DDI Codebook tools

Data documentation tools are also described in the literature. The two major codebook-based tools are the Nesstar suite (Digital Curation Centre 2013) and Dataverse (King 2007).

International Household Survey Network tools

The International Household Survey Network (IHSN) with funding from the World Bank, has integrated the use of a suite of generic, and open source DDI Codebook-based tools (International Household Survey Network 2018). The efforts by the IHSN and the World

Bank have facilitated the rapid uptake of these tools worldwide including in low resource settings. The software suite includes a metadata editor called Nesstar Publisher (Digital Curation Centre 2013) and a data cataloguing software called the National Data Archive (NADA) (International Household Survey Network 2016). Accompanying manuals and guides are also available, giving information on how to use the tools.

Despite these resources, a lot of decision making and customisation needs to be done at a local level for effective use. The World Bank launched a massive training and mentoring program to initiate the use of the IHSN toolset in developing countries' national statistical offices (IHSN 2013; Anne Thomson, Graham Eele, and Felix Schmieding 2013). This training program was called the Accelerated Data Program - ADP (IHSN 2013). As a result of the ADP, most national census and national surveys were documented and catalogued, what was not assessed is whether use of the existing data increased due to this program (Anne Thomson, Graham Eele, and Felix Schmieding 2013).

ADP had a specific focus on national official statistics. Consequently, individual HDSS projects wanting to use the tools need to dedicate time, staff and funds to the work of adapting the tools for local settings.

One potential drawback to the use of IHSN tools that HDSS may have to deal with in the future is the fact that the development of Nesstar Publisher has been discontinued (personal communication from DDI developer). This has implications for the maintenance of the metadata already created and the future preparation of metadata. The World Bank and IHSN are working on another metadata editor based on DDI 2.5 (Welch and Asghar 2018) but it is unclear if it is going to be compatible with the Nesstar Publisher produced DDI. If not compatible, to make sure there is continuity, tools will need to be developed for conversions of current Nesstar Publisher metadata to work with the new editor. This challenge is not unique to HDSS though as there are many other users of Nesstar Publisher. HDSS can leverage any solutions advanced by the wider Nesstar Publisher user community. It is also in the interest of the World Bank to develop a metadata editor that is compatible with Nesstar metadata since more than 10 years' worth of data on their Microdata Library (The World Bank Group 2019) have been documented in Nesstar Publisher.

Nesstar server

The use of the paid version of Nesstar suite of programs comprising a metadata editor and an online data cataloguing software called Nesstar server has generally been confined to the high income countries. A list of some of the Nesstar users is available on the Nesstar website

(Norwegian Centre for Research Data 2016). Among them are the Norwegian Centre for Research Data and the UK data services, formerly the UK data archive.

Dataverse

Dataverse network is an open source data repository software (Mercè Crosas 2011; King 2007). Unlike the Nesstar Publisher and NADA combination, it has a broader scope aiming to publish, reference, extracting and analyse research data (Mercè Crosas 2011). It comprises a central repository infrastructure and offers distributed ownership for data authors through virtual web archives called dataverses. Dataverse is currently widely used for publishing data either through the Harvard Dataverse, a repository accepting data from all researchers worldwide and from all disciplines, or from the individual installations around the globe (Mercè Crosas 2011). The low entry barrier (data from all researchers worldwide and from all disciplines) facilitates depositing of data but poses the challenges of interoperability as depositors bring their data in various formats which are not easy to integrate (Wilkinson et al. 2016). The Kenya Medical Research Institution Wellcome Trust Programme is the example of Dataverse African users closest to the ALPHA network (Robert W. Snow 2017; Snow et al. 2017; Ouma, Okiro, and Snow 2018; Irish et al. 2019). However, none of the data directly coming from the HDSS affiliated to ALPHA (Odhiambo et al. 2012) are on this Dataverse.

Other project specific tools

Besides IHSN and Dataverse network tools, other implementers have developed their own, this includes the Inter-university Consortium for Political and Social Research (ICPSR) and Integrated Public Use Microdata Series (IPUMS) among others.

2.4.2 DDI Lifecycle Tools

Colectica toolset

The Colectica toolset (Colectica 2019a) are the main Lifecycle-based tools for metadata entry and management in the generic category. Colectica has two free versions which are Colectica for Excel (Colectica 2019c) and Colectica Reader (Colectica 2019d). Colectica for Excel is a light weight version which enables dataset and variable-level documentation and the reading of DDI 3 documentation in Excel. Colectica Reader is a tool for viewing and validating DDI 3 metadata. It validates against a DDI 3.2 Schema, highlights missing metadata elements and inconsistent references (Colectica 2019a). On the commercial side, Colectica Designer is used for creating DDI Lifecycle, Colectica Questionnaires for survey specification, Colectica

Repository and Portal for storing, version control and cataloguing of the data and the metadata (Colectica 2019a).

The complexities of DDI Lifecycle and the costs of purchasing and maintaining this software have caused the use of Colectica tools to be generally confined to well-resourced nations in Western Europe, Canada, New Zealand and USA, as shown on the list of past and existing users (Colectica 2019b). No African group has used Colectica, to the best of the author's knowledge, for documenting or cataloguing data.

Open source Lifecycle tools

Metadata technology North America (Metadata Technology North America Inc 2019) has built free and commercial versions of generic software for transitioning between versions of DDI (DDI Alliance 2018c).

While they are generally declared to be open source, Questasy (CentERdata 2019), DDI on Rails (Hebing 2015a), DDIEditor (Jensen 2012) and the Rogatus suite (I. Barkow and Schiller 2013) among other DDI Lifecycle tools, were all designed for use within specific projects with potential for wider use. Questasy was created for use in the Dutch LISS panel and has also been used in other studies (CentERdata 2019), DDI on Rails was originally designed for use in the SOEP panel (Hebing 2015b), the Danish Data archive developed DDIEditor (Jensen 2012) and the Rogatus toolset was conceptualised within the DIPF, the German Institute for International Educational Research (Ingo Barkow 2015). All the institutions involved have indicated that they aim to make their tools sufficiently generic. However, as it stands, any user wanting to adapt them will need to edit the source code to suit their needs. A listing of other available DDI based tools is available on the DDI Alliance website (DDI Alliance 2018c) distinguished by whether they are commercial or free and other basic metadata, most of them have particular focus on questionnaire development, conversion between DDI formats and so on. The degree to which they are truly generic is a subject for debate, but at least their authors have made them available as DDI based tools.

Due to being based on Codebook or Lifecycle, these tools inherit the limitations of these DDI versions when it comes to capabilities that are required for documenting the provenance of harmonised datasets.

2.5 Data documentation practices among data harmonisation projects

There are many public health research data harmonisation projects reported in literature (O'Neill et al. 2019; Cooper et al. 2011; Cui et al. 2018; Fortier et al. 2010; Reniers et al. 2016; Herbst et al. 2015; Næss et al. 2007). However, as (Fortier et al. 2017) of the Maelstrom research (Maelstrom Research 2019) point out in their paper, until their proposal, there had not been a formalised guide to ensure high quality retrospective harmonisation. They went on to propose a comprehensive guide in the same paper comprising of 6 major steps. These steps are (0) defining the research questions and objectives, (1) gathering pre-existing knowledge, (2) evaluating harmonisation potential (3) data processing, (4) evaluating the quality of the harmonised data, (5) disseminating and preserving the harmonisation products. In addition, Maelstrom has also built a software suite for harmonisation and dissemination of multi-studies data (Doiron et al. 2017; Bergeron et al. 2018). Regarding structured metadata, they incorporate the DDI Lifecycle metadata standard. The extent to which they use Lifecycle is not clear from their publications.

Another significant effort regarding tools for data harmonisation documentation is the work on the CharmStats software within the GESIS Institute for Social Sciences (Winters and Netscher 2016). CharmStats is primarily based on transformations done using the SPSS software and seeks to support data harmonisation projects by organising, documenting and publishing them. The current version of CharmStats (1.1) does not directly handle the software packages relevant to ALPHA, Stata and Pentaho, data nor is it based on the existing metadata standards. Stata files are handled by first converting them to SPSS. However, there are plans to accommodate both Stata and DDI in version 1.2 of CharmStats (Winters and Netscher 2016).

Among data harmonisation projects, the CLOSER project (O'Neill et al. 2019) is arguably among the most comprehensively documented ones. CLOSER is a data harmonisation project constituting 8 UK birth cohorts (O'Neill et al. 2019). It has developed DDI Lifecycle-based documentation created from Colectica software and other home grown and open source software (CLOSER 2019). The CLOSER project data documentation has particular strengths in the area of questionnaire documentation for the data collection instruments from the 8 UK birth cohorts. This structured documentation has culminated in the CLOSER Discovery platform (CLOSER 2019).

All the work accessed from the existing literature have valuable lessons and tools that are of benefit to the broader work that ALPHA is involved in. They are however weak in the area

of structured metadata for data transformations. None of them is providing for provenance metadata in the manner aimed for in this project. Among them all, even the best documented of these, have their metadata based on the DDI Lifecycle version. The capabilities and drawbacks of the current Lifecycle version have been considered in section 2.3.5.

The other issue with the cited projects is that they are based in the high income nations and better financially resourced for data management as compared to partners within the ALPHA network. The solutions they offer may not smoothly work in the locations where ALPHA is operating.

2.6 Metadata standards implementation in HDSS studies

There are a number of publications alluding to the importance and lack of documentation for HDSS data (Chandramohan et al. 2008; Pisani and AbouZahr 2010). These data play an important role of bridging the data gap caused by incomplete vital registration and statistics systems (Setel et al. 2007). They contribute towards breaking the link between material and information poverty (Sankoh and Byass 2012). For the health benefits derived from HDSS data to be maximised, the data do not only need to be collected, processed and archived with care, they need to be also integrated into the wider data network (Wilkinson et al. 2016). The HDSS studies that are implementing structured documentation are mainly using the IHSN tools to create DDI Codebook (Ifakara Health Institute 2019; African Population and Health Research Center 2015; Africa Health Research Institute 2018). Though fully considered in Chapter 3, it is important to briefly mention the documentation practices related to the CiB infrastructure here. The CiB uses Codebook to document the finalised harmonisation products (Herbst et al. 2015). Besides the tool specific metadata relating to the data harmonisation routines performed via CiB, there is no other metadata created to describe the harmonisation processes. Structured documentation of the harmonisation processes is therefore still unexplored in the CiB.

2.7 Discussion

This chapter has provided a broad overview of the area of structured data documentation in the public health domain with a focus on publications relevant to the documentation of HDSS data. It has included the current state of the commonly used standards and their capabilities, the available tools and practices in data harmonisation collaborations and HDSS studies. It has given evidence of the importance of structured documentation for meaningful data exchange, sharing and reuse among humans and computer agents. It has also pointed

to our lack of understanding of three aspects of documentation relevant for describing ALPHA harmonised datasets.

Regarding documentation of HDSS primary data which is used to create ALPHA harmonised datasets, the literature shows the end to the documentation in form of data catalogues, without showing the means to that end. We saw examples of HDSS that have documented and catalogued their data, but there is no guidance on how to implement the documentation, what steps, consideration and choices to make in customising the tools and standards in the HDSS context.

Regarding documentation of data transformations, the literature acknowledges the importance of documenting data provenance but does not suggest sufficient provision for documenting provenance in the existing standards.

While the literature tells us about the importance of tools to facilitate the uptake of metadata standards, none of the available tools provides a solution for ALPHA datasets off the shelf. Therefore, though there are many advances in the data documentation technologies and successful application of these standards in various contexts, the three issues listed above remain unaddressed in the HDSS and their collaborative contexts. The ensuing chapters are a contribution towards addressing these outstanding issues. In Chapter 4, the implementation of metadata standards for documenting primary HDSS datasets is tackled, Chapters 5 and 6 address the extension of the DDI standard to better address data management and Chapter 7 analyses the requirements for a user-friendly provenance metadata browser.

3. STUDY SETTINGS – ALPHA NETWORK

3.1 Introduction

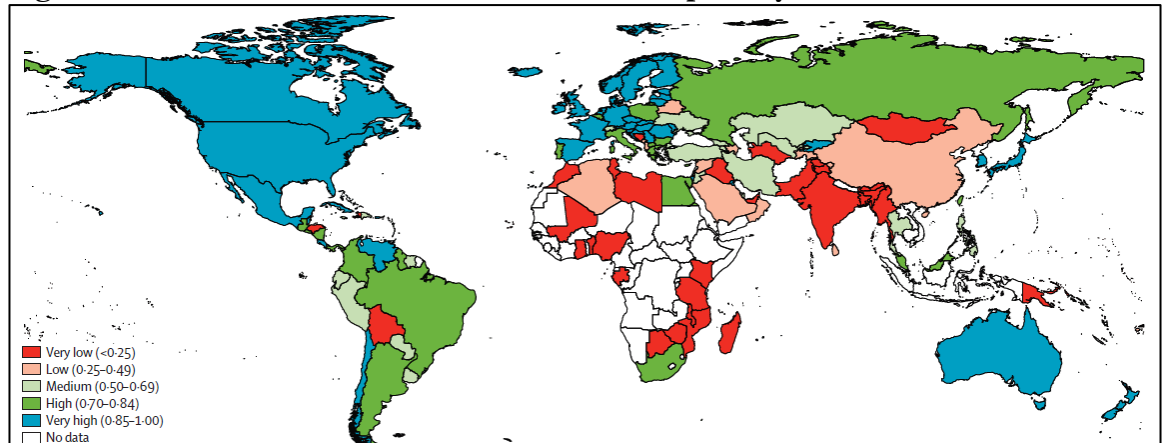
This chapter provides the context within which the study was done – the ALPHA network. The independent health research centres in the network operate HDSS. Therefore, HDSS are considered in the next section. The HDSS description include the rationale for their establishment, the core concepts underpinning their operations and the commonly used HDSS reference data model. Second, an overview of the ALPHA network is given which includes its formation and membership composition. Next, its data management and documentation practices are described including its adoption of the Centre in a Box technology. The chapter then ends with a summary.

3.2 Health and Demographic Surveillance System

An HDSS is a combination of field and computing procedures for collecting demographic, health risks, exposure and outcomes data from a defined population, within a defined geographical area on a longitudinal basis (Sankoh and Byass 2012; INDEPTH Network 2002). The data collected within HDSS have been used to show demographic and epidemiological trends, and for capturing the health related indicators missed by national health care services. HDSS however have the limitation that they are not nationally representative as they typically cover small geographical areas (INDEPTH Network 2002; Reniers et al. 2016).

3.2.1 Rationale for establishing HDSS in LMIC - defective civil registration and dearth of reliable vital statistics

Civil registration, the continuous and universal recording of occurrence and characteristics of vital events of births, deaths, marriages and divorces in a country (United Nations. Statistical Division 2001) is a source of vital statistics. Vital statistics provide the requisite information for development and health sector planning and evaluation (Setel et al. 2007). However, civil registration and vital statistics (CRVS) are incomplete and or defective in most LMIC (Mikkelsen et al. 2015). Figure 12 shows the dearth of CRVS in LMIC. Beyond poor CRVS, there is a general lack of population- based data on health across the LMIC (Sankoh and Byass 2012). In the short to medium term, alternatives have to be used. These include national censuses, nationally representative surveys and HDSS (INDEPTH Network 2002). ALPHA network mainly works with data from HDSS, therefore, this project does not consider national census data or the national health surveys in any substantive way.

Figure 12: The dearth of vital statistics in LMIC especially sub-Saharan Africa

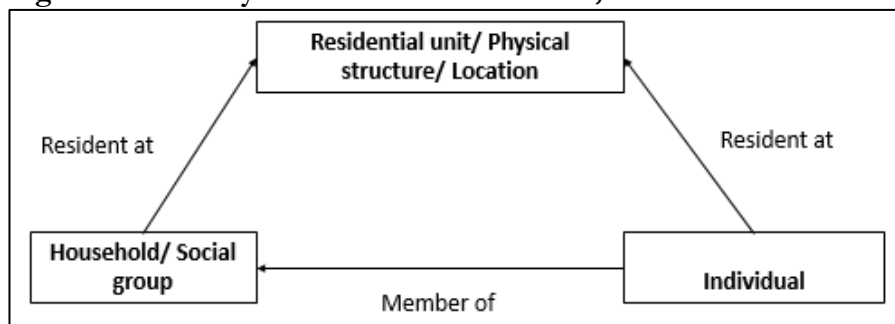
Source: (Mikkelsen et al. 2015)

3.2.2 HDSS core concepts

The main concepts circumscribing the existence and functioning of an HDSS are defined and thoroughly explained in a summary paper produced by the INDEPTH Network which has provided a forum for HDSS to discuss technical and scientific issues (INDEPTH Network 2002). I briefly summarise them in this section as they are foundational to the ALPHA network data management and analysis.

Initial HDSS setup

A typical HDSS starts with identifying a geographical area, clearly demarcated on the ground, where the study site will be located, also known as a Demographic Surveillance Area. This is followed by defining the population for surveillance using clearly defined inclusion/exclusion criteria. An initial census is carried out recording details of the primary entities for an HDSS shown in Figure 13 (individual, social unit – usually household and residential unit (physical structure) to which the individual belongs).

Figure 13: Primary entities: Residential unit, individual and social group

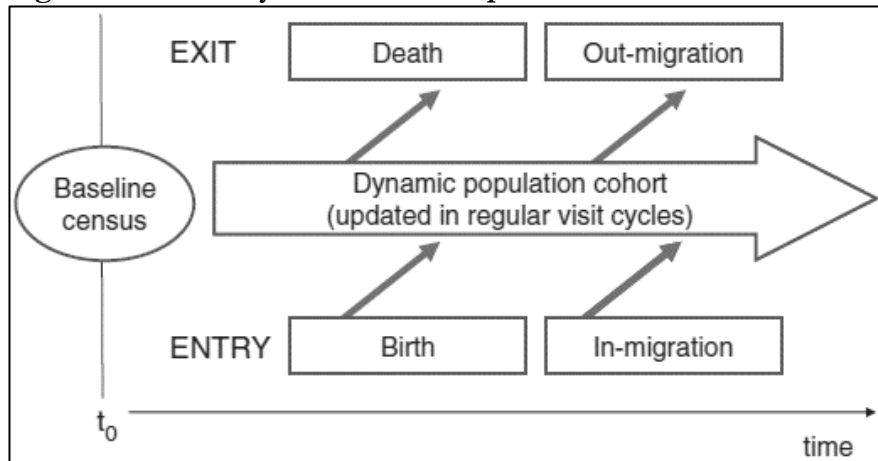
Adapted from (INDEPTH Network 2002)

Extensible unique identifiers are allocated to the primary entities. The initial census is also a basis for setting up a database system for managing the HDSS data.

Follow up visits/ Update rounds

After the initial census, follow up visits are required to keep track of the core events occurring to the primary entities - births, deaths and migrations. The main changes to be tracked are represented by the dynamic cohort schematic in Figure 14.

Figure 14: HDSS dynamic cohort representation



Source: (Sankoh and Byass 2012)

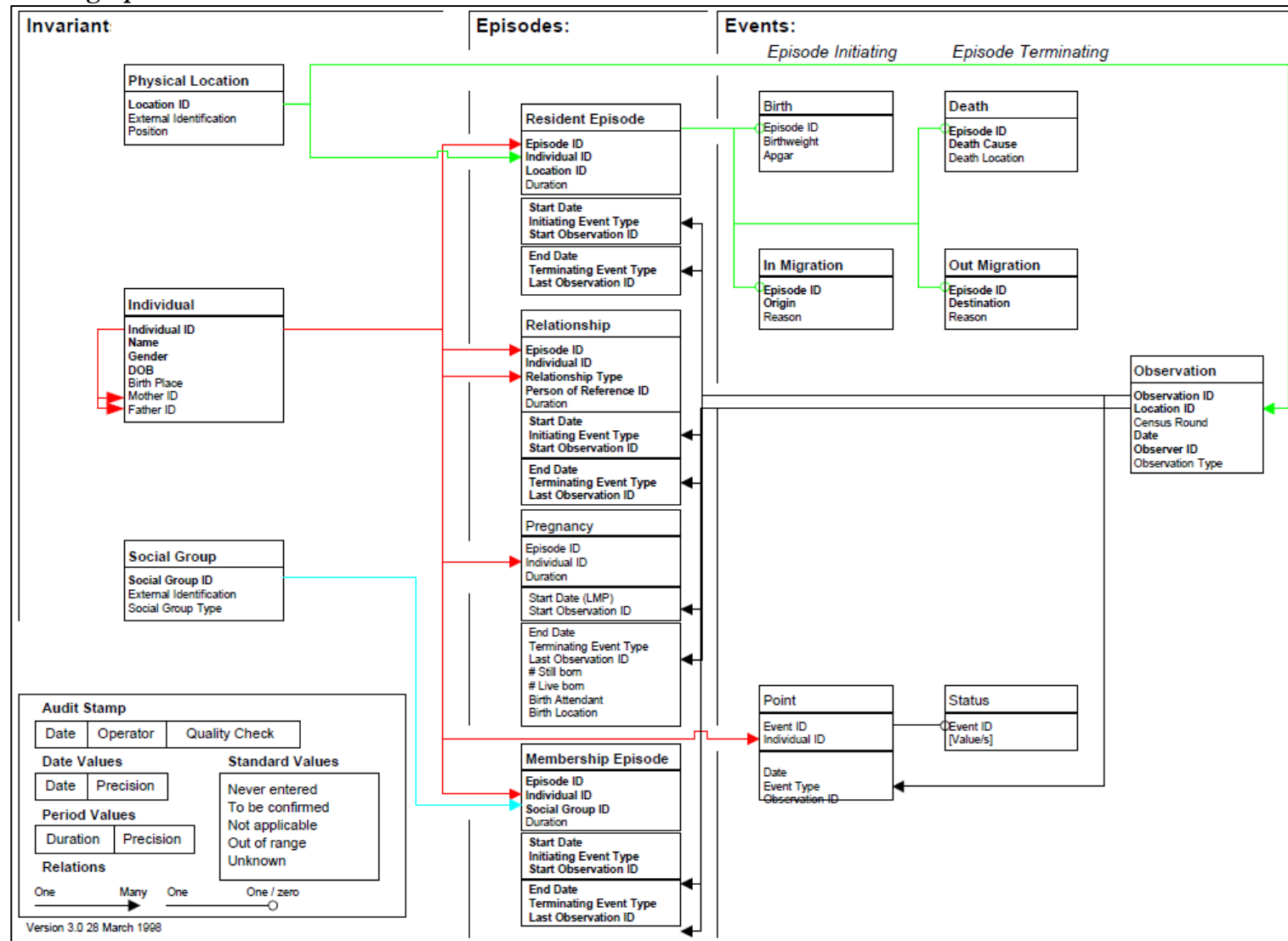
The HDSS population changes in a small number of ways: It is increased by new entrants in the form of births to resident mothers and immigrants coming from outside the surveillance area. It is decreased by deaths of resident members and out migrations to a destination outside the study area. Besides the migration that results in an alteration of the population size, individuals or social units also move within the study area, also referred to as internal migration (internal out-migration describes exiting one residence location to another and internal in-migration, the entrance of a residence location from another).

The majority of deaths of HDSS members occur outside health facilities due to the location of the study sites in areas that are typically poorly served. To ascertain cause of death, verbal autopsies are used with either a physician or a computer model (McCormick et al. 2016; Byass et al. 2013) assigning a likely cause of death (Sankoh and Byass 2012).

3.2.3 HDSS reference data model

The HDSS reference data model was first proposed by Benzler, Herbst, and MacLeod (1998). Over the years, several variants of this model have been developed. However, the original model, shown in Figure 15, is sufficient for purposes of this project.

Figure 15: Demographic surveillance reference data model



Source (Benzler, Herbst, and Macleod 1998)

It comprises the primary entities of an HDSS – Individual, social unit, physical location (shown on the left). It uses events to capture the way individuals enter or leave an area over time. A pair of events usually define an individual's residency in the area. One event starts/initiates a state - for example, a birth can initiate residence in the study area for the new born - and another ends/ terminates the state, for example, out-migration to a destination outside the study area terminates the state of being resident in the area. Episodes are used for pairing start and end events. Episodes are shown in the middle part of Figure 15. Core events are to the right of the diagram. Time thresholds are used to track episodes. An out-migration is only recorded if the absence lasts longer than an explicitly set threshold. The time thresholds for defining episodes differ from project to project ranging from 6 weeks to 3 months (INDEPTH Network 2002). Beyond residence, other episodes maybe of interest to an HDSS, for example, marital union, membership to a social group and so on. HDSS also record point events such as child birth (Benzler, Herbst, and Macleod 1998)- for example, delivery of a child. These are isolated in the sense of not bracketing episodes of interest. The model also includes observation recordings capturing the location, the date and the identity of the person making a recording. This helps to maintain good data quality.

It is within this HDSS framework that the HIV related repeated surveys and other studies used within ALPHA are performed. The next section provides an overview of the ALPHA network, it assumes the basic ideas provided in the current section.

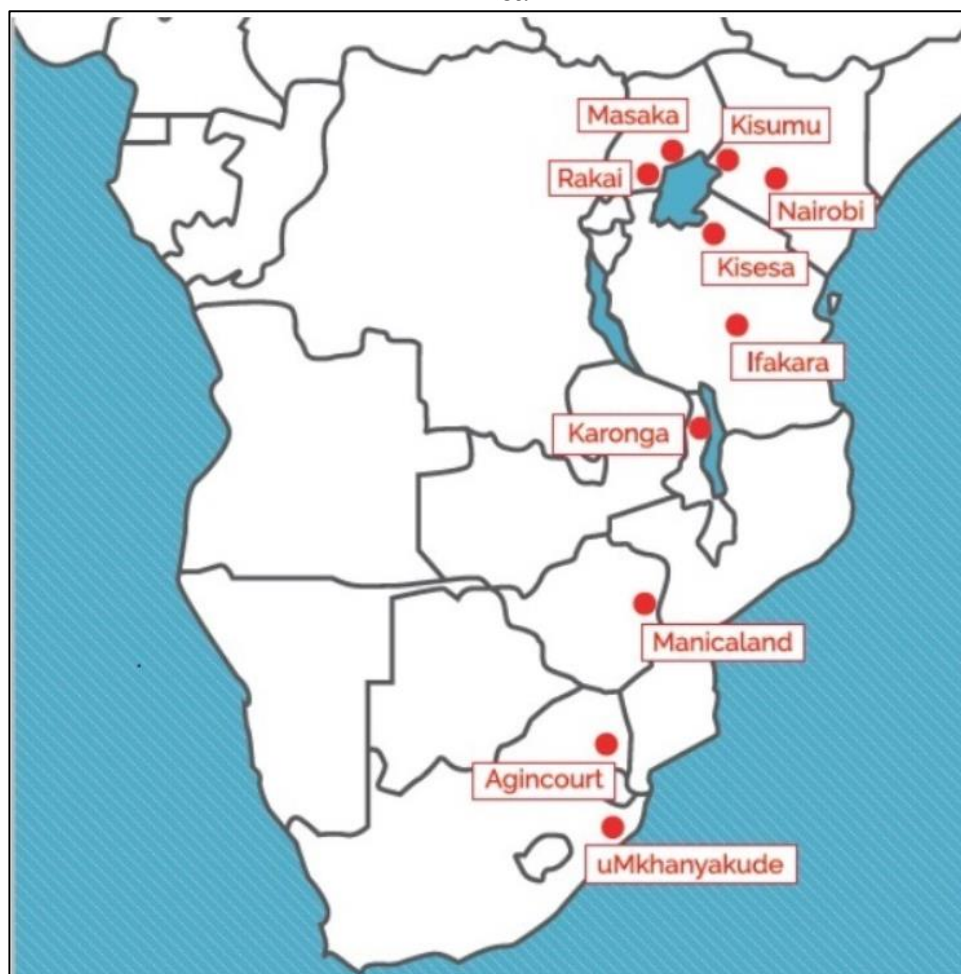
3.3 ALPHA network overview

The ALPHA network is adequately described in the two dedicated publications (Maher et al. 2010; Reniers et al. 2016) and also in many other ALPHA related publications listed here (<http://alpha.lshtm.ac.uk/publications/>). The listed publications cover various aspects of research themes characterising the network. This section however provides a brief overview concentrating on the details needed to understand the subsequent chapters of this thesis.

ALPHA represents an innovative research programme focussing on broadening the evidence base of HIV epidemiology for informing policy, strengthening analytical capacity for HIV research and fostering collaboration between network members (Reniers et al. 2016). It is a collaboration of ten autonomous health research institutions based in Eastern and Southern Sub Saharan Africa (Figure 16) and the London School of Hygiene and Tropical Medicine in the global north. The network members have published their individual study profiles

(Beguy et al. 2015; Kishamawe et al. 2015; Kahn et al. 2012; Geubbels et al. 2015; Odhiambo et al. 2012; Tanser et al. 2008; Crampin et al. 2012; Asiki et al. 2013; Gregson et al. 2017), these provide more details about each member. The network members share common interest in HIV epidemiology. Most ALPHA members are also members of the INDEPTH network of HDSS sites (<http://www.indepth-network.org/>). ALPHA was formed in 2005 (Maher et al. 2010), after a number of years of informal collaborations. Researchers in the department of population health at the London School of Hygiene and Tropical Medicine (LSHTM) coordinate the network.

Figure 16: Locations of ALPHA network member sites in eastern and southern Africa



Source: <http://alpha.lshtm.ac.uk/partner-study-institutions/>

A Scientific Advisory Board that includes researchers from UNAIDS, WHO and research leaders from the member institutions provides research oversight and direction.

The collaborating institutions were all established independently before the network came into existence, consequently, they were established for research aims that may be different

from those of ALPHA. Table 3 provides a basic description of the ALPHA network member sites. Reniers et al. (2016) provides more details on the network members.

Table 3: Selected Characteristics of ALPHA network member research centres

Short name	Year surveillance started	HIV % prevalence (Year)	Population size (year)	Name of study	Country	Year Joined ALPHA
Kisesa	1994	6 % (2011)	34000 (2011)	Magu Household Demographic Surveillance System managed by the TAZAMA programme at NIMR (Mwanza)	Tanzania	2005
Karonga	2002	8 % (2011)	35000 (2012)	Karonga Health and Demographic Surveillance System managed by Malawi Epidemiology and Intervention Research Unit (MEIRU), LSHTM	Malawi	2005
Masaka	1989	9 % (2011)	19000 (2011)	Kyamulibwa general population cohort managed by MRC/UVRI in Masaka district	Uganda	2005
Umkhanyakude	2000	29 % (2011)	96000 (2011)	Africa Centre Demographic Information System (ACDIS) managed by the Africa Centre in KwaZulu Natal	South Africa	2005
Nairobi	2002	*12 % (2007)	60000 (2007)	Nairobi Urban Health and Demographic Surveillance System operated by the African Population and Health Research Center http://aphrc.org/	Kenya	2010
Kisumu	2001	**15.4 (2004)	220000 (2012)	KEMRI/CDC Health and Demographic Surveillance System managed by Kenya Medical Research Institute and the Centers for Disease Control http://www.kemri.org/	Kenya	2010
Manicaland	1998	14 % (2008)	37000 (2008)	Manicaland HIV/STD Prevention Project managed by Biomedical Research and Training Institute (Harare), and Imperial College (London) http://www.manicalandhivproject.org/	Zimbabwe	2005
Agincourt	1992	19 % (2010-2011)	90000 (2011)	Agincourt Health and Demographic Surveillance Site, managed by University of the Witwatersrand http://www.agincourt.co.za/	South Africa	2010
Ifakara	1997	-	168000 (2007)	Ifakara Health and Demographic Surveillance System operated by the Ifakara Health Institute http://www.ihl.or.tz/	Tanzania	2010
Rakai	1989	11 % (2009)	40000 (2009)	Rakai Community Cohort Study, managed by Makerere University and Johns Hopkins School of Public Health	Uganda	2006

Columns 1, 5 and 7 (ALPHA 2013), Column 3 and 4 (Zaba et al. 2013a; J Madise et al. 2012; Gómez-Olivé et al. 2013; Odhiambo et al. 2012; Ifakara HDSS 2010), prevalence was mainly measured among adults aged 15 years and above

*Prevalence was for women aged 15-49 and men aged 15-54, **Prevalence was measured among study participants aged 13 – 34 years

The member institutions conduct longitudinal demographic surveillance of populations that range from 20,000 to 220,000 individuals (Reniers et al. 2016). They also conduct population-based surveys with HIV testing and verbal autopsies with relatives of the deceased to identify probable causes of death. ALPHA has significantly contributed to the monitoring of population-based estimates of HIV-associated mortality over the period of its existence. It has also provided estimates relating to the population level effect of antiretroviral therapy scale up and uptake of HIV diagnostic and AIDS care services. ALPHA provides estimates of the survival of people living with HIV without treatment. These estimates are used as inputs in the UNAIDS spectrum model (www.epidem.org) which generates global estimates of the epidemic.

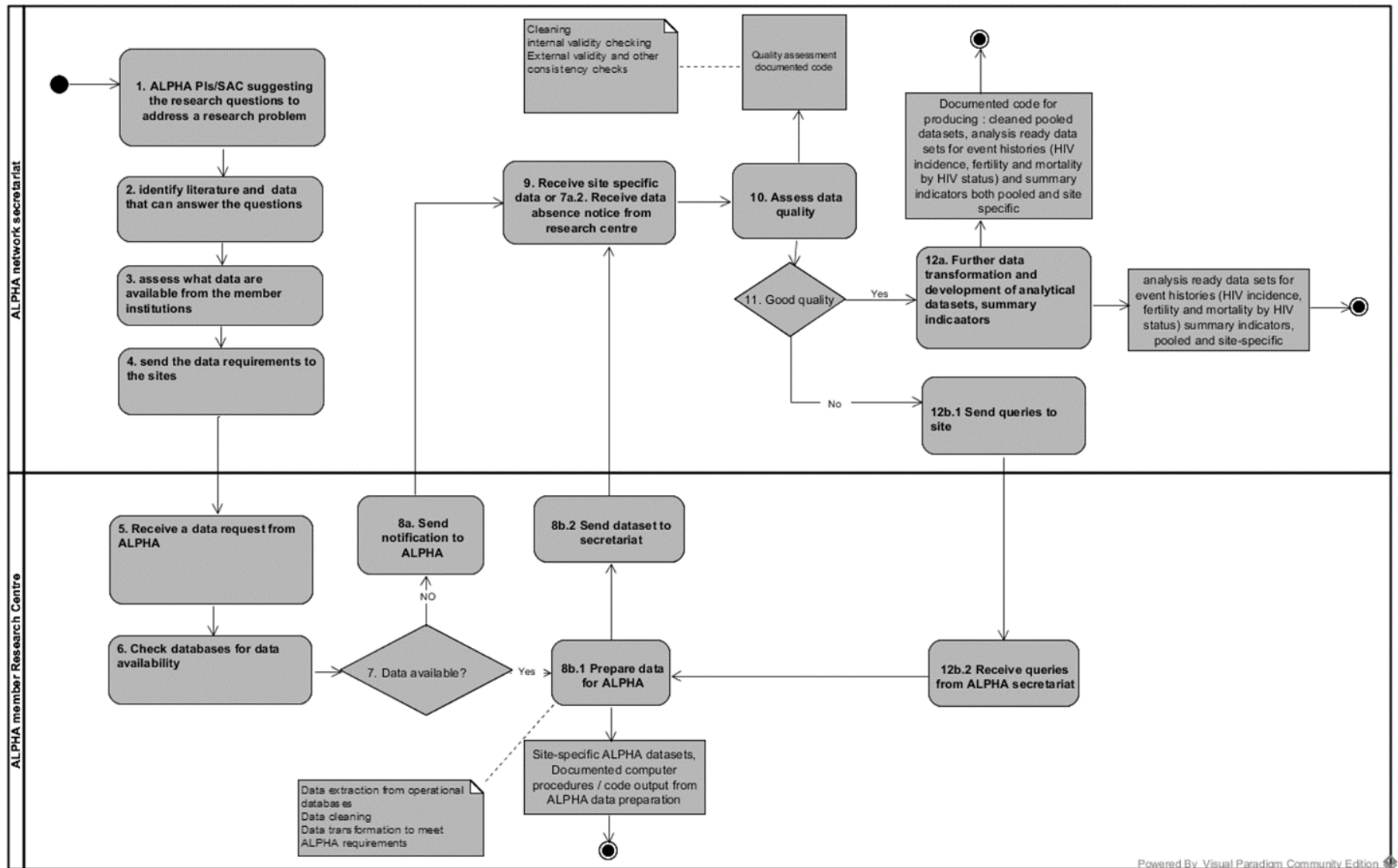
Work within the network is organised around thematic workshops. In the workshops, member institution researchers are taught aspects of data analysis for addressing research questions at hand. This teaching forms the basis for ALPHA site-specific and pooled analyses. One relatively new area of work in ALPHA is the linkage of records from the HDSS to the data from health facilities serving the population covered in the studies. This linkage takes deterministic (using unique identifiers) and probabilistic approaches such as those developed for the Agincourt study (Kabudula et al. 2014).

Until the Biomedical Resource grant awarded to ALPHA by the Wellcome Trust (Grant number 202917) in 2016, ALPHA had not had any funding for data management, only for specific topics of analysis. Thus, for the harmonised and pooled datasets, the network relied on the data management support provided by researchers based at the secretariat who already had data analysis and research findings publishing responsibilities.

3.4 ALPHA Data Production environment and processes

ALPHA does not collect primary data, rather it transforms data collected within its members for secondary analysis. Figure 17 provides an overview on how the individual member studies interact with the secretariat from the time that data are requested up to the time when the data are ready for analysis. The flow chart depicts two major phases of data processing. The first phase (steps 1 – 9) involves preparation of ALPHA harmonised datasets also known as ALPHA specifications from the individual members' source data. The second phase (step 10 onwards) develops analysis-ready files from the ALPHA specifications and is performed centrally by the secretariat at LSHTM. The latter phase of processing has frequently been the subject of the methods sections of ALPHA publications.

Figure 17: ALPHA data management overview



In addition, Slaymaker et al. (2017) provides a detailed description of the preparation of an exemplar ALPHA analysis ready HIV status and care continuum dataset.

This thesis focuses on the first phase and provides a summary of how the data are processed in this first phase. Each study contributes data to ALPHA in uniform format according to a prescribed structure. This is achieved by carrying out an ex-post or “after the fact” harmonisation of data from their original form into the form required by ALPHA.

The ALPHA specifications are developed to provide data that are suitable for answering specific demographic and HIV/ AIDS related research questions (Maher et al. 2010).

A full listing of the data that ALPHA has received from its members is found at <http://alpha.lshtm.ac.uk/metadata/>. Table 4 gives brief descriptions of the main ALPHA data specifications.

Table 4: ALPHA datasets

ALPHA data specification	Description
6.1: Residency episodes	Information on residence in the study area, including dates of birth, migration and death.
6.2b: HIV test data	History of HIV testing, including dates of tests, circumstance in which test was carried out, final test result, and whether or not the test result was returned to the participant.
7.1 Mother and father identifiers	contains the parental information whenever possible for all children/adolescents and also any parental links for adults if you have them
7.2 Reported births	contains for females only, one record for each birth reported by all mothers in DSS
7.4 Survey information	contains background characteristics for each individual (Men and women of all ages)
8: Verbal autopsy data	Verbal autopsy data
9.1 Self-reported data on HTC and ART use	Self-reported information, from periodic surveys, on use of HIV testing services, disclosure of HIV status, use of HIV care and treatment services, ART use and interruption of ART.
9.2: Clinic data	HIV clinic records on enrolment in care and ART history.
10.1: Sexual behaviour data	Sexual behaviour data
11.1: NCD Data	Longitudinal population-based African non-communicable diseases data on the burden, distribution and progression of the NCDs
SES Household	Household Socio-economic data

Sources (Slaymaker et al. 2017; Reniers et al. 2016), <http://alpha.lshtm.ac.uk/metadata/>, Dr Alison Price, Private communication at ANDLA workshops - Malawi, 2018 and Johannesburg, 2019

The gathered data include data on residency episodes, parent-child links, background characteristics of study participants, verbal autopsies for determining probable cause of death and use of HIV care and treatment services (from clinic data and self-reported).

Traditionally, the harmonisation processes were done using various versions of Stata between version 8 (StataCorp 2003) and version 19 (StataCorp 2019). Due to the complexity of the ALPHA data and their harmonisation processes, the use of these data by third parties has been limited as any external user interested in analysing them has had to work closely with

an ALPHA researcher to explain the intricacies of the data. Staff turnovers have also posed challenges relating to reproducibility of the data transformations over time. Often when there was a change in the staff responsible for ALPHA data preparation within a partner institution, the new staff could not follow their predecessor's Stata do-files leading them to starting all over thus consuming time and negatively impacting consistency of the data supplied to the network. ALPHA is currently working towards better management of the data processing and improvement of transparency with regards to the traceability of variables. This is being done through the use of Extract-Transform and Load (ETL) software called Pentaho Data Integration (PDI), a component of the Pentaho suite of software products (Pentaho Corporation 2018). The use of Pentaho in ALPHA is being done within the Centre in a Box (CiB) environment. The CiB is a self-contained and controlled data management and curation system (Herbst et al. 2015). This transition to the use of Pentaho from Stata is a follow up to the highly successful INDEPTH data management programme and the INDEPTH Data Repository (Herbst et al. 2015).

3.4.1 Centre in a Box major components

Herbst et al. (2015) describes the CiB environment fully. In brief, it comprises of a portable mini-server hardware which hosts a hypervisor. The hypervisor supports three virtual servers:

The first virtual server is a database server which host the database management system used by a member institution and replicates the institution's operational database. This facilitates the transfer of data to the analytical dataset production environment. The second virtual server is a data manager's desktop which hosts the PDI and Nesstar Publisher software applications. These applications are used for the preparation and documentation of the datasets. The third virtual operating environment is the system server. This server manages the CiB environment in terms of security, shared file system and a web server. The web server implements a local instance of the World Bank developed data cataloguing tool called National Data archive (NADA). This local instance serves the purpose of reviewing metadata before they are published on the data repository.

3.4.2 Pentaho ETL processes in ALPHA

Pentaho Data Integration information model

The hub of activity in the CiB happens on the data manager's desktop within PDI. Pentaho provides a graphical extract-transform-load (ETL) designer to simplify the creation of ETLs.

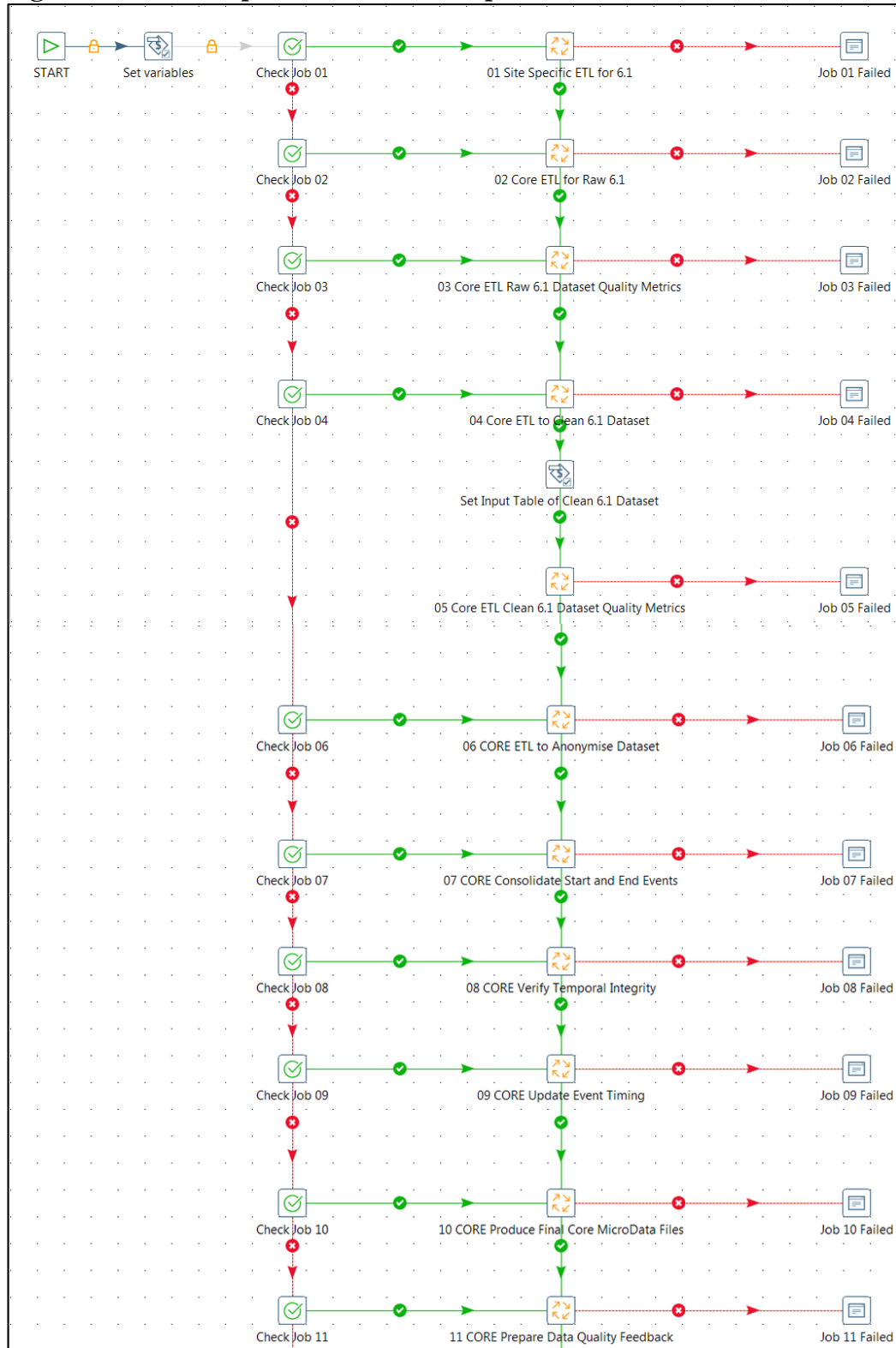
It has a rich library of pre-built components to access, prepare, and blend data from various sources (Pentaho Corporation 2018). It has powerful orchestration capabilities to coordinate and combine transformations. The Pentaho information model comprises of two levels at which ETL can be implemented; these are the job and the transformation levels. The transformations represent the lower level which decomposes the transformation into a sequence of Pentaho steps. The steps are connected via connections or, again, “arcs” that Pentaho calls hops. The Job level is the higher level. It shows the sequence in which transformations are carried out. Though users can take advantage of this provision to perform ETL tasks at two levels, Pentaho is flexible in that the user is not required to do so, an ETL can comprise of jobs only or transformations only or a mixture of jobs and transformations depending on the task at hand. PDI generates an XML document which records details of the ETL process as it gets developed.

ETL processes

The ALPHA data transformations done in Pentaho are organised as master jobs with sub-jobs within them. Each ALPHA specification has a PDI master job associated with it. This master job comprises of a list of sub-jobs performed in sequence. Each of the sub-jobs, in turn, comprises of transformations and other operations to accomplish particular tasks which when taken together, produce the ALPHA dataset. In this thesis, I have used the ETL for one particular ALPHA data specification, ALPHA 6.1 data specification (ALPHA spec 6.1) for illustration. The same methods and ideas also apply to for the documentation of all the other data specifications.

Figure 18 shows a picture of the entire ETL for creating ALPHA spec 6.1 from member studies’ source data. This dataset assembles data from demographic surveillance relating to residence episodes within the study area (Reniers et al. 2016; Slaymaker et al. 2017). The episodes of residency are bounded by starting events such as birth, in-migration, becoming eligible and terminating events such as death, out-migration. There is one master job for specification 6.1 for each of the member institutions. Within this master job, there is a member institution specific sub-job for transforming source data into a common intermediate format, the first sub-job –represented by the first square with orange arrows inside it in Figure 18 labelled “01 Site specific ETL for 6.1”. The rest of the ETL comprise of common sub-jobs used to process this intermediate data into the ALPHA data specification, these are the sub-jobs from “02 core for Raw 6.1” to “11 Prepare Data quality feedback”. The sub-jobs are sequentially executed.

Figure 18: ALPHA Specification 6.1 ETL process in Pentaho



Source (ALPHA spec 6.1 ETL for Nairobi HDSS provided by Tathagata Bhattacharjee)

Within each of the sub-jobs, there are Pentaho transformations and other entries such as sql scripts for creating tables. Drawing a crude analogy with Stata, a Pentaho transformation is

comparable to a do-file in the Stata package. Each of the transformations comprise of steps which are an equivalent of a command or function in Stata.

It should be noted that what the diagram in Figure 18 is displaying are the labels of the sub-jobs. The correspondences between these labels and the underlying sub-job names (used in Chapters 5 and 6) are given in Table 5.

Table 5: Sub-job Pentaho names and their labels

Sub-job name	Sub-job label
00 Generating Staging Tables	01 Site Specific ETL for 6.1
001 CORE Produce Raw 6.1 Dataset	02 Core ETL for Raw 6.1
002 CORE Data Quality Metrics	03 Core ETL Raw 6.1 Dataset Quality Metrics
003 CORE Data Cleaning	04 Core ETL to Clean 6.1 Dataset
002 CORE Data Quality Metrics	05 Core ETL Clean 6.1 Dataset Quality Metrics
004 CORE Data Anonymisation	06 CORE ETL to Anonymise Dataset
005 CORE Consolidate Start and End Events	07 CORE Consolidate Start and End Events
006 CORE Verify Temporal Integrity	08 CORE Verify Temporal Integrity
007 CORE Update Event Timing	09 CORE Update Event Timing
008 CORE Produce Final Core MicroData Files	10 CORE Produce Final Core MicroData Files
009 CORE Prepare Data Quality Feedback	11 CORE Prepare Data Quality Feedback

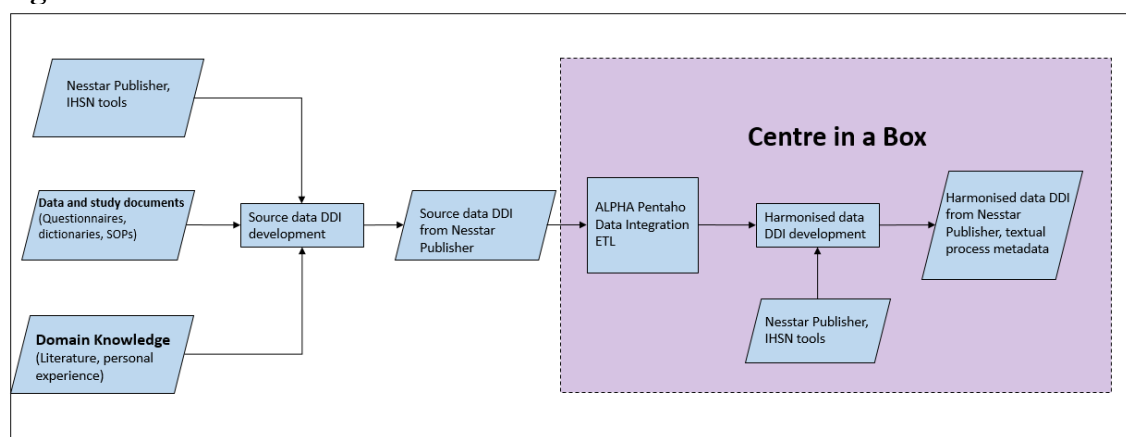
3.4.3 CiB Metadata management overview

Regarding documentation of the process prior to sharing data with the ALPHA secretariat, three major aspects come into play: (1) the source data feeding into the ALPHA ETLs, (2) the transformation processes and (3) the harmonised data specifications.

(1) Documentation of Source data:

The CiB does not have a formal way of integrating into its operations the structured metadata created by individual studies for their primary data. The majority of ALPHA members are producing DDI Codebook using Nesstar Publisher. Figure 19 shows how these metadata could be harnessed within the current CiB environment. Umkhanyakude, Kisesa, Karonga, Nairobi and Ifakara are some of the studies producing Codebook documentation.

Figure 19: Structured metadata catered for in the current CiB



In Chapter 4, the implementation of metadata standards within a typical HDSS is analysed with the aim of providing a recipe for an HDSS planning to start implementing metadata standards. This is of importance to ALPHA members since there is not such guide available. It is also of benefit to the wider HDSS community.

(2) Documentation of data transformations:

The CiB currently relies on Pentaho's proprietary metadata for process documentation. Pentaho is partially self-documenting. Its graphical interface provides diagrammatic expressions of the changes happening to data as the transformations progress. In addition, provisions are made for free-text descriptions of each step/ task. These metadata are stored in a Pentaho XML document. Annotated diagrams of the transformations in PDF format can also be generated in Pentaho. While comprehensive, these metadata are proprietary and not compliant with the existing metadata standards. They also tend to be deeply buried in other technical metadata used by the platform for its internal integrity and operations. In Chapters 5 and 6, frameworks and approaches are proposed for creating software agnostic and structured metadata for these transformations. They leverage the Pentaho information model and harness and extend the existing metadata standards and process models.

(3) Documentation of the harmonised datasets

The harmonised data are documented using Nesstar Publisher within CiB.

Once a specification has been created it can be imported into Nesstar Publisher. Figure 20 shows the various stages involved in importing a file into Nesstar. As shown at the bottom right corner of the diagram, files of various formats can be imported. If the imported file is in Stata or SPSS format, Nesstar Publisher automatically captures the variable names, the labels, the value codes and their labels. Beyond the automatically identified metadata, a Nesstar Publisher user is also able to manually capture metadata via the various sections of the interface. These include the description of the study to which the data file belongs, the bibliographical information for the DDI file (author, date, organisation and so on), the dataset as a whole and the variable level details. Figure 21 shows the variable level metadata provisions in Nesstar Publisher. These include variable definitions, universes, if it's a derived variable - what are the derivation instructions and, question texts and related instructions including skip instructions.

Also depicted by Figure 21 is the general look and feel of the Nesstar Publisher interface. It is widely regarded to be user-friendly and intuitive, partly owing to the simple structure of DDI Codebook.

Figure 20: Various stages of importing a file into Nesstar Publisher

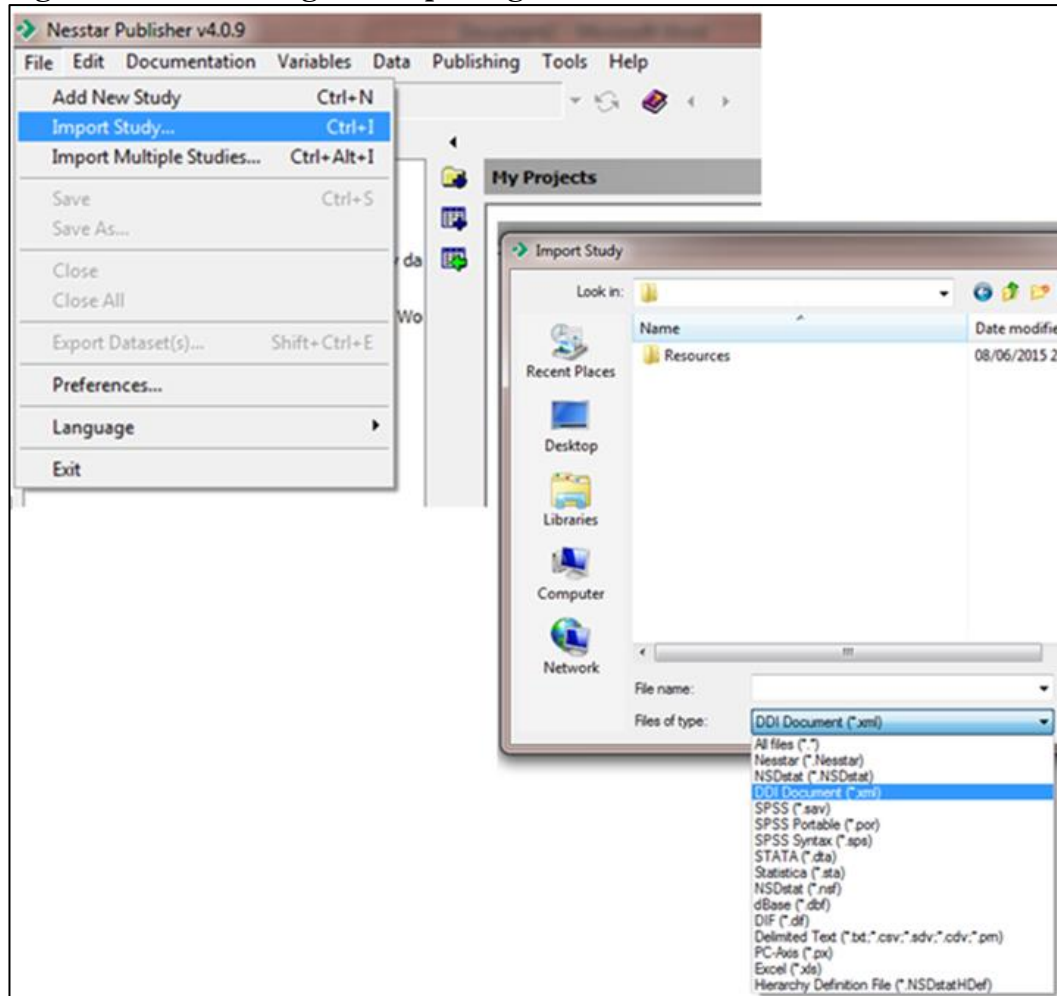
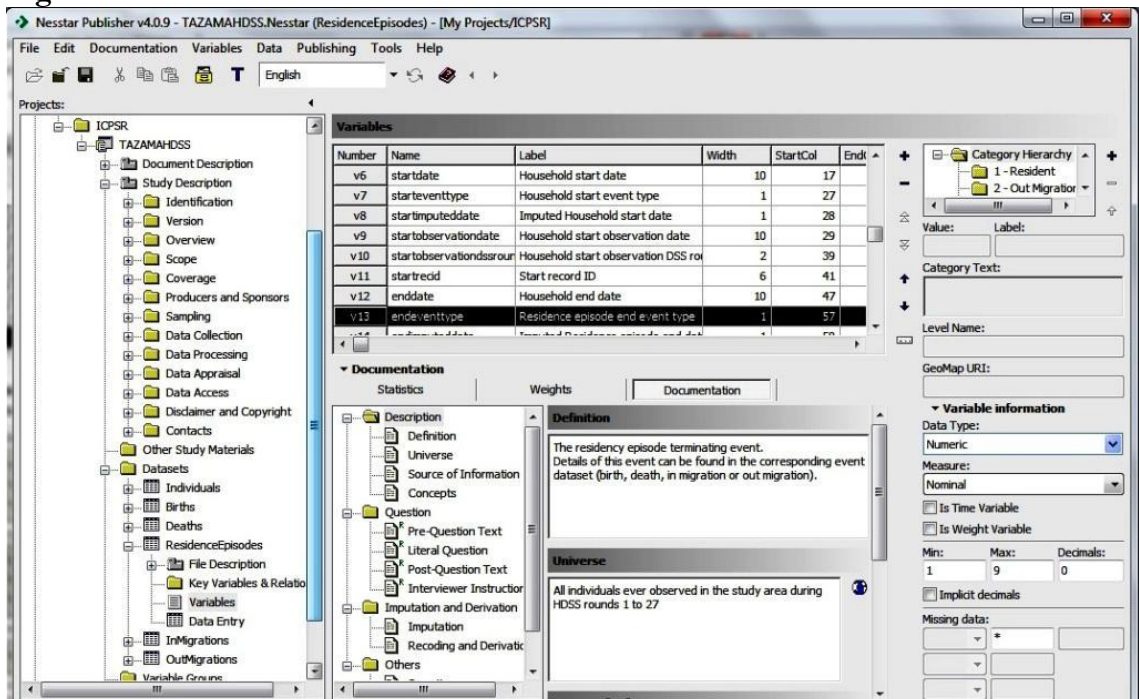


Figure 21: Variable level metadata in Nesstar Publisher



The result of the documentation process are a Nesstar file comprising the data and the enhanced documentation. Nesstar can export data in the various formats shown in Figure 20. It can export a human intelligible data codebook in PDF format. In addition, it can export a DDI Codebook in XML format. This DDI file and the data files are then imported into the NADA data repository for dissemination.

In Chapter 7, requirements for a user-friendly provenance metadata browser for ALPHA are analysed. This browser will need to integrate and present the source variables metadata, data transformations metadata and the documentation of the harmonised data for browsing and searching.

3.5 Chapter summary

This chapter has provided a description of the settings within which the study is being carried out. First, it looked at HDSS basics and reference data model. It has also provided a general overview of the ALPHA network and its data processing procedures including the current ongoing migration to Pentaho for ETLs. Lastly, it has considered the metadata currently available to annotate the data sources, the data processing and the resulting harmonised data products. It showed that while DDI Codebook metadata are available for the harmonised data via the CiB, the metadata for the source data are not actively incorporated and that the process metadata are predominantly ETL-tool specific and not compliant with the recommended international metadata standards.

A brief outline of the roadmap followed in this thesis from Chapter 4 to 7 is also given to point out how the thesis seeks to enhance the CiB metadata system.

Registry

T: +44(0)20 7299 4646
F: +44(0)20 7299 4656
E: registry@lshtm.ac.uk

RESEARCH PAPER COVER SHEET

PLEASE NOTE THAT A COVER SHEET MUST BE COMPLETED FOR EACH RESEARCH PAPER INCLUDED IN A THESIS.

SECTION A – Student Details

Student	Chifundo Kanjala
Principal Supervisor	Professor Jim Todd
Thesis Title	Provenance of “after the fact” harmonised community-based demographic and HIV surveillance data from ALPHA cohorts

If the Research Paper has previously been published please complete Section B, if not please move to Section C

SECTION B – Paper already published

Where was the work published?	IASSIST Quarterly		
When was the work published?	2016		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

**If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.*

The Creative Commons-Attribution-Noncommercial License 4.0 International applies to all works published by IASSIST Quarterly. Authors will retain copyright of the work.

SECTION C – Prepared for publication, but not yet published - N/A

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--

Student Signature: _____ **Date:** 20/08/2019

Supervisor Signature: _____ **Date:** 20/08/2019

4. OPEN-ACCESS FOR EXISTING LMIC DEMOGRAPHIC SURVEILLANCE DATA USING DDI

4.1 Abstract

The Data Documentation Initiative (DDI) specification has gone through significant development in recent years. Most Health and Demographic Surveillance System (HDSS) researchers in Low and Middle Income Countries (LMIC) are, however, unclear on how to apply it to their work. This paper sets out considerations that LMIC HDSS researchers need to make regarding DDI use. We use the Kisesa HDSS in Mwanza Tanzania as a prototype. First, we mapped the Kisesa HDSS data production process to the Generic Longitudinal Business Process Model (GLBPM). Next, we used existing GLBPM to DDI mapping to guide us on the DDI elements to use. We then explored implementation of DDI using the tools Nesstar Publisher for the DDI Codebook version and Colectica Designer for the DDI Lifecycle version.

We found the amounts of metadata entry comparable between Nesstar Publisher and Colectica Designer when documenting a study from scratch. The majority of metadata had to be entered manually. Automatically extracted metadata amounted to at most 48% in Nesstar Publisher and 33% in Colectica Designer. We found Colectica Designer to have stiffer staff training needs and software costs than Nesstar Publisher.

Our study shows that, at least for HDSS in LMIC, it is unlikely to be the amount of metadata entry that determines the choice between DDI Codebook and DDI Lifecycle but rather staff training needs and software costs. LMIC HDSS studies would need to invest in extensive staff training to directly start with DDI Lifecycle or they could start with DDI Codebook and move to DDI Lifecycle later.

4.2 Keywords

HDSS, open-access, metadata, DDI Codebook, DDI Lifecycle

4.3 Introduction

Investigators of HDSS studies in LMIC are realising the importance of preparing their existing data for open access. These data have been used to produce some of the key results leading to better understanding of HIV/AIDS among other diseases (Ghys, Zaba, and Prins 2007; Hallett et al. 2008; Porter and Zaba 2004; Todd et al. 2007; Zaba et al. 2013b; Ndirangu

et al. 2011; Streatfield et al. 2014; Desai et al. 2014). They have been used to shed light on sub-Saharan Africa mortality patterns (INDEPTH Network 2002; Sankoh et al. 2014). Providing open access will increase accessibility of these data to regional trainee scientists and the wider research community and thus maximise their public health benefit.

Human science research data documentation has gone through considerable methodological advances in recent years. One of these advances is the development of the Data Documentation Initiative (DDI), a specification that is commonly used for documenting observational survey data (Rasmussen and Blank 2007b; Wellcome Trust 2014). It uses the eXtensible Markup Language (XML) format (W3schools.com 2015) and has two main strands: DDI Codebook, originally called DDI 2, and DDI Lifecycle, originally called DDI 3. DDI Codebook is the simpler of the two and aims to describe a dataset in terms of its structure, contents and layout – a compilation of facts about a dataset mainly for archiving purposes. It has been used worldwide including in LMIC through the International Household Survey Network (IHSN) and the World Bank (International Household Survey Network 2013). The IHSN implementation of DDI Codebook was done using DDI-compliant software for metadata management called Nesstar Publisher (Digital Curation Centre 2013). Once data have been documented in Nesstar Publisher, the resulting documentation can be presented in various forms including PDF versions of the codebook and cataloguing of the data in web-based catalogues. A commercial data repository and catalogue created by Nesstar called Nesstar Server could be used. Alternatively open source software called National Data Archive can also catalogue data and DDI-compliant metadata. NADA was designed by the World Bank and the IHSN to facilitate archiving and sharing their national data (International Household Survey Network 2016).

DDI Lifecycle was developed from the premise that a dataset is an embodiment of a process that produced it, thus, it uses the data life cycle (Figure 3) as its conceptual model. It comprises modules which are packages of metadata each roughly corresponding to a stage in the data life cycle. There is one related to study conceptualisation, another related to data collection, another catering for archiving and so on. DDI Codebook metadata are still present in DDI Lifecycle and are spread throughout its modular structure. It also captures metadata that describe associations between groups of studies. A number of tools for implementing DDI Lifecycle are available. These include Colectica Designer, Questasy (de Bruijne and Amin 2009; de Vet 2013), DDI on Rails (Hebing 2015a), DDA DDI Editor (Jensen 2012) among others produced at the Gesis Leibniz Institute for Social Sciences in Germany (<http://www.gesis.org/en/institute/>) and the North American Metadata

Technology (<http://www.mtna.us/>). We used Colectica Designer because when we started the documentation work it was one of the few available DDI Lifecycle tools offering the most flexibility to meet our needs.

Closely related to the DDI Lifecycle is the Generic Longitudinal Business Process Model (GLBPM), which outlines steps taken in the process of producing longitudinal data for social and human sciences. The GLBPM is shown in Figure 7. Mapping an organisation's data production process to the GLBPM can determine what metadata to record at each step of data production since GLBPM has been mapped to DDI Lifecycle (I. Barkow, Block, Greenfield, Gregory, Hebing, Hoyle, and Zenk-möltgen 2013).

The LMIC HDSS studies have generally used metadata standards at the research network level as shown by the example of the INDEPTH Network data repository (INDEPTH Network 2013a). To the best of our knowledge, only a few individual HDSS studies, among them, the Africa Centre for Population Health (Africa Centre for Population Health 2015) and African Population and Health Research Center (African Population and Health Research Center 2015) have used DDI. For sites not using DDI, this has led to the documentation of a small subset of all the data that the studies generate, in many cases, less than 20% of the variables on which a typical HDSS collects data. This means that the strengths and limitations of the data are not properly understood by secondary users, making it hard for them to interpret their analyses.

To demonstrate the use of the DDI metadata standard to document 'legacy' data, we applied it to the existing Kisesa HDSS data. This task required consideration of the metadata editors to use, the amount of documentation needed when using DDI Codebook and DDI Lifecycle, staff training needs and approximate software costs.

4.4 Study settings and methods

4.4.1 Study settings

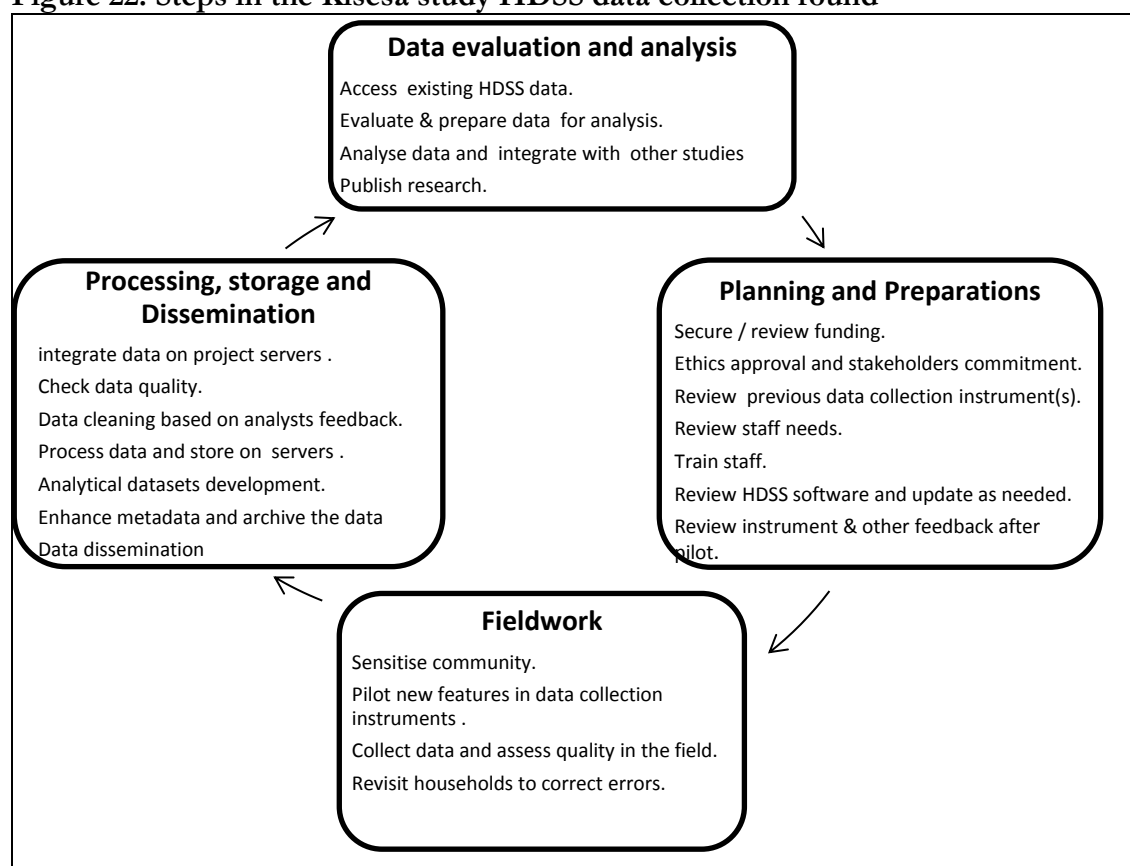
The TAZAMA project within the National Institute for Medical Research, Mwanza Tanzania runs the Kisesa open cohort study. It has been described in detail previously (Marston et al. 2012; Kishamawe et al. 2015; Urassa et al. 2001). The backbone of the Kisesa study is its HDSS. The population in the study area had grown to over 35,000 by 2014 (Kishamawe et al. 2015) from about 19,000 in 1994. Follow-up data collection rounds have been done at roughly six-month intervals recording new births, migrations and deaths. In addition, marriages, pregnancies and education are recorded. Paper questionnaires were used for data collection until round 25. Since round 26, HDSS data are collected electronically using

Portable Digital Assistants and CSPro applications. While the Kisesa study runs other studies including cause of death analysis and HIV serological studies, we focus on describing the documentation of the HDSS, which provides the sampling frame for all the nested TAZAMA studies. Once the HDSS documentation is understood, it will be easier to apply the principles to the studies that rely on the HDSS. The HDSS component is implemented in broadly similar ways across a range of studies (Sankoh and Byass 2012) so such studies can relate to the Kisesa experiences.

4.4.2 Study methods

The data production process involved in the implementation of a typical HDSS data collection round in Kisesa is illustrated in Figure 22.

Figure 22: Steps in the Kisesa study HDSS data collection round



At the top is the data evaluation and analysis phase prior to an HDSS round. Going clockwise, we have the planning and preparation phase followed by activities related to fieldwork, while the last box shows the steps related to office data processing, storage and dissemination. Each step was mapped to its closest equivalent within the GLBPM (I. Barkow, Block, Greenfield, Gregory, Hebing, Hoyle, and Zenk-möltgen 2013). We then used the

existing mapping from GLBPM to DDI Lifecycle (I. Barkow, Block, Greenfield, Gregory, Hebing, Hoyle, and Zenk-möltgen 2013) to guide us on the likely DDI metadata elements to use for documenting HDSS data.

Once the mapping exercise was completed, we used Nesstar Publisher to produce DDI Codebook and Colectica Designer for DDI Lifecycle. For Nesstar Publisher, we used the IHSN metadata template and the step-by-step guide (Dupriez and Greenwell 2007), while for Colectica Designer we used the information model provided with the Colectica online documentation (Colectica 2015b). The actual documentation was done in three overlapping phases: preparation, data documentation, and creation of an internal data catalogue.

In the preparation phase, we piloted the use of Colectica Designer and Nesstar Publisher. In Colectica Designer, we created an HDSS series as a group within which all the HDSS data from the numerous data collection rounds could be documented. For rounds 26 and 27, a study metadata package was created, using guidance provided by the Colectica user's guide (Colectica 2015a). We gathered and entered foundational metadata including concepts, affiliated organisations and universes for variables, and added metadata pertaining to study-level, data collection, data processing, dataset and variables. This pilot showed that the levels of training and finances required to do this work using locally recruited staff were not sustainably available for the project. On the other hand, DDI Codebook seemed accessible from both our pilot work and examples from other studies (INDEPTH Network 2013b), and its use was agreed. Two recent graduates from quantitative backgrounds were recruited and trained in the use of Nesstar Publisher – this initial training took two weeks. Data in the project's databases that required documentation was identified and relevant details - lists of the database tables and locations of the databases on the project's servers - recorded.

We started the documentation phase by importing the data into Nesstar Publisher where additional metadata were added. Metadata not available in the data files were extracted from questionnaires, ethical clearance documents, funding proposals and other supporting documents and entered manually in Nesstar Publisher. After documentation, we went on to catalogue the data.

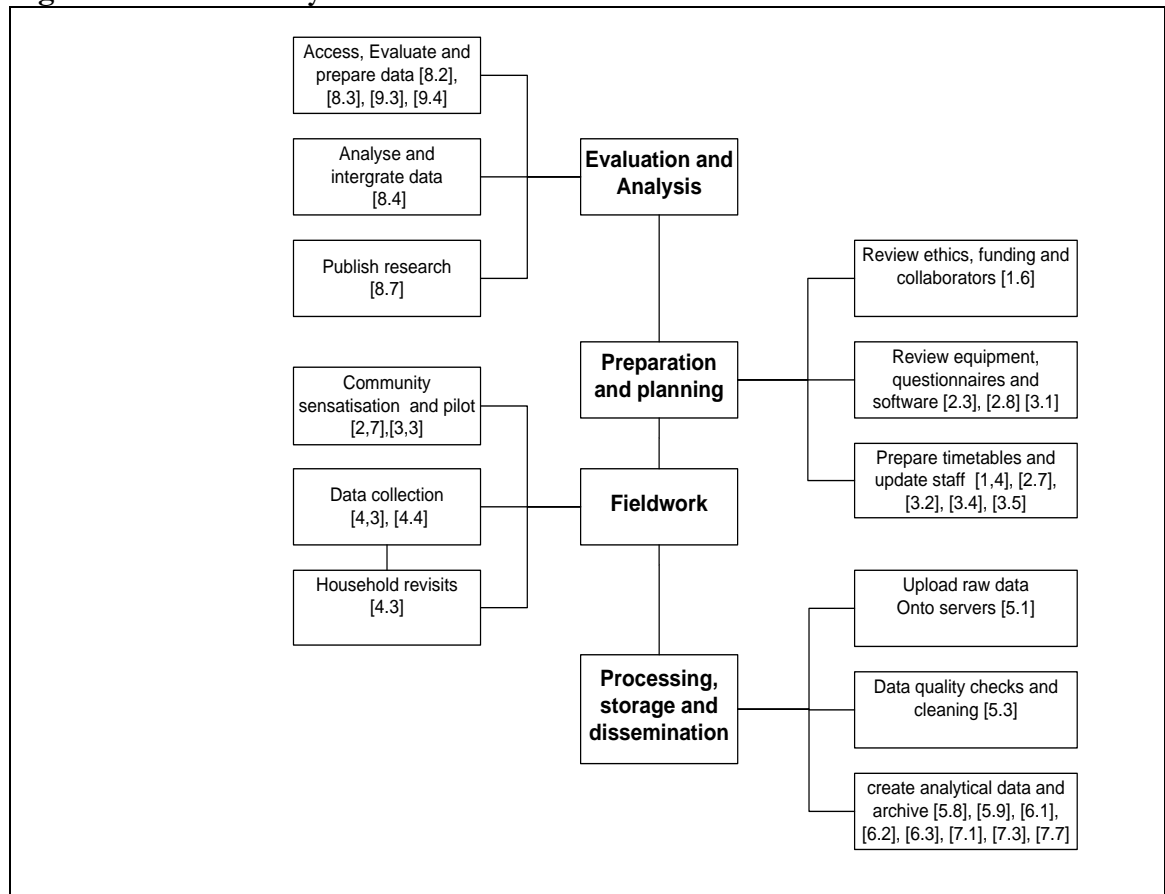
Finally, the metadata, data, supporting documents and publications based on the data were brought together into the data catalogue. The DDI Codebook files were transferred from Nesstar Publisher to NADA and we subsequently configured NADA to suit our needs. The design of the catalogue provides for demarcation of collections of the data and their associated documentation – in this case we created a collection dedicated to Kisesa HDSS data.

4.5 Results

4.5.1 Mapping Kisesa study HDSS data production to GLBPM

The results of mapping one round of the Kisesa study HDSS data production process onto the GLBPM are presented in Figure 23. The GLBPM steps are shown in square brackets.

Figure 23: Kisesa study HDSS Documentation in Nesstar Publisher



The activities within the *Evaluation and analysis* phase corresponded to the two GLBPM steps Research / publish (8) and Retrospective Evaluation (9). The *Preparation and Planning* phase corresponded to the GLBPM's first 3 steps which are Evaluate / specify the needs (1), Design / redesign (2) and Build / rebuild (3). The *Fieldwork* phase corresponded to the Design/ redesign (2), the Build / rebuild (3) and the Collect steps (4). The *Processing, Storage and dissemination* phase corresponded to the Process / analyse (5), the Archive / preserve and Curate (6) and the Data Dissemination/Discovery (7) steps. Data dissemination is done via the data catalogue at the project offices and through correspondence with the project head for remote access.

In line with the two properties of the GLBPM that it is not exhaustive and non-linear, not all sub-steps were used in the mapping. Of the 53 sub-steps in GLBPM, 28 were found to be relevant to the Kisesa HDSS. The excluded sub-steps fell into 3 broad categories: those that were not supported within the Kisesa data management system, those that were not applicable to the HDSS round under consideration and those that do not apply to the HDSS type of studies. The examples of sub-steps not currently supported include 5.4 – imputing missing data, 7.5 – support for data citation, and 7.6 enhance data discovery, among others. Most of the sub-steps in step 1 mainly applied to the initial census and first follow-up round of the HDSS and were not frequently revisited in the subsequent rounds of the study. Since the HDSS involves the entire population within a geographically demarcated area, it does not apply any sampling so the sampling and weighting sub-steps are not applicable.

4.5.2 Implementation of DDI in Nesstar Publisher and Colectica Designer

Table 6 contains counts of some of the main items involved in the documentation of Kisesa study HDSS data. This gives an idea of the scale of the documentation involved.

Table 6: Counts of items involved in the documentation of Kisesa HDSS

Item	Quantity
HDSS data collection rounds	27
Data files	38
Questionnaires	27
Computer Assisted Interviews	2
Paper Questionnaires	25
Variables ¹	1216

We had completed data documentation for 27 HDSS rounds at the time of writing. Starting from the baseline round to round 20, there is one data file per round. Rounds 21 onwards have either two or three data files for each round, with one file holding household-level data and the other holding individual household members' data. In round 26, the questionnaire comprises a hierarchical set of 36 household-level questions and 53 individual-level questions, generating 41 and 67 variables respectively in the household and individual data files, including derived and administrative variables. In round 27 there are 54 household-level questions and 104 individual level questions, generating 62 and 118 variables respectively. In developing the metadata repository, we extracted data from MS Access and SQL Server databases into Stata 12. Within Stata, we added notes, variable and value labels as needed.

¹ In this case we are counting the instance of each variable within a data collection round as a distinct variable even though many of the variables remain unchanged across data collection rounds

The resulting Stata files were then imported into Nesstar Publisher and Colectica Designer (only two rounds, 26 and 27 for pilot). Some metadata were automatically extracted from the Stata files: categories, codes, variable names and labels, data file and variable notes. We compiled counts of the metadata items we considered to be important for HDSS. The results are given in Table 7.

Table 7: Nesstar Publisher (NP) and Colectica Designer (Colectica) documentation

	Round 026		Round 027	
	NP	Colectica	NP	Colectica
Foundational metadata				
Universes	108	15	180	2 (15 referenced from round 26)
Categories	41	32	67	8 (32 referenced from round 26)
Codelists	41	32	67	8 (32 referenced from round 26)
Concepts	14	14	16	2 (14 referenced from round 26)
Organisations	12	12	12	All 12 referenced from round 26
Automatically entered	81	62	134	16
Study-level metadata				
HDSS studies Group	-	53	-	
Each HDSS Round	41	49	41	37 (12 referenced from round 26)
Automatically entered	-	-	-	-
Data Collection metadata				
Methodology	4	5	4	5
Instrument	11	1237	11	331 (1233 referenced from round 26)
Collection events	5	8	5	7 (1 referenced from round 26)
Data Processing	Attach batch edit programs as external resources/ other materials			
Automatically entered	-	-	-	-
Datasets metadata				
Dataset	20	20	20	20
Variables	2808	2160	4680	2090 (1512 referenced from round 26)
Automatically entered	1404	432	2054	720
Total automatically entered items	1485	494	2188	736
Total number of metadata items	3105	3637	5103	2212

The four broad categories into which we classified the metadata are foundational, study level, data collection-related and datasets metadata. Universes were identified both at study and variable levels. In Nesstar Publisher, even in cases where a number of variables shared the same universe, that universe had to be entered for each variable due to lack of mechanisms for reuse of metadata. This is a limitation of DDI Codebook not of Nesstar Publisher. In contrast, in Colectica Designer, we entered each unique universe once and referred to that universe each time it applied, which explains why there are many more universes in Nesstar Publisher than in Colectica Designer. The ability to reuse metadata across studies meant we only needed 2 additional universes during documentation of round 27, since most of them had been entered in round 26. Reuse of metadata also led to the reduction in categories, codes, concepts and organisations that needed to be entered for round 27 for Colectica Designer. Categories and codes were automatically extracted from Stata files but the concepts, universes and organisations had to be entered manually. Automatically extracted

foundational metadata contributed 38 per cent of all the foundational metadata needed in Nesstar Publisher and about 60 per cent of the foundational metadata in Colectica Designer. Regarding Study-level metadata, DDI Codebook does not have the concept of grouping studies so we had no counts of metadata items for Nesstar Publisher in Table 7 in the “HDSS studies group” row. In Colectica Designer studies are grouped together in what is called a Series. We put the HDSS rounds together in an HDSS series, documenting each round as a separate study. The amounts of metadata required for HDSS at study level are comparable for Nesstar Publisher and Colectica Designer. There was little reuse of study-level metadata across studies as most of the metadata provided at study-level are specific to the particular study.

The data collection section is the one where a lot more metadata are provided for in DDI Lifecycle compared to DDI Codebook. Methodology description and collection events had similar metadata requirements for both Nesstar Publisher and Colectica Designer. However DDI Lifecycle provides far more metadata and structure related to instrument description. It was possible for us to build digital versions of HDSS paper questionnaires or CSPro data entry applications for rounds 26 and 27 from Colectica Designer. The paper questionnaires that we built were similar to the ones that would have been used during the actual data collection if rounds 26 and 27 had used paper questionnaires. However, the data collection applications for CSPro generated by Colectica Designer did not represent their final state, and more work would need to be done to include loops and skips as there are no inbuilt functions to do these in CSPro so they are implemented using user-defined functions. DDI Codebook mainly provides textual description and bibliographic information for a questionnaire, thus there are few metadata elements for HDSS questionnaire documentation in Nesstar Publisher.

The Datasets metadata section is divided into metadata relating to a dataset as a whole and variable-level metadata. This is where we entered most of the metadata in Nesstar Publisher. In both Colectica Designer and Nesstar Publisher, variables within a given data file are linked to their source questions where applicable. The same source questions entered during instrument development are referred to in Colectica Designer.

Here we also see comparable amounts of metadata between Nesstar Publisher and Colectica Designer in round 26 and due to metadata reuse, fewer items are needed for round 27 in Colectica Designer, mainly to cater for variables not present in round 26. We distinguished between metadata that editors automatically extracted and those that we manually entered. In round 26, 48% of the metadata were automatically entered from Stata files for Nesstar

Publisher and 20% for Colectica Designer. Round 27 had a similar percentage of automatically extracted metadata in Nesstar Publisher (44%) while in Colectica designer automatically extracted metadata went up to 34%.

Further details on staff training needs and the software costs are shown in Table 8.

Table 8: Training materials and software costs

	Nesstar Publisher	Colectica Designer
Pages of documentation in user manual read by documentalist	80 pages	100 pages
Pages of training material prepared for metadata entry staff	PowerPoint presentations – 80 slides, 30 pages Handbook.	Handbook under development - 40 pages Power point presentations - 250 slides
Self-study time and courses taken by documentalist	1 – 2 months initial Self-study - IHSN toolkit, Nesstar Publisher user's guide and DDI codebook online documentation	1 week DDI lifecycle Training and One day DDI / Colectica Workshop 4 - 6 months DDI Lifecycle self-study and practical work in Colectica Designer
Time taken to train metadata entry staff	2 weeks initial, 3 months during work	Not done
Cost of metadata preparation software	Nesstar Publisher - Free	Colectica Designer Monthly license - US \$65 per seat (logged in user) Annual license - US \$59 per month Perpetual license – US \$2000 per seat
Cost of archiving service	NADA - free Nesstar Server - commercial fee not specified on website	Colectica Repository US \$5000 - US \$74000 depending on selected options

The documentalist used a combination of short courses and self-study of online resources to get started with DDI and its metadata editors. Knowledge of DDI Codebook and Nesstar Publisher was acquired using the IHSN resources in form of a toolkit comprising sample documentation in Nesstar Publisher and a step-by-step DDI Codebook documentation guide (Dupriez and Greenwell 2007). In addition, the DDI Codebook online documentation on the DDI Alliance website² was used. For Colectica Designer, the documentalist attended a one-week introduction to DDI Lifecycle course, and a one-day introduction to DDI Lifecycle and Colectica course. In addition, he spent between 4 to 6 months of self-study of DDI Lifecycle resources available on the DDI Alliance website mainly in the form of DDI Lifecycle documentation, conference presentations and working papers. Parallel to that, practical activities were also carried out in Colectica Designer.

To prepare metadata entry staff, we spent two weeks on initial Nesstar Publisher training. It then took 3 months of close supervision to get them comfortably working independently.

² <http://www.ddialliance.org/>

Regarding software costs, Nesstar Publisher is available for free while Colectica is commercial software with pricing at the time of writing as given in Table 8.

Most of the online resources were accessible to the documentalist but difficult to understand for metadata entry staff at our disposal. The documentalist made the online resources that he had accessed available to the metadata entry team with follow up explanations to help them understand the content.

4.6 Discussion

We investigated the implementation of DDI on the existing Kisesa HDSS data. In particular, we paid attention to the identification of the steps involved in the Kisesa HDSS data production and their relationship to the GLBPM, the choice of DDI tools to use, the amount of metadata to be entered, the staff training needed and the software costs involved. We used Nesstar Publisher and Colectica Designer as our DDI Codebook and DDI Lifecycle tools respectively.

Our first finding is that the number of metadata items that had to be entered in Nesstar Publisher and in Colectica Designer were comparable when an HDSS round was documented from scratch. Documenting a subsequent round reduced the amount of metadata entry drastically in Colectica Designer due to reuse of metadata from the earlier round. This is supported by the fact that we needed to enter 3105 items in Nesstar Publisher and 3637 items in Colectica Designer when we documented round 26 from scratch. Round 27 required 5103 in Nesstar Publisher and 2212 in Colectica Designer.

Our second finding is that though the metadata editors automatically extracted some metadata from the Stata files we used, we still had to manually enter the majority of the metadata in both Colectica Designer and Nesstar Publisher. This is supported by the observation that metadata automatically extracted from Stata files for round 26 catered for 48% of the metadata entered in Nesstar Publisher while it was about 14% for Colectica Designer. In round 27 it was 43% and 33% respectively. In each case we still had to manually enter more than half of the metadata that we considered necessary. Nothing in the Colectica or Nesstar software documentation indicate these packages' capabilities in terms of automatically extracting metadata from R. If not provided for, probably an R user can export to either SPSS or Stata and then import into the metadata management software in order to benefit from automatic metadata extraction.

Our third finding is that more staff training and stiffer financial demands were required to implement Colectica Designer than Nesstar Publisher. This is supported by the time taken

to get the training done for the staff and the reported software costs. The documentalist spent about 2 months of initial study of DDI Codebook, the IHSN toolkit and the Nesstar Publisher user's manual before embarking on the preparation of training materials for metadata entry staff. It then took another one to two months to get the training material ready. For comparison, Colectica Designer took a week of formal training by DDI Alliance-affiliated DDI Lifecycle developers, an introduction to Colectica pre-conference workshop and 4 – 6 months of online DDI Lifecycle resources searching and study. Concurrent to the self-study, the documentalist was having practical sessions learning the Colectica Designer software. With respect to costs, Nesstar Publisher and the NADA software are free, whereas Colectica Designer is commercial and so are the Colectica repository and portal (the data and metadata storage system and its web application for cataloguing the data).

Regarding the mapping of the Kisesa HDSS data production process, we mapped this process to 28 sub-steps of the GLBPM. The GLBPM sub-steps we did not use are in one of the three categories: not supported within the Kisesa HDSS data production process, not suitable for the round of HDSS under consideration or not applicable to the HDSS type of studies. This mapping helped to describe the Kisesa HDSS data production process in a standardised and coherent manner.

We faced some challenges during the mapping. For some activities, we could not find the exact sub-steps to map them to. The mapping also required input from a wide range of staff involved in the data production process who often could not give immediate response as they needed to first study the GLBPM. In those cases, we made efforts to gather their understanding of the steps they were responsible for and we centrally mapped their feedback onto the GLBPM. This procedure is in contrast to that used by another study that worked on a similar mapping but to a different reference model (Ausborn, Rotondo, and Mulcahy 2014). Gathering input from staff on their responsibility and then mapping centrally takes away the need for the concerned staff to understand GLBPM.

The generic tools for data documentation that we used, arguably among the best currently available, still involve a lot of manual entry of metadata and parsing through free-text documents, in the form of questionnaires, protocols or reports, in search of study-level metadata, involved organisations, the concepts being measured and so on. This requires trained documentation personnel who understand DDI, especially if DDI Lifecycle is to be produced, having necessary skills to work out study concepts from proposals, questionnaires and publications. This does not mean that the DDI Lifecycle standard is unsuitable, however; it just means that its complexity makes it difficult to use generic tools for most of the steps

within the GLBPM. In practice, HDSS studies clearly do not need to leverage most of the additional features of DDI Lifecycle; however, there are some parts of the standard that would be advantageous (referenceability, versioning and comparison, for example). The generic tools seem to be most useful once the DDI content has been created. This seems to suggest that a sensible next step would be to consider development of bespoke software solutions, funds permitting. The bespoke tools would cover the parts of the documentation process that involve manual metadata entry. Much of the Data Dissemination and Discovery (step 7 in the GLBPM) could be supported by using the generic tools. The question of generic versus bespoke tooling therefore needs to be explored for each of the other process steps in the GLBPM.

We have only considered two metadata editors but there are other DDI Lifecycle editors in development that are free -- for example, DDI on Rails (Hebing 2015a), the Danish Data Archive's DdiEditor (Jensen 2012) and Questasy (de Bruijne and Amin 2009). It would be worthwhile to carry out a more extensive exploration of the wider range of tools to see if any of the ones we did not consider would offer distinct advantages in the documentation of HDSS data. We chose Colectica Designer over the others as it was arguably the most generic at the time we were starting our documentation work. Questasy, which was originally designed for the CentERdata at Tilburg University in the Netherlands, is now being developed further to make it more generic (Edwin de Vet, scientific programmer at CentERdata, personal communication). DDI on Rails was not yet available when we started. Other HDSS studies have taken this route of documenting their existing HDSS data using DDI Codebook. These include the Africa Centre (AC) for Health and Population Research in South Africa (Dr. Kobus Herbst, personal communication) and the Africa Population and Health Research Centre (APHRC) in Nairobi, Kenya (APHRC, 2014). These two studies are larger than the Kisesa study, covering populations of 85,000 (Tanser et al. 2008) and 65,000 (Beguy et al. 2015) respectively compared to Kisesa's 35,000. The AC HDSS currently acts as a platform for 5 research programmes; each with its own sub-studies. Since its inception in 2002, the APHRC has had more than 15 projects, using its HDSS as a platform, compared to 4 sub-studies in Kisesa. They are also better resourced in terms of IT and programming staff, compared to Kisesa. But even with this level of sophistication they have not yet adopted the more advanced technology offered by the DDI Lifecycle approach, which has hitherto been used only by studies in more developed countries, such as the MIDUS study in the USA (Radler, Iverson, and Smith 2013), the CLOSER project in the UK (Gierl and Johnson 2012), Statistics Denmark (Nielsen, Iverson, and Smith 2013), and Statistics New Zealand

(Brown et al. 2012). It would appear that this technology will not be rapidly adopted by HDSS in LMIC.

One important finding, which was not part of the original remit of this investigation, is awareness of how much harder it is to include in the study documentation a questionnaire that has been developed for collecting data on an electronic device rather than on paper. HDSS, which moved to electronic data collection using specialist software like CSPro, need to be aware that for documentation purposes they need to develop paper versions of the questionnaire for explanatory purposes, or supply the code and its interpretation (e.g., as screen shots) as part of the documentation package. This has many benefits among them facilitation of future studies preparations by analysts.

4.7 Summary

In summary, our study shows that at least for a typical African HDSS, it is not so much the difference in the amount of metadata to be entered but rather, the staff training requirements and the software costs that producers should consider when deciding between DDI Codebook and DDI Lifecycle. If available staff expertise is capable of learning and implementing DDI Lifecycle, an HDSS could directly start with DDI Lifecycle; otherwise, they would better start with DDI Codebook and then move on to DDI Lifecycle at a later stage. The Kisesa study is used as an example but the general principles would apply to other African HDSS studies.

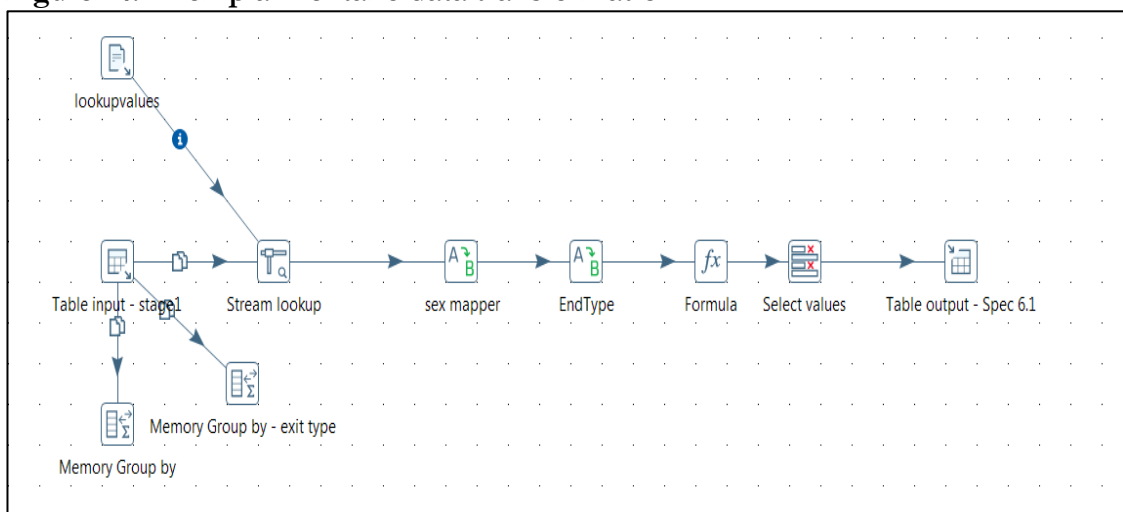
5. HIGH LEVEL, STRUCTURED METADATA FOR ALPHA DATA TRANSFORMATIONS:

Foundational content for the African Demographic and Epidemiological Surveillance Business Process Model (ADESBPM)

5.1 Introduction

In their seminal paper on bespoke infrastructure for HDSS data harmonisation and curation (Herbst et al. 2015), Herbst and colleagues demonstrated the setting up and maintenance of a predominantly open source, secure and robust environment for HDSS data harmonisation and curation suitable for LMIC settings – the CiB technology. Inspired by the success of the CiB within the INDEPTH network, ALPHA is adopting the CiB and extending its scope to cater for the network’s various datasets, the ALPHA specifications. Use of the CiB technology will ameliorate the reproducibility and transparency problems of the ALPHA data transformations alluded to in section 3.4. The CiB combines inbuilt Pentaho documentation of data transformation processes and DDI codebook annotation of the data specifications to produce a data lineage record. The Pentaho software provides a graphical interface that shows the various data processing tasks and their interconnectedness in flow diagrams such as the one in Figure 24. In addition, it allows for textual descriptions of tasks at various levels of granularity.

Figure 24: Exemplar Pentaho data transformation



However, from a data documentation best practice perspective, inbuilt Pentaho provenance metadata present three challenges: (i) They do not adequately cater for the research domain

of interest, (ii) they are tool specific and (iii) in cases when Stata scripts are executed within Pentaho, the script content is a blackbox, not accessible to software agents seeking automated mining of metadata.

Depending solely on tool specific/ proprietary metadata has well documented draw backs (Rasmussen 2014; Corti and Gregory 2011; Van den Eynden et al. 2011; Locke and Lowe 2007). These include limited access to and understanding of the harmonised data for non-Pentaho users both internal and external to ALPHA, and limited “future proofing” of the metadata and data usability (Corti and Gregory 2011; Van den Eynden et al. 2011).

This problem has not been addressed in the literature within the context of harmonisation of LMIC longitudinal demographic and HIV or epidemiological surveillance data. A business process model mapped to GSBPM/ GLBPM would facilitate international business level communication of HDSS data management. On the other hand, the envisioned model would define and describe activities relating to HDSS event data harmonisation in more concrete terms than the GLBPM, it would be a specialisation of the GLBPM. Besides the activities, it would capture the information objects flowing between these activities as inputs and outputs for tasks. The information objects would be in the form of the events of interest to HDSS practitioners and data users alluded to in chapter 3, the sequence and timing of those events and the assumptions and rules pertaining to censoring, loss to follow-up among other concepts important for defining exposure to risk.

This chapter and chapter 6 investigate procedures for the liberation of provenance metadata for ALPHA datasets from being exclusively Pentaho specific. Chapter 5 takes a high level perspective developing a specialisation of the GLBPM comprising of an integrated description of sub-processes and information objects pertinent to HDSS event histories. Chapter 6 focuses on the more granular details of the transformations. The contributions of these two chapters are to enhance CiB metadata by making them tool agnostic, HDSS domain sensitive and compliant with international standards.

5.1.1 Aims and objectives

The aim of this chapter is to explore the development of ALPHA data transformations documentation calibrated to the African population-based longitudinal demographic and epidemiological surveillance domain - represented by ALPHA ETLs - through:

1. Identifying the main activities defined in the ALPHA ETLs for specification 6.1
2. Mapping the identified activities to the GLBPM.
3. Specialising the mapped GLBPM steps to the HDSS domain.
4. Assembling inputs and outputs associated with the identified business processes

The activities and their overviews will constitute the foundational content for a process model for African demographic and epidemiological surveillance domain, the ADESBPM.

5.1.2 Related work

The literature on the documentation of data transformations was described in chapter 2. It includes current DDI Lifecycle approach (DDI Alliance 2015b; Marker et al. 2009), the Validation and Transformation Language (VTL) (SDMX Technical Working Group 2018c) and the Structured Data Transform Language (SDTL) (C2Metadata 2017). The current DDI Lifecycle specification uses textual descriptions of the transformations and inclusion or reference to the original code used to create the transformation. SDTL is considered in detail in the next chapter as it is concerned with the granular details of data transformations. At a business level, the solutions reported in the literature are not exclusively aimed at provenance documentation, they cover a broader scope, the entire data life cycle. This chapter explores the application and adaptation of these models for description of ALPHA ETLs.

The GSBPM (UNECE 2018b) has been widely adopted by national statistics offices across the high income nations (Brancato and Simeoni 2012; UNECE Secretariat 2009; Ausborn, Rotondo, and Mulcahy 2014). Often, the national statistics offices have mapped their own business processes to the model. In some cases, attempts have been made to specialise the GSBPM to meet local contexts. In the majority of these cases, no attempts have been made to add structure to the specialisation. The Australian Bureau of Statistics (ABS) is one of the few examples where the specialisation has been done in a structured fashion (UNECE 2018c). The ABS example is the best to date in terms of contextualising the GSBPM in a structured format. It is however being done within the official statistics domain is backed up by the use of off the shelf commercial software currently not affordable for ALPHA. There is therefore, no equivalent within the demographic and epidemiological surveillance domain which ALPHA could emulate. It remains unclear, how to document the data transformations performed within the CiB in a more generic, structured and HDSS domain aware manner. Such documentation is needed to shed light on the provenance of the ALPHA datasets.

Limited understanding of the ALPHA data transformation processes impedes the usability of those data. Besides hindering usability, the lack of structured metadata constrains the ability of computer programs to automate repetitive data processing tasks thus burdening human resources and increasing data production costs.

In an effort to improve our understanding on the documentation of the ALPHA data transformations, the current chapter investigates the application of DDI and the Generic

Longitudinal Business Process Model (GLBPM), for ALPHA data provenance documentation.

5.1.3 Chapter overview

The foregoing section gave an introduction to the chapter identifying the research problem and the aims of the work presented in this chapter. The ensuing section 5.2 relates the contents of this chapter to the previous chapters 3 and 4, it thus shows how the current chapter advances the story of ALPHA datasets documentation by supplementing source data documentation with high level documentation of ALPHA data transformations. Figure 25 depicts how the current chapter expands on foundational materials presented earlier. Section 5.3 goes on to describe the methods employed in determining the metadata content and structure for describing ALPHA data transformations, the mapping of ALPHA ETLs to the GLBPM and the specialisation of the identified sub-processes of the GLBPM. Thereafter, the mechanism employed to capture metadata not provided for in Pentaho is outlined. Section 5.4 presents the results. The last two sections comprise a discussion and a summary of the Chapter.

5.2 Connecting Chapter 5 to Chapters 3 and 4

Figure 25 gives the bigger picture within which this chapter is placed. The rectangular blocks represent processes and the trapezia represent objects. The colour coding is used to identify the thesis chapter in which a particular object/ process was first introduced. The same colour coding is also followed in Table 9 which is providing an annotation for Figure 25.

The objects and processes in blue represent the state of the art regarding implementation of structured documentation in ALPHA member institutions. As described in Chapter 3, documentation within member institutions is mainly supported by the International Household Survey Network tools, Nesstar Publisher and NADA catalogue. The object “*Generic Longitudinal Business Process Model (GLBPM)*” and the process “*Mapping source data processing to GLBPM*” – both in green - were first introduced in Chapter 4. They represent the adoption of a process perspective in analysing and describing data management activities in a typical ALPHA member institution.

Figure 25: Input objects, processes and output objects in the ALPHA ETL and their structured documentation: High level

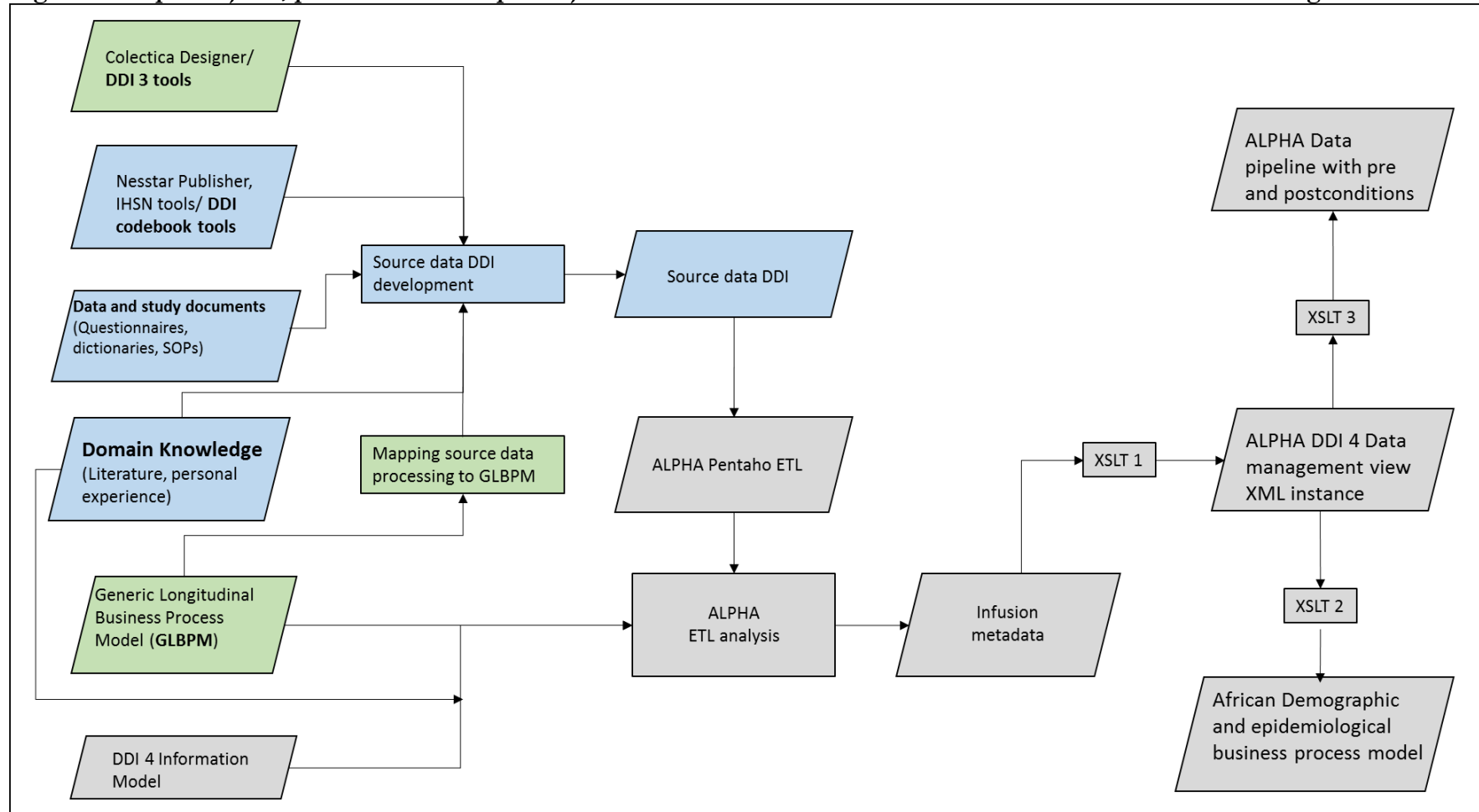


Table 9: Description of the objects and processes involved in ALPHA ETL – High level view

Object/ process	Description	Inputs	Outputs	Chapter*
Data and study documents	Questionnaires, protocols, manuals, data dictionaries, edit checks etc	X	Source data DDI	3, 4
DDI Codebook tools	Nesstar Publisher, IHSN tools, Dataverse etc	X	Source data DDI	3, 4
DDI 3 tools	Colectica designer, Rogatus or other DDI 3 editors	X	Source data DDI	4
Source data DDI	DDI 2 or DDI 3 metadata for the data produced in the member institutions and used to prepare ALPHA data	Data and study documents, Domain Knowledge, GLBPM, DDI 3 tools, Nesstar publisher	ALPHA Pentaho ETL	3, 4, 5
Mapping source data processing to GLBPM	Process: Analysing member institution source data management process and finding equivalent steps in the GLBPM	GLBPM, Member institution source data management processes	Source data DDI	4
ALPHA Pentaho ETL	ALPHA data transformations performed on member institutions source data within the Pentaho data integration platform	Source data DDI	Infusion metadata	5
Domain Knowledge	HDSS literature and data management experiences of the metadata producer	x	Source data DDI, Infusion metadata	3, 4,5
GLBPM	Generic Longitudinal business process model	x	Source data DDI, Infusion metadata	4, 5
ALPHA ETL Analysis	Process: Reviewing the Pentaho Jobs and transformations used for ALPHA data transformations	Domain knowledge, ALPHA Pentaho ETL, DDI 4 Information model, GLBPM	Infusion metadata	5
DDI 4 Information model	A library of objects encompassing the entire DDI 4 without specific views or schemas	x	Infusion metadata, DDI 4 Data management view XML instance	5
Infusion metadata	Contextual metadata to supplement those in Pentaho ETL	GLBPM, Domain Knowledge, DDI 4 Information model	DDI 4 Data management view XML instance	5
XSLT 1	XSLT application for transforming the infusion metadata into DDI 4 Data management view XML instance	Infusion metadata, DDI 4 Information model	DDI 4 Data management view XML instance	5
DDI 4 Data management view XML instance	A DDI 4 functional view (Specific application) relating to ALPHA ETL constructed from DDI 4 information model classes () specialisation	Infusion metadata, DDI 4 Information model	ALPHA Data pipeline, ADESBPM	5
XSLT 2	XSLT application for transforming DDI 4 Data management view XML instance into the ADESBPM	DDI 4 Data management view XML instance	ADESBPM	5
ADESBPM	African Demographic and Epidemiological Surveillance Business process Model – a specialisation of the GLBPM	DDI 4 Data management view XML instance	x	5
XSLT 3	XSLT application for transforming DDI 4 Data management view XML instance into the ALPHA Data Pipeline	DDI 4 Data management view XML instance	ALPHA Data pipeline	5
ALPHA Data pipeline	A list/ sequence of business processes working one after the other to accomplish the goal of transforming source data to ALPHA specifications	DDI 4 Data management view XML instance	x	5

The mapping to GLBPM was used as input in deciding on suitable elements of DDI to use. Objects and processes introduced in Chapter 5 are in grey. They show the development of “*infusion metadata*” – described in section - and their further processing into an instance of a DDI 4 data management view which was then processed to present a data pipeline with processes and their inputs (pre conditions) and outputs (post conditions) and foundational content for the ADESBPM.

5.3 Methods

5.3.1 Determining metadata content and structure for ALPHA ETLs

Literature on African demographic and epidemiological surveillance systems and personal experiences provided the bulk of the domain metadata content. Additional metadata content was extracted from the Pentaho files. The GLBPM and the DDI standard (mainly classes in DDI 4 model) provided a standardised vocabulary suitable for describing longitudinal data production processes and the structure for the metadata.

Literature on African demographic and epidemiological surveillance systems

I drew on the cohort profiles for the member institution used in this study (Beguy et al. 2015), the HDSS data reference model related publications (Sankoh and Byass 2012; INDEPTH Network 2002; Benzler, Herbst, and Macleod 1998; Bocquier et al. 2017), the INDEPTH network Centre in a Box technology (Herbst et al. 2015) and personal experiences working with HDSS data for the metadata content relating to ALPHA data transformations description.

Analysis of the Pentaho transformation shows what was done to transform data. What is not clear is the reasoning behind the decisions to transform data in one way rather than the other. The literature provided theoretical underpinnings and rationale for particular decisions.

Pentaho information model and the ALPHA ETLs

This study analysed ETL routines for creating ALPHA spec 6.1 shown in Figure 18. The details of this and the other ALPHA data specifications were given in Chapter 3, Table 4 and are available at <http://alpha.lshtm.ac.uk/metadata/>.

I only present the metadata resulting from the analysis of the common sub-jobs as there is substantial variation in the member specific sub-job which deemed it unfit for the procedures used in this study. As such, alternative approaches to documenting its transformations maybe more appropriate.

To give an idea of the breadth of the reviewing and analysis involved - the numbers of sub-jobs, transformations and steps analysed, Table 11 shows the quantities in each of the jobs and the number of steps in the transformations in each job.

Table 10: Numbers of job entries and transformation steps in the ALPHA 6.1 ETL

Sub-job	Number of transformations ³	Number of steps
CORE Produce Raw 61 Dataset	6	42
002 CORE Data Quality Metrics	12	130
003 CORE Data Cleaning	17	105
002 CORE Data Quality Metrics	12	130
004 CORE Data Anonymisation	2	14
005 CORE Consolidate Start and End Events	14	61
006 CORE Verify Temporal Integrity	7	28
007 CORE Update Event Timing	13	86
008 CORE Produce Final Core MicroData Files	5	44
009 CORE Prepare Data Quality Feedback	7	14

Generic Longitudinal BusinessProcess Model

The GLBPM provided the overarching framework within which the metadata for ALPHA ETLs were developed. The full model distinguishes nine phases in the longitudinal data production process which are: “Evaluate needs”, “Design”, “Build”, “Collect”, “Process”, “Archive”, “Disseminate”, “Research/ Publish” and “Retrospectively Evaluate”. Each phase is divided into a number of sub-processes. For example,

“Evaluate needs” phase: “Define research questions”, “Evaluate existing data”, “Define concepts”, “Establish outputs”, ...

“Processing” phase: “Integrate data”, “Classify and recode”, “Impute missing data” ...

However, since it is not a rigid tool requiring use of all phases, I focused only on the phases that are relevant to the ALPHA ETL context.

5.3.2 ALPHA ETL through GLBPM lenses: Mapping and specialising GLBPM

A bottom up approach was followed to build structured metadata potentially reusable within a data specification ETL life cycle, across data specifications and member institutions. The ETL implementation were reviewed first and then abstraction from the Pentaho-specific details was done to generate the reusable domain metadata. These metadata form a basis for a specialisation of the GLBPM into a domain specific business process model called the African Demographic and Epidemiological Surveillance Business Process Model. This model

³This number is not only that of transformations, it also includes entries such as an SQL scripts.

aims to capture structured metadata for data production processes in the African population based longitudinal cohorts within the demographic and epidemiological surveillance domain.

Mapping ALPHA ETL to GLBPM

The first step was to identify the main non-overlapping activities involved in creating the ALPHA 6.1 specification. Each of the main activities, represented by an ALPHA ETL Pentaho sub-job, was mapped to a corresponding GLBPM step or sub-step. The sub-job names were often suggestive of the sub-process they mapped to in the GLBPM but not sufficient to definitively identify the mapping. There was need for reviewing the content of the sub-job to determine what tasks were being accomplished. Once the review provided enough information, the GLBPM sub-process which described the tasks the closest was tagged as a mapping.

Since the GLBPM is designed to cater for a wide range of longitudinal and panel studies (I. Barkow, Block, Greenfield, Gregory, Hebing, Hoyle, and Zenk-möltgen 2013; Hebing 2015b), using its vocabulary to describe ALPHA ETL facilitates communication of ALPHA data management processes with users outside the network.

Specialising mapped GLBPM sub-processes

The GLBPM does not fully capture the intricacies of the ALPHA transformations in enough detail needed by a network member or other domain experts for full understanding. Further, because it is aimed at a broader user community, the GLBPM is also too generic and does not use terms that are familiar to the ALPHA data managers and researchers. To mitigate these two problems, the identified GLBPM steps and sub-steps were specialised to contextualise them within the population-based longitudinal demographic and epidemiological surveillance domain. To do this specialisation, a deeper review of the contents of each of the sub-jobs was performed with the aim of understanding and describing the details of the jobs including the transformations contained therein.

Contents of each sub-job were described by 3 attributes: (i) an overall purpose statement, (ii) an “algorithm overview” - which is a summary of the tasks in a sub-job provided through outlining the steps involved – and (iii) by input and output data records. The choice of these attributes is mainly based on process modeling theory and examples of application of the Generic Statistical Business Process Model (GSBPM) a process model for National Statistics production (Brancato and Simeoni 2012). The GSBPM is the ancestor of the GLBPM.

Input and output data records

The purpose statement and algorithm overview were complemented by data records representing the inputs processed in the sub-job and the outputs resulting from the processing. Each sub-job was associated with its input data records (pre-conditions) and output data records (post-conditions). In some cases, an output was presented in more than one format. For example, in addition to a table added to a staging database, another version of the data was produced in Excel or Stata format. This existence of “sibling” datasets was also recorded in the metadata. In other cases, output data had slightly different structures due to actions such as dropping of variables in one of the outputs. To handle this, “exceptions” were documented to specify the difference in the outputs.

5.3.3 Metadata infusion file - template for domain metadata capturing

Thus far, the methods described have focused on “what” metadata were captured. This section turns to “how” the metadata were captured.

Being partially self-documenting, Pentaho supports the capture of some of the required metadata, for example, through transformations overview and textual descriptions of the steps. Work was done with the support of a data/ metadata systems architect to determine the extent to which Pentaho natively supported the recording of the needed metadata. To capture the metadata not provided for in Pentaho, the metadata architect used GLBPM, the DDI 4 information model and the metadata content decisions made based on the literature review and the author’s HDSS data management experience. An initial version of a bespoke metadata template, in XML format, was developed together with its schema. This template was named the metadata infusion file. The⁴ author and the metadata architect then refined the initial version to the current working version through iterative steps of reviews and modifications.

The infusion file was then used to capture master-job level overview metadata, demographic surveillance concepts, algorithm overviews, purpose statement for each sub-job, the inputs and outputs from each sub-job among other metadata and bibliographic metadata for attributing the various aspects of the ALPHA ETL to the appropriate contributors.

The structure for the infusion file comprising of overarching and sub-job specific components such as purpose statement, mapping to steps or sub-steps in the GLBPM, algorithm overviews and pre and post conditions form ALPHA ETLs’ contributions to DDI

⁴ The author provided domain expertise while the architect provided DDI and XML expertise in developing the metadata template

4 development. ALPHA ETLs' analysis and documentation is a use case for the development of a data management functional view for DDI 4.

5.4 Results

This section presents the results of the characterisation of the ALPHA ETLs from the perspective of the GLBPM. It provides the results for the mapping and the specialisation of the mapped steps.

5.4.1 Results of literature review on African demographic and epidemiological surveillance systems

The literature provided the context within which decisions were made to transform data in particular ways across the ETL. This context is captured across the ensuing sections. In Table 11 we see a mapping between the sub-jobs and the HDSS reference data model described by Benzler, Herbst, and MacLeod (1998) and in (INDEPTH Network 2002). In addition, we see the relationships between the literature and the data transformation activities in the ETL in the algorithm overviews presented in section 5.4.3. The algorithm overviews draw from the HDSS reference data model. In addition, they draw from the procedures described by Bocquier et al. (2017) for creating and quality assessing residency episodes datasets from demographic and health surveillance data.

5.4.2 Mapping ALPHA ETL sub-jobs to GLBPM

The results of mapping the sub-jobs to the GLBPM are shown in Table 11 and in Figure 26. In each row of Table 11, a sub-job is related to the HDSS reference data model, a purpose statement and a GLBPM step.

In Figure 26, the ALPHA ETL is superimposed on the GLBPM. The blue circles show which steps of the model are involved in the ETL and the sky blue and green arrows show the sequence in which the steps are performed in the ETL.

The ALPHA ETL only mapped to the “Process/ Analyse” phase of the GLBPM. Since our interest in this study is the ETL process, we do not have mappings to the phases to do with study design, development of data collection instruments, data collection (Phases 1 – 4). Neither do we map to the later phases of the model (Archive, disseminate, publish and evaluation (Phases 6 – 9)).

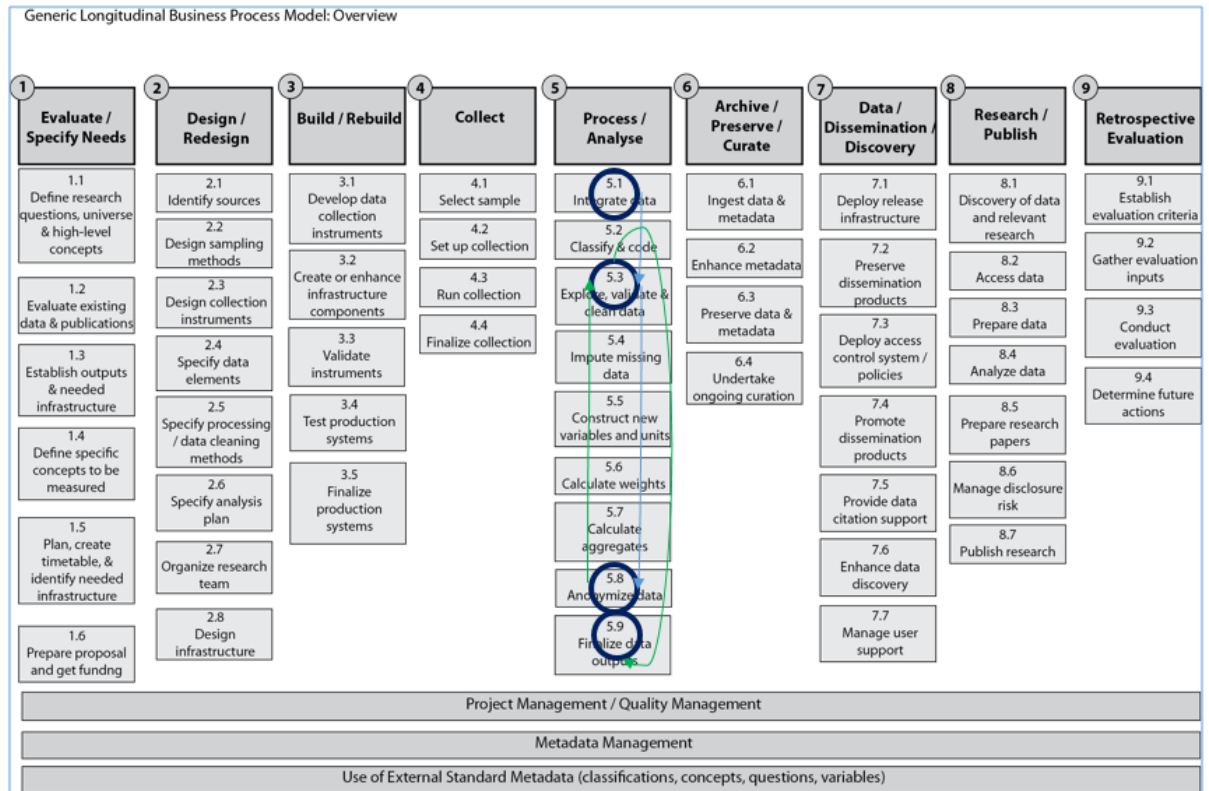
The GLBPM steps have very generic descriptions. While it communicates the essence of the tasks in an internationally understood manner, this mapping to steps of GLBPM tends to be too simplistic, it hides the complexity of the ETL. For example, most of the Pentaho sub-jobs map to the “Explore, validate and clean data” step of the GLBPM while they are doing

clearly distinguishable tasks. By itself, the mapping to the model steps does not give enough detail needed to fully understand, let alone reproduce the process.

Table 11: Mapping ALPHA ETL sub-jobs to GLBPM

Sub-job	HDSS reference model link	Sub-job purpose	GLBPM
CORE Produce Raw 6.1 Dataset	Transform relevant entities of the HDSS reference data model into raw ALPHA 6.1 Spec 6.1)	Produce an unedited ALPHA 6.1 Specification from the intermediate data	5.1 Integrate data
002 CORE Data Quality Metrics	Validate sex, dob, events order and events dates	Assesses the quality of the data in the raw specification created in business process, 02 Core ETL for Raw 6.1, on the basis of a set of quality metrics	5.3 Explore, validate and clean data
003 CORE Data Cleaning	Clean Event Dates and Event ordering	Applies cleaning procedures to correct some inconsistencies identified in the quality assessment business process (03 Core ETL Raw 6.1 Dataset Quality Metrics). This business process does not clean all the errors identified, those requiring the attention of the member centre are compiled in preparation for sending to the member centre	5.3 Explore, validate and clean data
002 CORE Data Quality Metrics	Validate sex, dob, events order and events dates	Reruns data quality metrics to assess the quality of the cleaned 6.1 using the same set of quality metrics used in 00 CORE Data Quality Metrics	5.3 Explore, validate and clean data
004 CORE Data Anonymisation	Anonymise individuals' IDs, Physical Locations IDs and Mothers' IDs	Randomises the individual and household identifiers to anonymise the data	5.8 Anonymise data
005 CORE Consolidate Start and End Events	Quantify duplicate events and drop single/unpaired events	Quantifies proportion of records that are duplicates in terms of unique-identifier, event and event date, cleans the duplicates and drops individuals with single events	5.3 Explore, validate and clean data
006 CORE Verify Temporal Integrity	Validate and clean event histories	Assesses the ordering, in time, of dates for consecutive/ successive events, compiles those with illogical timing, quantifies their proportion and drops individuals with wrongly timed successive events	5.3 Explore, validate and clean data
007 CORE Update Event Timing	Smooth event histories	Assesses and corrects migration event sequences. A movement out of the study area is defined as an external-outmigration (OMG) if the time between the external-outmigration and the subsequent external-immigration (IMG) is above a defined period of time (threshold) - e.g. six months.	5.3 Explore, validate and clean data
008 CORE Produce Final Core MicroData Files	Create calendar of events	Produces the final dataset in "events" format. Each row represents an event of interest (baseline recruitment, birth, external in-migration, Internal in-migration, Found after lost to follow up, etc) together with other data relating to the individual / event	5.9 Finalize data outputs
009 CORE Prepare Data Quality Feedback	Compile data quality metrics	Compiles data quality assessment report to be shared with the member centre	

Figure 26: Mapping of ALPHA ETL to the GLBPM



5.4.3 Specialising the mapped GLBPM steps

One way to address the limitations associated with only mapping to GLBPM, is to provide more details to describe the sub-process being mapped to GLBPM that is, specialising the mapped steps.

The results of specialising the mapped GLBPM steps by further describing the sub-jobs are given in Table 12 for four example sub-jobs. As shown in the last column, we add more details using “Algorithm overviews”. Algorithm overviews are summary descriptions of the steps taken to do the tasks comprising the sub-job of interest. The algorithm overviews aim to express the sub-process in a software agnostic manner, addressing the question of *what is done* without specifying *how it is done* in a specific software, in this case, Pentaho.

Consequently, if properly prepared, algorithm overviews give a metadata user a representation of the steps involved without requiring the user to know the syntax of the software used to implement the tasks. In addition, the specialisation gives the detail of each job using a language (concepts and terms) that are familiar to the users working in the demographic and epidemiological surveillance domain. This makes the processes easier to understand, evaluate, modify or reproduce.

Table 12: Specialisation of GLBPM steps for four exemplar ALPHA ETL sub-jobs

Sub-job	GLBPM	Algorithm overview (Specialisation)
CORE Produce Raw 61 Dataset	5.8 Anonymise data 5.1 Integrate data	<ol style="list-style-type: none"> 1. Generate anonymised unique-identifiers 2. Create a mapping between original and anonymised ids 3. Store the ids mapping information where it can be accessed internally in the future 4. Create raw spec 6.1 from staging data
002 CORE Data Quality Metrics	5.3 Explore, validate and clean data	<ol style="list-style-type: none"> 1. Compile a list of quality metrics relevant to the data specification 2. Create events consistency matrix showing the logical ordering of event sequences 3. Identify in the data, events that start a residency episode (birth, external-immigration, enumeration, becoming eligible for a study, found after being lost to follow-up, Internal-immigration) 4. Identify in the data, events that end a residency episode (external-outmigration, death, became ineligible for study, lost to follow-up, internal-outmigration, present in the study (right censored)) 5. Review the identified start events and distinguish between legal and illegal ones 6. Review the identified end events and distinguish between legal and illegal ones 7. Review all transitions between two events and distinguish between legal and illegal ones 8. Compile illegal, missing or unknown sex 9. Compile illegal, missing or unknown DOB 10. Calculate numbers of legal and illegal start events, end events, event transitions, sex values, out of range DOBs and missing sex and DOBs
003 CORE Data Cleaning	5.3 Explore, validate and clean data	<ol style="list-style-type: none"> 1. Check if the first event to be ever recorded for each individual is enumeration, birth or external-immigration 2. If first event is an internal-immigration change it to an external-immigration 3. Classify all first events other than enumeration, birth or external-immigration as illegal first events 4. Check if the marked as first event is a birth, an enumeration or an immigration from outside DSA 5. Drop individuals with illegal start events 6. Check if last events are external-outmigration, death, present in study site 7. If last event is an internal-outmigration change it to an external outmigration 8. Classify all last events other than external-outmigration, death, present in study site as illegal last events 9. Drop individuals with illegal end events 10. Identify current and next event and their dates 11. Check if a birth event is followed by a birth, an enumeration, external-immigration or internal-immigration 12. Check if a death event is followed by an event other than a NULL 13. Review all other transitions in the data and record violations of consistency matrix 14. Drop individuals with illegal transitions 15. Drop individuals with unknown sex or DOB
004 CORE Data Anonymisation	5.8 Anonymise data	<ol style="list-style-type: none"> 1. Bring together original and anonymised IDs in the cleaned spec 6.1 2. Create cleaned spec 6.1 with only anonymised IDs 3. Preserve an internal mapping of original IDs to the anonymised IDs

5.4.4 Input (pre-condition) and output (post-condition) data records

Thus far, the results have shown the mapping to GLBPM steps and the purpose statements for each sub-job. This section turns to the input and output data records for each sub-job.

Figure 27 shows a snippet of the preconditions and postconditions for the sub-job 002 CORE Data Quality Metrics. The picture also shows the format of the data involved and their location.

The input and output data records show us what data were required for a sub-job and what data were produced from it.

Figure 27 : Input and output data records for sub-job 002 CORE Data Quality Metrics

```
538 <preconditions>
539 <precondition>
540 <recordname>Raw_61_Event_Format</recordname>
541 <location>6.1/CORE/001 CORE Produce Raw 61 Dataset.kjb</location>
542 <type>SQL</type>
543 <aggregate>>false</aggregate>
544 </precondition>
545 </preconditions>
546 <postconditions>
547 <postcondition>
548 <recordname>StartingEvents</recordname>
549 <location>6.1/CORE/002 CORE Data Quality Metrics.kjb</location>
550 <type>SQL</type>
551 <aggregate>>false</aggregate>
552 <datadescription>
553 <sibling>StartingEvents</sibling>
554 <uri>6.1/CORE/02A CORE Illegal Start Events.ktr</uri>
555 <description>Excel Output</description>
556 </datadescription>
557 </postcondition>
558 <postcondition> [10 lines]
559 <postcondition> [10 lines]
560 <postcondition> [5 lines]
561 <postcondition> [5 lines]
562 <postcondition> [5 lines]
563 <postcondition> [5 lines]
564 <postcondition> [5 lines]
565 <postcondition> [5 lines]
566 <postcondition> [5 lines]
567 <postcondition> [5 lines]
568 <postcondition> [5 lines]
569 <postcondition> [5 lines]
570 <postcondition> [5 lines]
571 <postcondition> [5 lines]
572 <postcondition> [5 lines]
573 <postcondition> [5 lines]
574 <postcondition> [5 lines]
575 <postcondition> [5 lines]
576 <postcondition> [5 lines]
577 <postcondition> [5 lines]
578 <postcondition> [5 lines]
579 <postcondition> [5 lines]
580 <postcondition> [5 lines]
581 <postcondition> [5 lines]
582 <postcondition> [5 lines]
583 <postcondition> [5 lines]
584 <postcondition> [5 lines]
585 <postcondition> [5 lines]
586 <postcondition> [5 lines]
587 <postcondition> [5 lines]
588 <postcondition> [5 lines]
589 <postcondition> [5 lines]
590 <postcondition> [5 lines]
591 <postcondition> [5 lines]
592 <postcondition> [5 lines]
593 <postcondition> [5 lines]
594 <postcondition> [5 lines]
595 <postcondition> [5 lines]
596 <postcondition> [5 lines]
597 <postcondition> [5 lines]
598 <postcondition> [5 lines]
599 <postcondition> [5 lines]
600 <postcondition> [5 lines]
601 <postcondition> [5 lines]
602 <postcondition> [5 lines]
603 <postcondition> [5 lines]
604 <postcondition> [5 lines]
605 <postcondition> [5 lines]
606 <postcondition> [5 lines]
607 <postcondition> [5 lines]
608 <postcondition> [5 lines]
609 <postcondition> [5 lines]
610 <postcondition> [5 lines]
611 <postcondition> [5 lines]
612 <postcondition> [5 lines]
613 <postcondition> [5 lines]
614 <postcondition> [5 lines]
615 <postcondition> [5 lines]
616 <postcondition> [5 lines]
617 <postcondition> [5 lines]
618 <postcondition> [5 lines]
619 <postcondition> [5 lines]
620 <postcondition> [5 lines]
621 <postcondition> [5 lines]
622 <postcondition> [5 lines]
623 <postcondition> [5 lines]
624 <postcondition> [5 lines]
625 <postcondition> [5 lines]
626 <postcondition> [5 lines]
627 <postcondition> [5 lines]
628 <postcondition> [5 lines]
629 <postcondition> [5 lines]
630 <postcondition> [5 lines]
631 <postcondition> [5 lines]
632 <postcondition> [5 lines]
633 <postcondition> [5 lines]
634 </postconditions>
```


Table 13 gives a summary of the numbers of data records involved as preconditions and postconditions for the sub-jobs in the ETL for specification 6.1.

Table 13: Numbers of pre and post conditions for each sub-job in specification 6.1 ETL

Sub-job	Pre-conditions	Post-conditions
CORE Produce Raw 61 Dataset	1	3
002 CORE Data Quality Metrics	1	12
003 CORE Data Cleaning	1	9
002 CORE Data Quality Metrics	1	11
004 CORE Data Anonymisation	3	1
005 CORE Consolidate Start and End Events	1	9
006 CORE Verify Temporal Integrity	1	4
007 CORE Update Event Timing	1	9
008 CORE Produce Final Core Microdata Files	3	3
009 CORE Prepare Data Quality Feedback	7	7

5.4.5 Metadata infusion file

APPENDIX A shows the schema of the infusion file. In summary, the infusion file contains business level metadata for a study or an ALPHA dataset as a whole. These metadata include an overview, concepts, methodology, design overview, data pipeline (overall process) among others. The entire process is divided into business processes that map to Pentaho sub-jobs. Each business process is characterised by an algorithm overview, a human intelligible step by step description of the business process. This description is complemented with pre-conditions and post conditions.

One value of the ADESBPM is that it will help to generate infusion type metadata in the future with less human assistance and more machine assistance. Metadata and processes reuse will be automated to generate drafts of infusion metadata that will then be updated by humans.

Figure 28: Snippet of an algorithm overview within the infusion metadata file

```
<algorithmoverview>
  <step id="5.1" name="Compile a list of Quality Metrics">Compile a list of quality metrics relevant to the data specification</step>
  <step id="5.2" name="Create events consistency matrix">Create events consistency matrix showing the logical ordering of event sequences</step>
  <step id="5.3" name="Compile residency starting events">Identify in the data events that start a residency episode (birth, external-immigration, enumeration,
    becoming eligible for a study, found after being lost to follow-up)</step>
  <step id="5.4" name="Compile residency ending events">Identify in the data events that end a residency episode (external outmigration, death,
    became ineligible for study, lost to follow-up, internal-outmigration, present in the study (right censored)) </step>
  <step id="5.5" name="Compile legal and illegal start events">Review the identified start events and distinguish between legal and illegal ones</step>
  <step id="5.6" name="Compile legal and illegal end events">Review the identified end events and distinguish between legal and illegal ones</step>
  <step id="5.7" name="Compile legal and illegal transitions">Review all transitions between two events and distinguish between legal and illegal ones</step>
  <step id="5.8" name="Compile illegal, missing or unknown sex">Compile illegal, missing or unknown sex</step>
  <step id="5.9" name="Compile illegal, missing or runknown dob">Compile illegal, missing or unknown dob</step>
  <step id="5.10" name="Compile quality metrics">Calculate numbers of legal and illegal start events, end events, event transitions, sex values, out of range DOBs
    and missing sex and DOBs</step>
</algorithmoverview>
```


5.5 Discussion

The work done in this chapter had two goals: (i) to harness domain knowledge and Pentaho inbuilt provenance metadata in generating high level, tool agnostic and domain sensitive documentation for ALPHA specification 6.1 ETL. (ii) To format the generated provenance metadata in compliance with international metadata standards.

To achieve the two goals, the Pentaho ETL for spec 6.1 was analysed and mapped to the GLBPM. Next, each of the mapped steps was specialised into a description comprising of African longitudinal population-based demographic and epidemiological surveillance concepts and terms. Input and output data records were also linked to their corresponding specialised steps/ sub-steps. Further, the information models for both the Pentaho data integration software and the DDI 4 specification were used to develop a structure for the generated metadata. The metadata content and structure were encapsulated in an XML metadata infusion file.

The first finding is that all the sub-processes involved in creating Specification 6.1 from an intermediate dataset could be mapped to the GLBPM. Second, a domain sensitive specialisation of the GLBPM applicable steps/ sub-steps was achieved. The third result is that input and output data records (pre and postconditions) were identified from the ETL for each of the ETL sub-processes and were linked to the GLBPM mappings. The fourth result is the addition of structure to the generated provenance metadata. The metadata were formatted in compliance with the proposed DDI 4 information model. All these results taken together represent an enhancement of CiB provenance high level metadata in a tool agnostic, structured and domain sensitive fashion.

This work contributes to the literature and professional practice in a number of ways. It represents a first attempt, in literature, at the characterisation of Pentaho ETL using generic process models within the demographic and epidemiological surveillance data harmonisation context. Relating to documentation standards, it adds to our understanding of structured documentation of data harmonisation.

The ALPHA data provenance documentation is being used to test future candidate versions of DDI. It is being used as a use case for the data management components proposed in the DDI 4 prototype information model. Consequently, decisions made during the course of this work regarding the structure of the provenance metadata have augmented the proposed DDI 4 process model and the DDI 4 data management functional view (Greenfield 2018; Greenfield, Kanjala, and Gregory 2019). The augmented model is shown in section 7.2, Figure 35. Hitherto, this process model was more inclined towards data collection

documentation with limited data transformation documentation capabilities (Greenfield, Kanjala, and Gregory 2019). In addition, the results from this ongoing testing will be submitted to the DDI scientific governance group in December 2019.

Further, in work going on in the DDI Alliance and beyond the remit of this thesis, parts of DDI 4 model have been incorporated into the latest version of DDI Lifecycle, version 3.3 to be released by the end of 2019. This is being done as the future roadmap of DDI remains to be determined. So DDI 3.3 has an information model and an RDF representation (Jay Greenfield – DDI Developer, personal communication). The big question is whether DDI can become free of its XML roots which restrict its ability to interoperate with other standards. This is being deliberately decided through a series of workshops.

In terms of practice, DDI 4 has provided a methodology for creating tool agnostic provenance metadata from the ALPHA Pentaho data harmonisation processes. This methodology could potentially be used for process reuse across specifications or network member studies. The expression of data transformation activities using a domain-specific business process model improves communication and understanding of the activities across the network and between network members and external domain experts wishing to use the data.

Similar to its predecessor, the Generic Longitudinal Business Process Model (GLBPM), the ADESBPM is valuable as a reference model through which implemented processes can be mapped, understood and compared (I. Barkow, Block, Greenfield, Gregory, Hebing, Hoyle, and Zenk-möltgen 2013). Such models are invaluable in ALPHA since the network is seeking to standardise data management and exchange across its members. Thus, a reference model can foster common understanding of ETL processes across the network.

The ADESBPM is a more nuanced description of the ALPHA ETL process compared to GLBPM. However, though more concrete and domain sensitive, it still inherits the limitations of the GLBPM regarding technical implementation: it is a conceptual model, not a technical implementation blueprint (I. Barkow 2016). And as such, it lacks the specificity required for direct operationalisation of the modelled activities in production systems. ADESBPM will also ultimately need to accommodate much more than the ALPHA processes. Accommodating ANDLA (ANDLA 2019) and other data specifications will be one task, but there are many others. The hope is that even though the learning process presented in this thesis has been bottom up, the model may have captured core business processes that pertain to the construction of records and datasets that represent event data

in general. This would make the ADESBPM applicable to the entire demographic and epidemiological surveillance domain.

The mapping to and specialisation of the GLBPM steps to local contexts is similar to efforts by national statistical offices around globe. These NSO's have mapped their data management processes to the GSBPM. However, unlike many of the NSO's, this project does not end with textual descriptions as specialisations, it adds structure to the specialisation. This is more in line with the ABS approach (UNECE 2018c) though not as elaborate as the ABS. In 2019 UNECE decided to grow the ABS approach by wedding the GSBPM with GSIM in a model where each GSBPM sub-process has GSIM information object inputs and outputs (Jay Greenfield – DDI Developer, personal communication). This is leading to more structured documentation when it comes to following the vicissitudes of data between sub-processes in the business process model. However, it is still the case that with this effort it remains to be the case that only textual descriptions of the actual transformations are supported.

Since this chapter has a high level perspective, it targets a different scope from VTL and SDTL which describe transformations at more granular details.

The ETL used is only for one specification, 6.1. This specification is predominantly on demographic surveillance. There is need to repeat the process using ETL for the specifications that cover epidemiological surveillance. This will ensure the methodology is representative of what is required for demographic and epidemiological surveillance data harmonisation processes. This is important because ultimately, the aim is to have a domain specific business process model - the ADESBPM - which should capture both demographic and disease surveillance data harmonisation activities.

Though the methods used here provide more concrete process description than the GLBPM due to the specialisation, they are still a simplification of the reality in the ETLs and they lack the specificity required to drive production systems. The next chapter takes a more granular view to address this specificity gap.

6. LOWER LEVEL, STRUCTURED METADATA FOR ALPHA DATA TRANSFORMATIONS:

Going the last mile - operationalising the business process model

6.1 Introduction

The previous chapter began to address the problems related to solely depending on tool specific provenance metadata for ALPHA data harmonisation routines. It focused on the development of high-level tool agnostic and structured metadata.

Two data provenance related aspects of particular interest to ALPHA and external users of the data remain unaddressed at the end of chapter 5. The first one is that the high-level metadata are silent about the changes happening to variables as the data are being transformed. The second one is that it is unclear how statistical indicators (data aggregates) created in the course of the harmonisation, such as data quality metrics, could be documented in a tool agnostic way.

Thus, the provenance metadata pertaining to variable level details of transformations and the statistical indicators of data quality are still tool specific. It is important for these too to be documented in more generic ways at a level more granular than that provided in chapter 5 to bridge the gap between modelling and implementation. Filling this specificity gap would aid ALPHA in its endeavour to produce an unbroken record of lineage for the harmonised data.

Existing literature on structured metadata for variable level data transformations can be classified into three approaches. These are (i) the approach taken in the current versions of DDI (DDI Alliance 2015b; Marker et al. 2009), (ii) the Validation and Transformation Language (VTL) and (iii) the Structured Data Transform Language (SDTL) (C2Metadata 2017).

DDI allows for a mixture of textual process descriptions and the inclusion of the source code in the documentation. Under DDI Codebook this is limited to the “*recoding and derivation*” element of the standard which allows for free text description of data transformations relating to a variable. It also allows the addition of the source code. DDI Lifecycle has much more structure for process descriptions compared to DDI Codebook, but in essence, they both leave users with proprietary software code as the description of the lower level data transformation detail. Thus, the disadvantages of using proprietary code for documentation are not overcome in this solution.

SDTL and VTL

Besides these DDI capabilities, literature has advanced our understanding on this topic in two fronts: the VTL and SDTL.

VTL and SDTL are promising methods for tool agnostic and structured documentation of granular details of data transformations such as those carried out in ALPHA. However, neither VTL nor SDTL have been applied before within the HDSS domain. In terms of provenance metadata for variable level transformations, this chapter's main contributions are to provide empirical results of testing out the SDTL for describing the variable level details of the ALPHA data harmonisation processes capturing both dataset level and variable details of the transformations.

Chapter 2 gave a comparison between VTL and SDTL and pointed out that using VTL we can produce scripts in various languages as output. In the case of ALPHA, the scripts are already available as either Stata setup files or Pentaho ETLs. ALPHA rather needs a translation of these into a software agnostic form. SDTL offers that facility since the C2Metadata project has developed a Stata parser for converting Stata to SDTL. Moreover, the same project is also working on expressing the SDTL in natural language (Ionescu 2018). What the C2Metadata project does not provide is the mapping between Pentaho transformations and steps and SDTL.

Structured documentation of statistical indicators

Relating to the documentation of statistical indicators and aggregate data, users of the indicators need to be able to trace back from a representation of an output, through the transformation process, to connect to the input microdata. Production of aggregated data as an output is common whenever data are validated and analysed, as such, aggregated data are commonplace in ALPHA workflows. These aggregates manifest as quality metrics, disease prevalence, incidence or other rates (mortality, fertility, sexual partner acquisition, marriage dissolution) among many.

The documentation of process outputs is one of the landscapes where DDI and SDMX could be jointly applied in a complementary manner as depicted in Figure 5 and by (Gregory and Heus 2007). Though both standards cater for both microdata and aggregates, DDI is more suitable for microdata while SDMX is more appropriate for aggregates.

Turning to the documentation of statistical indicators or other aggregated outputs, literature shows that the Statistics Data and Metadata eXchange (SDMX) standard, an ISO standard (ISO 17369) for describing statistical data and their metadata for efficient exchange and sharing (SDMX Technical Working Group 2018a) has been the main player. The realm

within which SDMX has been in use is official statistics. There is no evidence of its use for describing data aggregates generated from demographic and epidemiological surveillance data. This chapter investigates the use of SDMX for describing data quality metrics generated as part of ALPHA ETLs.

6.1.1 Aim

This chapter has two aims:

1. to characterise, at a more granular level, the ALPHA ETL tasks using SDTL
2. To explore the implementation of SDMX for documenting data quality indicators produced within ALPHA ETL.

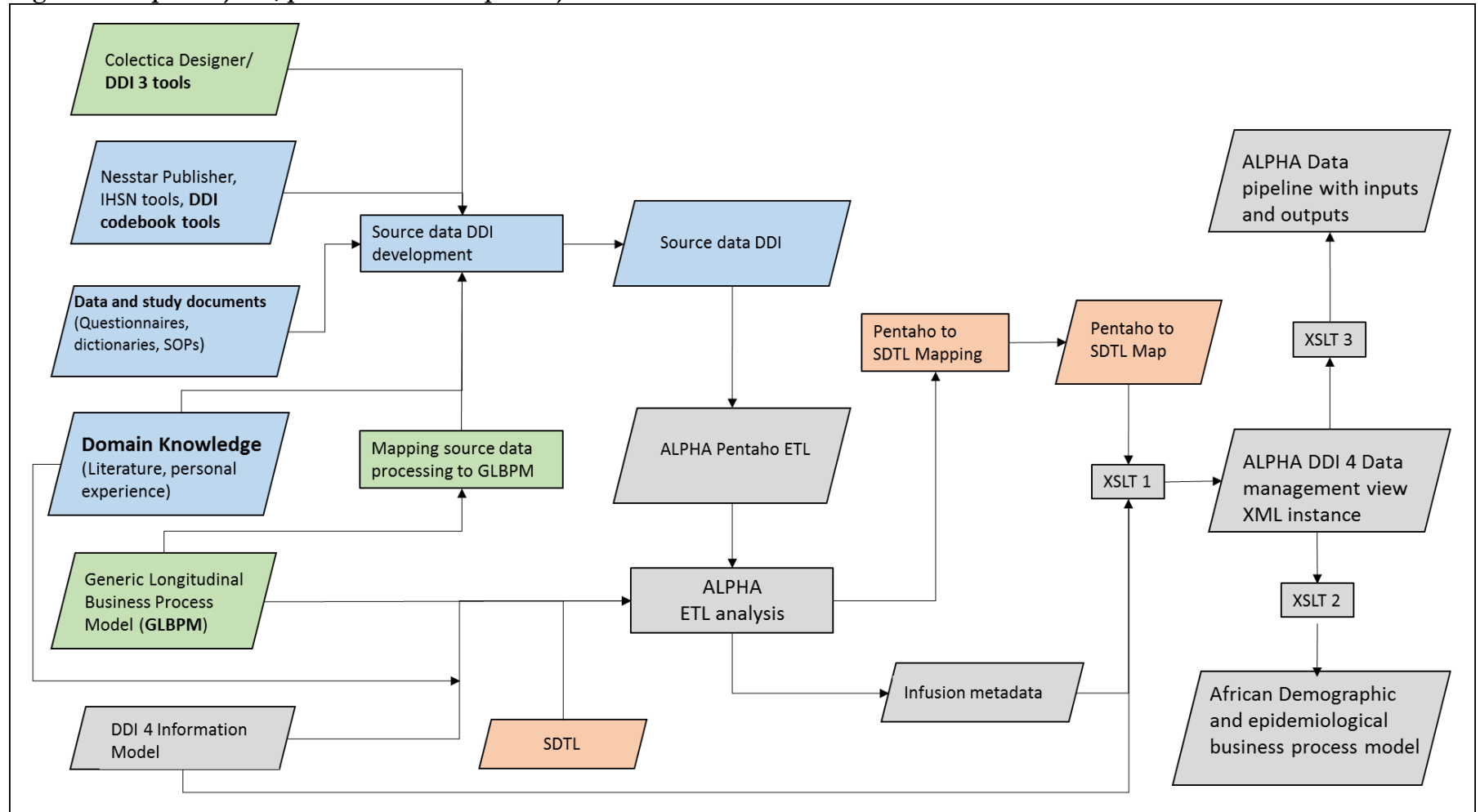
6.1.1.1. Objectives

- To identify SDTL elements that are equivalent to the Pentaho steps used within the data quality assessment sub-job in the ALPHA ETL.
- To assess whether there are Pentaho steps in the data quality assessment sub-job that do not have equivalents in SDTL
- To compile the Pentaho steps with no equivalents in SDTL, if any are found, for feedback to SDTL developers
- To document ALPHA 6.1 specification related data quality metrics for events, sex and date of birth data using SDMX.

6.1.2 The bigger picture

This chapter adds the orange objects in Figure 29 – *SDTL, Pentaho to SDTL map* and the *Pentaho to SDTL mapping* - to the objects in Figure 25. Here, the ETL is described in a more granular way to facilitate operationalisation. The individual tasks performed in the Pentaho steps are mapped to SDTL elements. SDTL scripts are then produced using this Mapping. The contents of the scripts could be infused into the DDI 4 data management view instance. In a nutshell, Chapter 5 provided a mechanism for prospective and retrospective description of the ETLs at a high level. In chapter 6, the granular details are documented after the ETL exercise. The contents of chapter 4, 5 and 6 taken together represent an unbroken record of data lineage for ALPHA specifications.

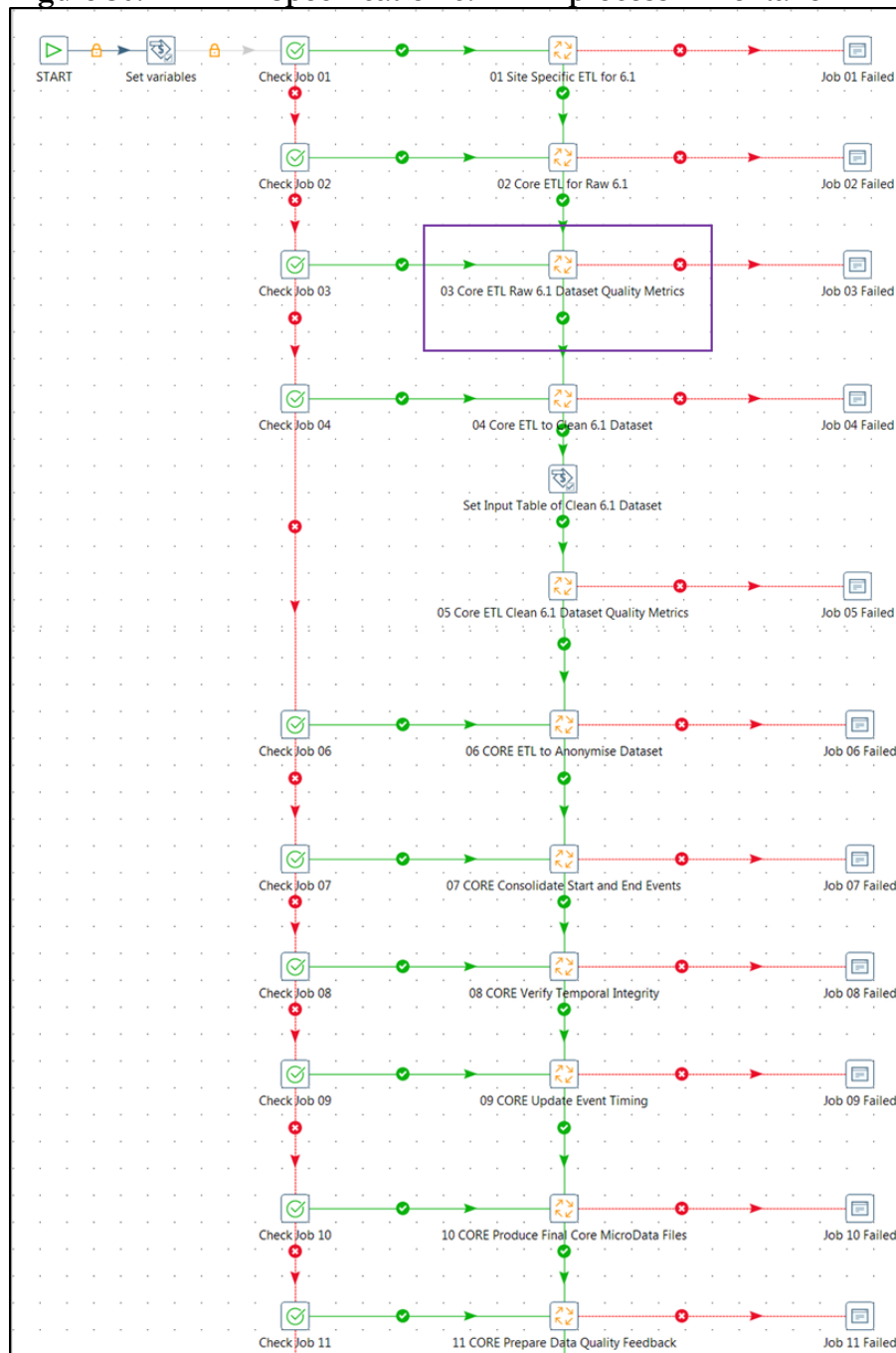
Figure 29: Input objects, processes and output objects in the ALPHA ETL and their structured documentation



6.1.3 Chapter overview

The next section is on the methods, it first describes the Pentaho job “002 CORE Data Quality Metrics” found in the ALPHA 6.1 specification’s ETL and used in this Chapter. It also describes the relevant aspects of SDTL and SDMX used to produce structured metadata for the process and outputs which include aggregates used to quantify data quality metrics. The methods are then followed by a presentation of the results and then summary and discussion sections.

Figure 30: ALPHA Specification 6.1 ETL process in Pentaho



6.2 Methods

Figure 30 shows a picture of the entire Pentaho ETL for creating ALPHA 6.1 specification. There is one master job for specification 6.1 for each of the ALPHA members. Within this master job, there are 11 sub-jobs depicted by the squares with orange arrows inside them. The first sub-job is labelled “01 Site specific ETL for 6.1” and the last one is “11 Prepare Data quality feedback”. The sub-job highlighted in purple represents the data quality assessment done after creating the specification. This is the sub-job considered in detail in this chapter to illustrate Pentaho to SDTL mapping and documentation of indicators.

Chapter 5 described the data quality assessment sub-job using an algorithm overview shown Table 14. The input for this sub-job is the raw 6.1 specification and the outputs are data quality indicators.

Table 14: Algorithm overview for 002 CORE Data Quality Metrics

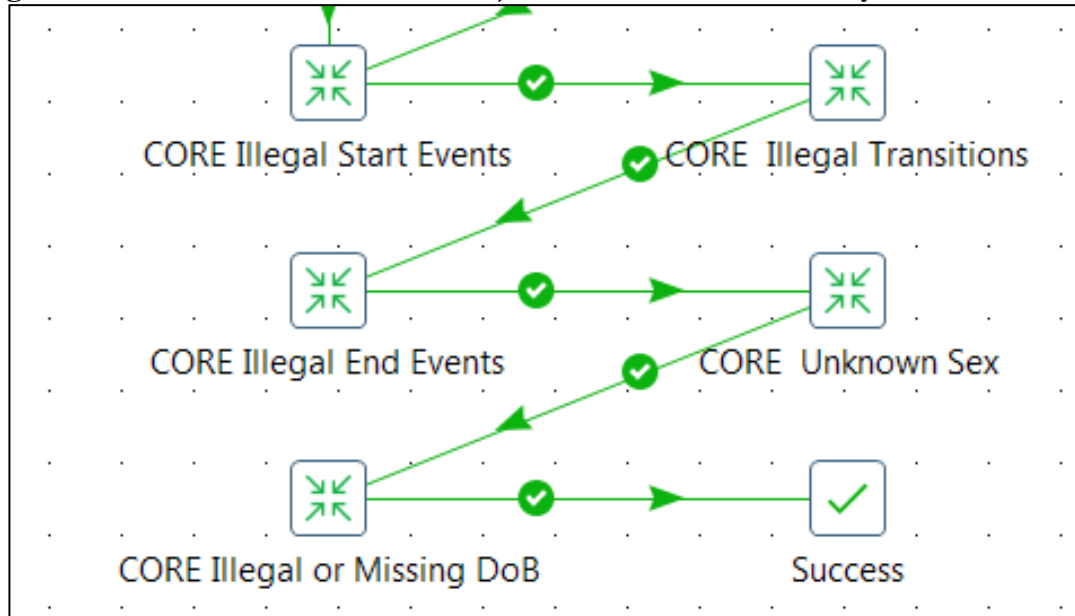
Sub-job name	GLBPM step	Algorithm overview
002 CORE Data Quality Metrics	5.3 Explore, validate and clean data	<ol style="list-style-type: none"> 11. Compile a list of quality metrics relevant to the data specification 12. Create events consistency matrix showing the logical ordering of event sequences 13. Identify in the data, events that start a residency episode (birth, external-immigration, enumeration, becoming eligible for a study, found after being lost to follow-up, Internal-immigration) 14. Identify in the data, events that end a residency episode (external-outmigration, death, became ineligible for study, lost to follow-up, internal-outmigration, present in the study (right censored)) 15. Review the identified start events and distinguish between legal and illegal ones 16. Review the identified end events and distinguish between legal and illegal ones 17. Review all transitions between two events and distinguish between legal and illegal ones 18. Compile illegal, missing or unknown sex 19. Compile illegal, missing or unknown DOB 20. Calculate numbers of legal and illegal start events, end events, event transitions, sex values, out of range DOBs and missing sex and DOBs

This chapter goes further to break the steps in the algorithm overview down to variable level transformations.

6.2.1 Data quality assessment sub-job analysis

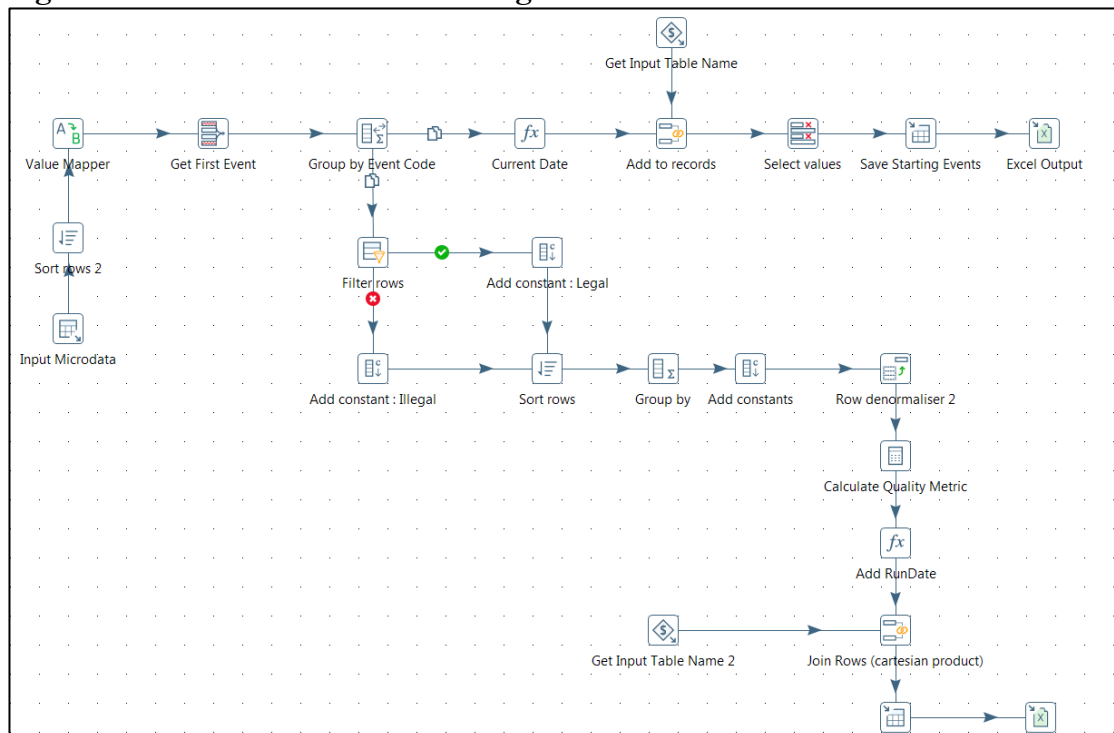
The data quality assessment sub-job is identified as “002 CORE Data Quality Metrics” in the Pentaho ETL and it has five “transformations” within it shown in Figure 31. A Pentaho transformation is roughly equivalent to a setup script for a statistical package such as a Stata do-file. While in a statistical package script there are commands, a Pentaho transformation comprises of “steps” and “hops”. The hops link consecutive steps and each step is an atomic task (a task that cannot be broken down any further). Figure 32 shows the steps in the

Figure 31: Transformations in the sub-job 002 CORE Data Quality Metrics



The nodes in the diagram are the steps and the lines are the hops.

Figure 32: Transformation CORE Illegal start events



6.2.2 Pentaho to SDTL Mapping

Each and every step in each of the five transformations shown in Figure 31 was analysed, and human intelligible description of these were prepared. This description would be useful to provide together with the more generic one coming from C2Metadata as it gives more context for the step under consideration. Next, the SDTL equivalent of each step involved was sought. The steps which had equivalents in SDTL were then mapped. Thereafter, SDTL code was developed from the mapping. This exemplar SDTL code is presented in a format called JavaScript Object Notation (Json) format ('JSON' 2019). Json is a data interchange format for exchanging data between platforms (Taylor 2014). The development of the Json code was merely for illustration of what code would be produced from the mapping. C2Metadata have Json schemas to use for automated development of the code. These would be used to generate SDTL once equivalents of the Pentaho steps have been identified and mapped.

I also compiled a list of the Pentaho steps for which I could not find an equivalent within SDTL to give as feedback to the language's developers.

6.2.3 Structured documentation for indicators

The quality indicators calculated in the data quality assessment sub-job include numbers of legal and illegal residency start and end events, event transitions, sex values and dates of birth. As stated in Chapter 2, SDMX has an information model that captures data structure, the metadata structure and data exchange related characteristics. The work presented here is only concerned with describing the data structure of the data quality metrics. This structured documentation of aggregates is complementing the rest of the metadata being provided for the harmonised microdata.

The data structure is being used to define the characteristics of the data of interest by identifying and defining SDMX concepts, dimensions, and data points (observations). (Stahl and Staab 2018). A DSD for the quality indicators was developed by listing relevant concepts underlying the indicators, the involved dimensions, their types (dimension, time, measure) and key values (either categorical or discrete) associated with each dimension. The data points were represented as unique combinations of dimensions and key values.

6.3 Results

6.3.1 Mapping Pentaho steps to SDTL

The total number of steps in the transformations used in the data quality assessment sub-job was 154. Of this total, several steps were applied repeatedly. The number of distinct steps was 21. 17 of the steps had equivalents in SDTL. No equivalents were found for 4 steps. Table 15 shows the 17 steps, their textual descriptions and their equivalents in SDTL.

Table 15: Pentaho steps and their SDTL equivalents

Step name	Step description	Step type	SDTL
Input Microdata	Imports Raw ALPHA specification 6.1 in Event Format data from a staging database	TableInput	Load
Value Mapper	Map numerical event codes to string ones suggestive of the event: e.g, 1 to ENU - for enumeration, 2 to BTH for a live birth etc	ValueMapper	Recode
Get First Event	Remove duplicates in terms of ids to remain with only the first occurrence of the id number and first event (since previously sorted by id and eventnr)	Unique	Select
Current Date	Adds current date to the data stream	Formula	Compute
Select values	Select variables to output and provide metadata for the eventcode variable	SelectValues	Select
Excel Output	Send a table to excel with a record of starting events	ExcelOutput	Save
Filter rows	Separate between legal and illegal starting events	FilterRows	Select
Add constant : Illegal	Catch numbers of individuals with illegal start events from the filter and add a column stating that they are illegal events	Constant	Compute
Group by	Calculate the total numbers of illegal and legal start events	GroupBy	Aggregate
Calculate Quality Metric	Calculate the total number of start events by adding legal and illegal events and the percentage of this total that are illegal events	Calculator	Compute
Save Quality Metrics	Send the starting events quality metrics table for a particular site to a staging database	TableOutput	Save
Microsoft Excel Writer	Create an excel table of the starting events quality metrics for a particular site	TypeExitExcelWriterStep	Save
Get Next Event	move backwards (by lagging n steps) or forward (by leading n steps) across ordered rows	AnalyticQuery	
CrossTab with NextEvent	Cross tabulation of an event and its corresponding next event	Denormaliser	ReshapeWide
If field value is null	Replace NULL values with 0 for the specified column/variable	IfNull	Dolf
Get CurrentDate	Extracts the current date from computer's system	SystemInfo	Compute
Status=Before 1850		SetValueConstant	Dolf

The 4 steps for which no equivalents were found in SDTL are the Sort Rows, Get variable, Join Rows (Cartesian product) and Dummy (Do Nothing) steps. The Sort Rows step sorts the input data by specified variables. The Get Variable step calls a named input from Pentaho user-defined or system built-in variables. The Join Rows step which creates a Cartesian product for two specified arrays was used in the ALPHA ETL to create tabulations of quality

metrics. The Dummy (Do Nothing) step was used as a place holder to catch and or unify data streams before their next step.

6.3.2 Development of SDTL code from the mapping

The steps Input, Value Mapper, Sort, Select, Recode, Denormalizer and Save were developed into SDTL code to illustrate the translation of Pentaho to SDTL. The code in json format for the Recode, ReShapeWide and Save commands are shown in Figure 33 and Figure 34.

Figure 33: Exemplar SDTL for the recode step translated to its namesake in SDTL

```
32 {
33   "command": "recode",
34   "sourceInformation": {
35     "lineNumberStart": 1764,
36     "lineNumberEnd": 1827,
37     "originalSourceText": " <step> \n <name>Value Mapper</name> \n <type>ValueMapper</type>
38   },
39   "recodedVariables": [
40     {
41       "source": "event",
42       "target": "EventCode"
43     }
44   ],
45   "rules": [
46     {
47       "fromValue": [
48         "1"
49       ],
50       "to": "ENU"
51     },
52     {
53       "fromValue": [
54         "2"
55       ],
56       "to": "BTH"
57     },
58     {
59       "fromValue": [
60         "3"
61       ],
62       "to": "IMG"
63     },
64     {
65       "fromValue": [
66         "4"
67       ],
68       "to": "ENT"
69     },
70     {
71       "fromValue": [
72         "11"
73       ],
74       "to": "OBS"
75     },
76     {
77       "fromValue": [
78         "12"
79       ],
80       "to": "DTH"
81     },
82     {
83       "fromValue": [
84         "13"
85       ],
86       "to": "OMG"
87     },
88     {
89       "fromValue": [
90         "14"
91       ],
92       "to": "EXT"
93     },
94     {
95       "fromValue": [
96         "15"
97       ],
98       "to": "OBL"
99     },
100  ],
101 }
```


SDTL model is setup to declare commands without equivalents in SDTL as unsupported. Thus, within SDTL code one can list all their commands from their platform and tag the ones with no equivalents as unsupported. This is the case with Sort commands for example. Exemplar SDTL for the recode step translated to its namesake in SDTL. The original source text was truncated for the image to fit on the page.

Figure 34: Reshape wide and save commands in SDTL

```

117 {
118   "command": "ReshapeWide",
119   "sourceInformation": {
120     "lineNumberStart": 1471,
121     "lineNumberEnd": 1530,
122     "originalSourceText": " <step> \n <name>Row denormaliser 2</name> \n <type>Denormaliser</ty
123   },
124   "KeepItems": [
125     {
126     },
127   ],
128   "IdVar": "QMetric",
129   "IndexVar": "Start"
130 },
131
132 {
133   "command": "save",
134   "sourceInformation": {
135     "lineNumberStart": 887,
136     "lineNumberEnd": 978,
137     "originalSourceText": " <step> \n <name>Excel Output</name> \n <type>ExcelOutput</type> \n
138   },

```

6.3.3 Structured documentation of data quality indicators

The data quality metrics were for residency start events, events transitions, residency termination events, sex values and date of birth values. The measures include the number of illegal events, legal events, total events, percentage of total events that are illegal.

Table 16 lists the identified concepts and their definitions.

Table 16: SDMX concepts for quality metrics

Concepts	
Concept	Definition
Centre ID	The centre which produced the data being examined
Metric Table	The table being examined
Run Date	The date on which the table was examined
Quality Metric	The aspect of the data being assessed
Illegal	Count of illegal cases
Legal	Count of legal cases
Total	Total number of cases
Metric	The percentage of illegal cases out of the total
Type of Measure	SDMX mechanic used in tables that contain multiple measures

Table 17 shows the dimensions of the aggregates. In Table 18, all the data points and their keys are provided. The keys show what dimensions were combined to what measure to obtain a particular observation.

Table 17: Dimensions for the quality metrics

Dimensions			
Dimension	Dimension Type	Key Value Type	Dimension Concept
Centre ID	Dimension	Uncoded - String	Center ID
Metric Table	Dimension	Uncoded - String	Metric Table
Run Date	Time	Time Stamp	Run Date
QMetric	Dimension	Coded – (Start Date, End Date, DoB Values, Sex Values, Transitions)	Quality Metric
Type of Measure	Measure	Coded - (Illegal, Legal, Total, Metric)	Type of Measure

Table 18: Keys and data points for the quality metrics

Observations	
Key (Centre ID + Metric Table + Run Date + QMetric + Type of Measure)	Observation Value
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:29 EAT + Start Date + Illegal	13662
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:29 EAT + Start Date + Legal	43108
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:29 EAT + Start Date + Total	56770
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+14/06/2018 10:45:29 EAT + Start Date + Metric	24
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:31 CAT + Transitions + Illegal	44296
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:31 CAT + Transitions + Legal	128298
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:31 CAT + Transitions + Total	172594
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:31 CAT + Transitions + Metric	25
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:33 EAT + End Date + Illegal	55072
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:33 EAT + End Date + Legal	1698
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:33 EAT + End Date + Total	56770
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:33 EAT + End Date + Metric	97
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:36 EAT + Sex Values + Illegal	0
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:36 EAT + Sex Values + Legal	172594
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:36 EAT + Sex Values + Total	172594
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:36 EAT + Sex Values + Metric	0
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:38 EAT + DoB Values + Illegal	0
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:38 EAT + DoB Values + Legal	172594
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:38 EAT + DoB Values + Total	172594
ALPHA011 + ALPHA011_Raw_61_Event_Format_Version1+ 14/06/2018 10:45:38 EAT + DoB Values + Metric	0

6.4 Discussion

This chapter set out to contribute towards addressing two problems. The first is our incomplete knowledge regarding documentation of data transformations details beneath the reach of business process models such as those considered in chapter 5 (the specificity gap). The second is the extension of the use of the SDMX standard into the domain that ALPHA represents. Transformation tasks performed in Pentaho ALPHA ETL were mapped to SDTL and exemplar SDTL code was manually generated using additional metadata provided in Pentaho. Regarding the documentation of quality indicators, the values of the indicators were presented using SDMX elements - concepts, dimensions and measures. All the data points representing quality indicators for Start events, Transition events, End events, Sex values and Date of birth, were characterised using SDMX keys which comprised of the appropriate combinations of concepts and dimensions.

The results suggest that SDTL can describe the majority of the transformation tasks involved in ALPHA ETL. The data quality assessment sub-job used 21 distinct Pentaho transformation steps, after discounting for steps that were repeated used. Of the 21, only 4 steps could not be mapped. The results also showed that there was sufficient input in the Pentaho ETL to use for the development of SDTL code from the mapping, this suggests that the conversions from Pentaho to SDTL can be automated.

To the best of my knowledge, until its exploration in this chapter, SDMX had never been used in the African longitudinal population-based epidemiological and demographic surveillance. Though novel, its successful implementation for annotating the ALPHA data quality indicators is not surprising as the statistical methods used for presenting the indicators – crosstabulations – are also commonly used in official statistics.

However, the ground breaking aspect of both the results on SDTL mapping and SDMX implementation is that, at least for ALPHA and probably for many players in the demographic and epidemiological surveillance domain, these structured metadata for data transformations and indicators have been produced for the first time in this context. The implications of these results are far reaching.

Probably the biggest impact from ALPHA's perspective is related to process transparency. The structured documentation sets the stage for exploitation of the metadata by suitably programmed software. The software programs can be guided by the metadata to search and build provenance chains that link data points in outputs, regardless of whether these outputs are microdata or aggregates, to the inputs via the data transformations documentation. With

the availability of lineage records, process comparisons and standardisation across the network will become more feasible.

Extending the use of SDTL to the description of ALPHA Pentaho ETL means that we are closer to the possibility of initially designing an ETL process in Pentaho and then reusing it in statistical package or vice versa thus increasing flexibility in reuse of code.

This work also represents the liberation of Pentaho ETL metadata. Though Pentaho has a native, easy to understand, graphical interface that shows diagrams linking steps via hops as shown in Figure 31 and Figure 32, the metadata underlying these steps are hidden from the user and entangled within Pentaho specific XML files. The efforts in this chapter work towards extracting from those XML files only the contents essential for defining the lineage of the data. This implies leaving behind the functional metadata that responsible for display and other mechanics of how the interface works which do not add any value to the data lineage aspect outside Pentaho.

7. PROVIDING END USERS WITH ACCESS TO ALPHA PROVENANCE METADATA

ALPHA METADATA BROWSER USER REQUIREMENTS ANALYSIS

7.1 Introduction

Chapters 5 and 6 explored a means of developing standards-based documentation for ALPHA data harmonisation processes on top of the CiB's tool-specific metadata. These provenance metadata make the data to be better understood by users independent of knowledge of the production tools. The developed metadata are in XML format. This format is flexible for computer programs to manipulate the metadata and data. However, for human end users, the full utility of the developed metadata lies in the availability of user-friendly tools (Vardigan, Heus, and Thomas 2008) for browsing and searching the metadata and the harmonised datasets. Such tools are unavailable off-the-shelf. In order to build them, software developers need domain experts' perspectives on the desired functionality to guide their work. This study sought to perform a requirements analysis for an ALPHA provenance metadata browser from eliciting and synthesising requirements from experts in ALPHA and the CLOSER (Cohort & Longitudinal Studies Enhancement Resources) project (<https://www.closer.ac.uk/>)(O'Neill et al. 2019).

7.1.1 Objective

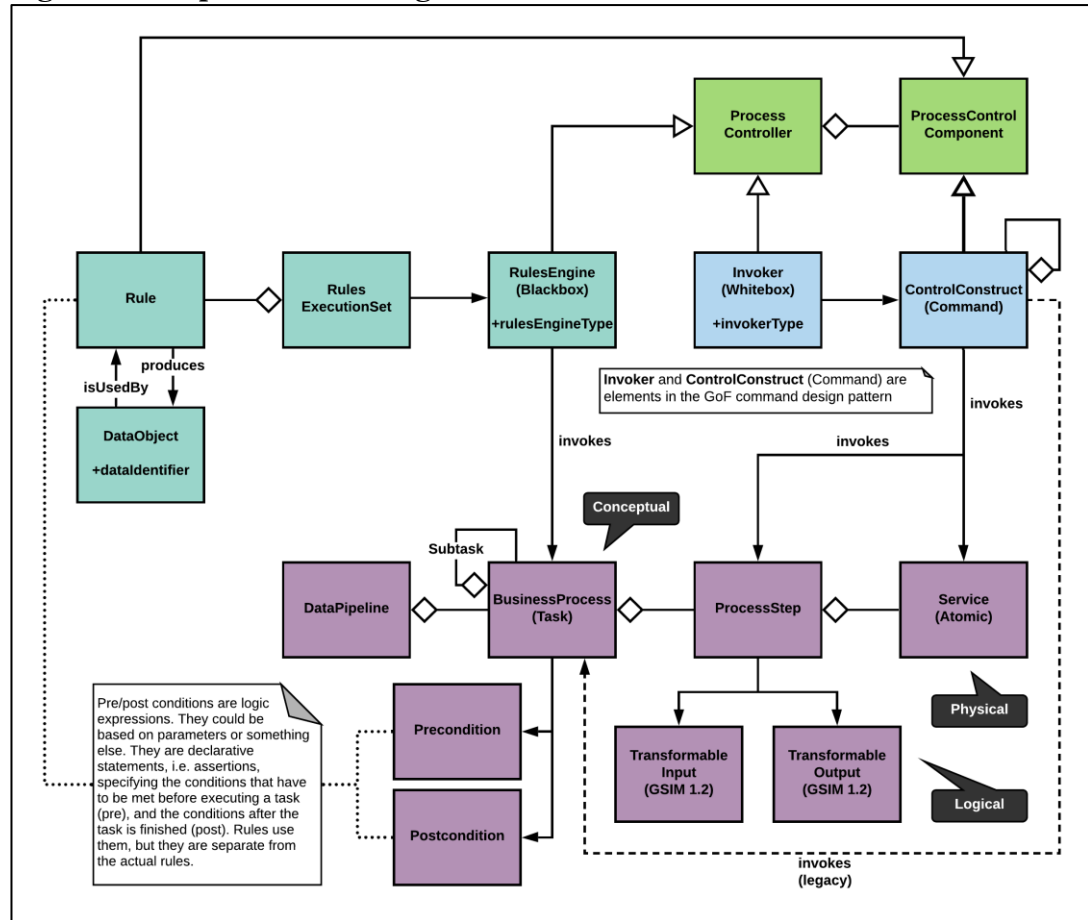
To gather and analyse the requirements for a user-friendly provenance metadata platform for ALPHA datasets.

7.2 Methods

7.2.1 Development of mock-up diagrams for use in elicitation study

As alluded to in chapter 5 in the discussion section, and in (Greenfield, Kanjala, and Gregory 2019), the developed ALPHA data harmonisation metadata served to augment the initial version of the proposed DDI 4 process model. In its original form, the process model was more inclined towards data collection, with limited data management documentation capabilities. This improved process model (Figure 35) and the metadata developed in Chapter 5 provided the inputs for the mock-up diagrams used in the requirement elicitation study. Consequently, the names of the displayed features are derived from components of this process model.

Figure 35: Proposed DDI 4 augmented Process model



Source: (Greenfield, Kanjala, and Gregory 2019)

The mock-up diagrams, comprising of the proposed metadata browser features, were the core of the interview discussions.

7.2.2 Mock-up diagrams of the proposed features: The details

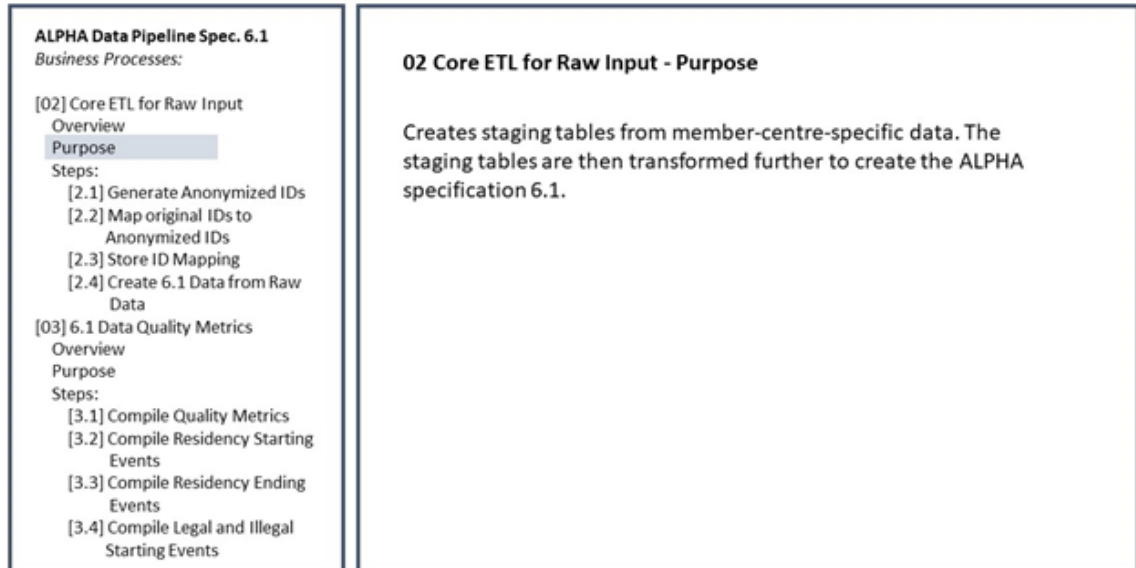
The entire process of creating a specific ALPHA dataset for a particular member institution of the network is called a *data pipeline* for that dataset. In the mock-up diagrams, the data pipeline for ALPHA specification 6.1 – residency data in Event format was used following on from the use of the same data specification in chapters 5 and 6.

Figure 36 to Figure 41 show the mock-up diagrams. The proposed features are described under four broad sub-groups (i) data pipeline, task, overview and purpose, (ii) task steps, concepts and setup scripts, (iii) task-centric view and (iv) dataset-centric view.

Data pipeline, task, overview and purpose

Figure 36 worked as an entry point for the description of the mock-ups serving the purpose of introducing the data pipeline, the tasks within a data pipeline also called *business processes* and the breakdown of each task into its *overview*, its *purpose* and its individual *steps*.

Figure 36: Data pipeline, its constituent tasks and their details



The left panel in the mock-ups listed the various components of a data pipeline while the bigger, right panel expounded on a highlighted feature providing its description and relevant links.

Task step, concepts and setup scripts

Each task within a data pipeline was broken down into *steps* (business steps). Figure 37 and Figure 38 show relevant metadata associated with a step within a task. These include a description of that step, the concepts related to that step and setup scripts containing the actual code used to perform the step. Though not shown in the picture, the original code or data transformation service (in the case of transformation performed in an ETL platform such as Pentaho) would then be described using SDTL.

Figure 37: Metadata elements for describing steps in a task

<p>ALPHA Data Pipeline Spec. 6.1 <i>Business Processes:</i></p> <p>[02] Core ETL for Raw Input Overview Purpose Steps: [2.1] Generate Anonymized IDs [2.2] Map original IDs to Anonymized IDs [2.3] Store ID Mapping [2.4] Create 6.1 Data from Raw Data</p> <p>[03] 6.1 Data Quality Metrics Overview Purpose Steps: [3.1] Compile Quality Metrics [3.2] Compile Residency Starting Events [3.3] Compile Residency Ending Events [3.4] Compile Legal and Illegal Starting Events</p>	<p>[3.2] Compile Residency Starting Events</p> <p>Description: Identify in the data, events that start a residency episode (birth, external-immigration, enumeration, becoming eligible for a study, found after being lost to follow-up).</p> <p>Concepts: This algorithm step references the following study concepts: residency birth migration</p> <p>Related Files: CompResSt.do (STATA Executable File) ResStart.sps (SPSS Executable File)</p>
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

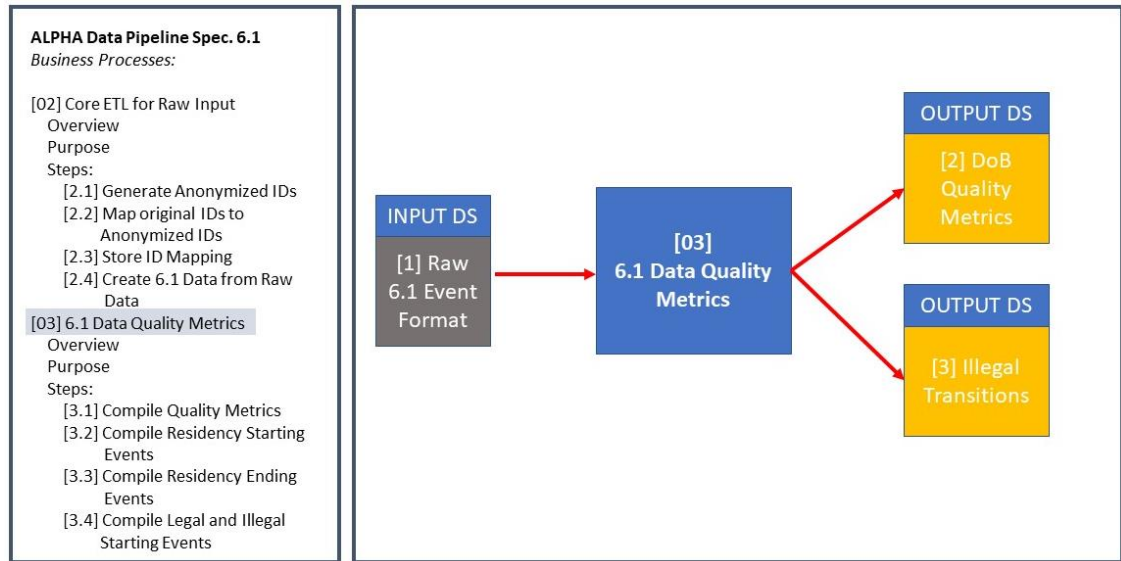
Figure 38: Provision for definition of concepts related to a step in a task

<p>ALPHA Data Pipeline Spec. 6.1 <i>Business Processes:</i></p> <p>[02] Core ETL for Raw Input Overview Purpose Steps: [2.1] Generate Anonymized IDs [2.2] Map original IDs to Anonymized IDs [2.3] Store ID Mapping [2.4] Create 6.1 Data from Raw Data</p> <p>[03] 6.1 Data Quality Metrics Overview Purpose Steps: [3.1] Compile Quality Metrics [3.2] Compile Residency Starting Events [3.3] Compile Residency Ending Events [3.4] Compile Legal and Illegal Starting Events</p>	<p>[3.2] Compile Residency Starting Events</p> <p>Description: Identify in the data, events that start a residency episode (birth, external-immigration, enumeration, becoming eligible for a study, found after being lost to follow-up).</p> <p>Concepts: This algorithm step references the following study concepts: residency birth migration</p> <div data-bbox="877 1198 1236 1444"> <p>The change of residence by a registered individual or social group (e.g., a household). There are two types of migration that occur among the registered population. These are internal and external migration.</p> </div>
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Task-centric view

The view in Figure 39 aims to show the relationship between a particular task, its input datasets (INPUT DS) and its output datasets (OUTPUT DS). It was used to elicit perspectives of interviewees on whether they wanted to see this association as part of provenance metadata.

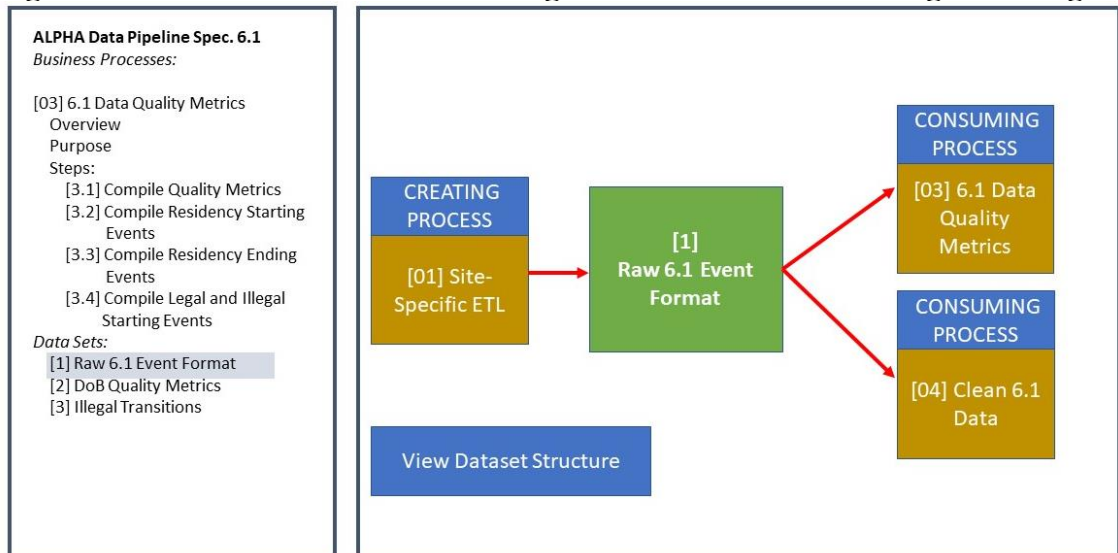
Figure 39: Task-centric view showing a task and its input and output data



Dataset centric view

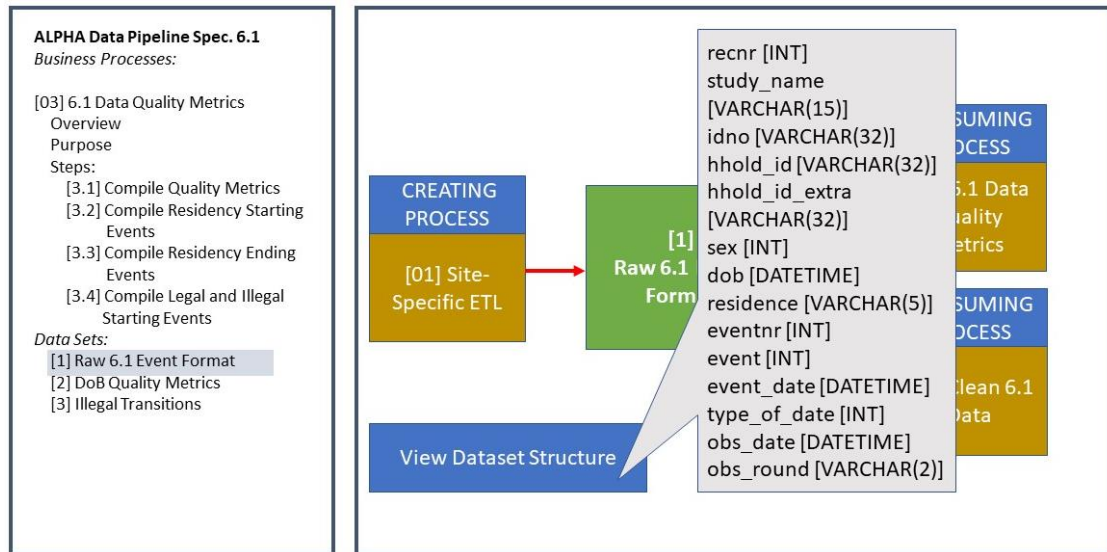
The data centric view aimed to present provenance metadata relating to a particular dataset. The diagram in Figure 40 shows association between a dataset (Raw 6.1 Event Format), a task(s) that create a the dataset (CREATING PROCESS) and task(s) that use the dataset as an input (CONSUMING PROCESS)

Figure 40: Dataset centric view – showing a dataset and tasks creating and using it



The last mock-up in Figure 41 shows variable level details for the dataset of interest in form of the names of the variables and their types.

Figure 41: Dataset-centric view: Dataset structure (variable name and type)



These mock-ups formed the core of a question guide used during the data collection described in section 7.2.4.

7.2.3 Recruitment of study participants

A convenience sample of 10 participants was drawn from data managers and researchers affiliated to the ALPHA and CLOSER projects. The sampling aimed to represent organisation diversity and depth of work experience by drawing among professionals who have a number of years of experience in data management and research. The ALPHA participants, as the producers and or internal users of the harmonised data, provided the viewpoint of users who are familiar with the specifics of the data harmonisation process. On the other hand, interviewees affiliated to the CLOSER project provided the viewpoint of experts familiar with the metadata standards and data harmonisation but not familiar with ALPHA data production. The CLOSER project has successfully conducted an ongoing data harmonisation project involving eight UK birth cohorts. These two groups of users, were considered suitable for identifying the requirements of both internal and external users.

7.2.4 Data collection

The data collection consisted of background material reading and a recorded Skype interview. An information pack was emailed to the study participants prior to the interview. In this pack there were the following items - (1) a study background document (APPENDIX C), (2) an information sheet (APPENDIX D), (3) a consent form (APPENDIX E) and (4) a question guide comprising of the 6 mock-up diagrams of the proposed features and accompanying questions (APPENDIX F).

Skype interview: Each participant was interviewed over Skype on the features in the mock-up diagrams using the semi-structured question guide. The participants graded each feature's importance on a provided scale and gave the rationale for their grading. Further, they listed any desired features not included in the mock-ups. All the interviews were recorded and transcribed verbatim.

7.2.5 Data analysis

Descriptors used to group the textual responses were primarily decided on in advance of the interviews, also called a priori coding (Lazar, Feng, and Hochheiser 2017). The codes were in the form of the proposed features and their groupings as shown in Figure 42. Data analysis followed steps commonly used in qualitative analysis. First, the recorded interviews were transcribed verbatim into MS Word documents. Second, the transcribed interviews were read end to end. This was followed by labeling of sections of texts in the transcriptions with descriptors from the coding scheme alluded to. The coding involved identification of scores allocated to the various features and the reasoning for or against the features. Features not in the mock-ups but perceived as vital were added to the requirements list using a prioritisation criteria. The criteria considered the perceived complexity of the task of adding the feature versus the time and funding resources. The coding was done using the NVivo software (*NVivo Qualitative Data Analysis Software* (version 12) 2018).

7.3 Results

In this results section, education and work experience characteristics of the participants are presented, the scoring of the various proposed features, the rationale for the scoring, some overarching aspects, feedback requiring structural changes to the metadata schema and cultural changes required among metadata producers.

7.3.1 Organisational diversity, education, work experience and roles of interviewees

Table 19 shows education and work experience levels for the interviewees. Their names and institutions were replaced with a numerical and alphabetical letters for privacy and confidentiality. Of the 10 interviewees, five were working in capacities involving both data management and research, two were researchers and three were data managers. Eight interviewees were affiliated to ALPHA (member institution or secretariat) while two were affiliated to CLOSER.

Figure 42: A priori Coding Scheme based on DDI 4 Process model and proposed features

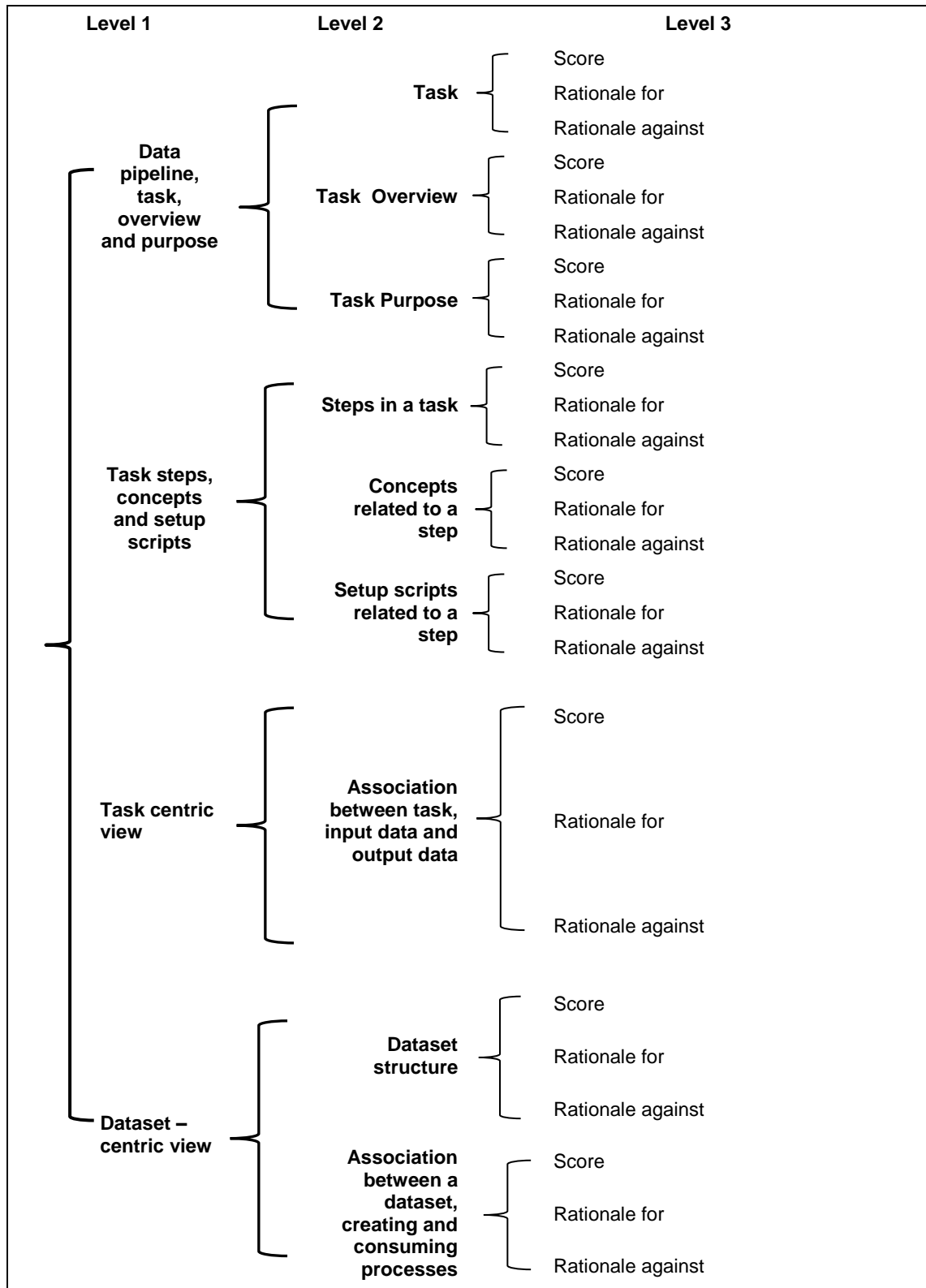


Table 19: Work experience of the interviewees

Interviewee	Institution	Education, Data management /Research experience	Role
1	A	PhD, More than 10 years of experience in data management and research	Both
2	B, I	MSc, More than 10 years of experience in data management and research	Both
3	C	PhD, and More than 20 years of experience in data management	Data management
4	D	MSc, More than 10 years of experience in data management and research	Both
5	E	PhD, Over 20 years of research experience in epidemiology	Researcher
6	F	PhD, Over 10 years of research experience in epidemiology	Researcher
7	G	MSc, 2 years data management experience and 3 years' software development	Data management
8	H	BSc, About 8 years' experience in research data management	Data management
9	I	PhD, Over 15 years research and data management experience	Both
10	I	MSc, Over 20 years data management experience and research	Both

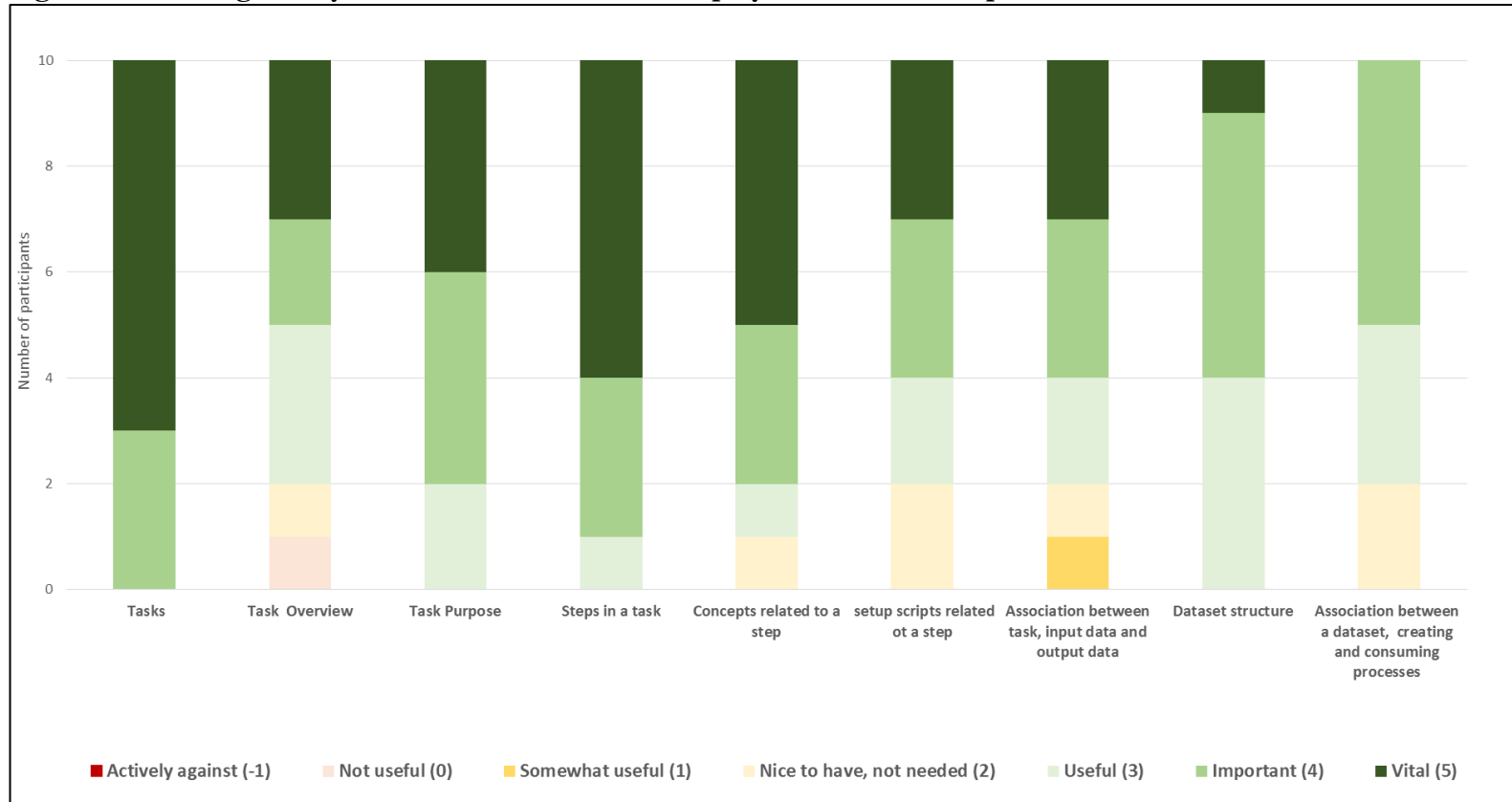
7.3.2 Scores for proposed features

Figure 43 gives the number of interviewees who gave a particular score for each of the features that were included in the mock-up diagrams. The results show that the interviewees generally considered the proposed features to be important for documenting the harmonised datasets. The scores for the feature called tasks in a data pipeline (*Tasks*) had the least variation among respondents, three respondents considered it important (4) while seven considered it vital (5). On the other hand, the *Task overview* feature had the most variation with scores ranging from “not useful” (0) to vital (5). The scoring was generally diverse, with 5 out of the nine graded features having scores ranging at least from as wide as Nice to have, not needed (2) to vital (5). Though the scores give an indication of the relative importance attached to the features by the respondents, it is only a partial picture if the reasoning behind the scoring is not considered. Further, the variation in these scores makes the partial picture more obscure. To help clarify the perspective of the participants, their rationales for choosing the scores as they did are presented next.

7.3.3 Rationale for or against having proposed features and suggested improvements

Table 20 gives a summary of the rationale provided for or against each of the proposed features. The content in this table relates to responses directly relating to the proposed features. Also included in the table, are the suggestions for improving the features.

Figure 43: Scores given by interviewees for features displayed in the mock-ups



No respondent registered being actively against any of the proposed features – as there is no feature that had the score “Actively against” (-1). In the majority of the cases, it was rather that the respondents wanted the features to be developed further. Examples of this include the suggestion to integrate *Task overview* and the *Task purpose* feature into one feature, the suggestion to have the dataset –centric and the task centric views together in one diagram at the data pipeline level and the suggestion to develop concepts further into site-specific concepts.

Table 20: Rationale for having and against proposed features and suggested improvements

Group	Feature	Rationale for having the feature	Rationale for not having the feature	Improvements or other comments
Data pipeline, task, overview and purpose	Task	Name and description can give summary of what task involves.		Make the language more generic
	Task Overview	Might guide for those less familiar with the subsystem and data	Not adding much value in the presence of steps and purpose	Needs to be integrated with purpose though difficult to judge it without seeing content.
	Task Purpose	Helps to communicate why a task is carried out		Could be integrated with overview
Task steps, concepts and setup scripts	Steps in a task	Seen as a good way to manage and transfer knowledge. Helps with high level understanding of the task	Considered by some interviewees to be less detailed than needed	Tasks have different levels of difficulty therefore some will require more detail than others Supplement with variable level metadata for targeted outputs Include site-specific rules, assumptions and cut offs (methodology).
	Concepts related to a step	Help to understand steps. Terms have different meanings so defining concepts reduces ambiguity.	Some data managers did not consider these to be essential and making real difference in their work. They considered them to be redundant in the presence of steps descriptions	Make them site and dataset specific to capture variations across sites. For instance, capturing how migration, residency, HIV Status, marriage/ cohabitation etc are defined in a particular dataset? Beyond concepts also capture assumptions, rules and cut offs applied to a dataset of interest.
	Setup scripts related to a step	Needed for reproducibility of processes and to allow alterations of cut offs, assumptions and rules, to alter the data specification. Internal users at member institutions and secretariat will need these for full understanding and reuse	Researchers not primarily interested in looking through setup scripts. There may be intellectual property rights concerns. Some producers may request payment for sharing their setup scripts. Most site specific scripts were prepared without sharing in mind, consequently they are not easy to follow.	Address queries requiring looking at the scripts on a case by case basis. Removal of redundant code leaving only code relevant to the task step of interest.
Task centric view	Association between task, input data and output data	Picture was considered more appealing than textual description found in task overview.	Redundant given that the overview contents overlaps with this picture.	Instead of doing a diagram for each task, make one overall diagram giving a view of associations between tasks and input and output datasets for the entire data pipeline.
Dataset – centric view	Dataset structure	Structure shows the variables in input or output datasets for a task, useful details in tracking changes	Currently has less variable level and dataset detail than required.	add more variable level details: labels, value codes and their labels for categorical variables, date formats among other details Provide for browsing structure in Stata/ Excel or other not in metadata browser
	Association between a dataset, creating and consuming processes	Its content is more in line with what goes into a protocol/ SOP, gives an idea what the pipeline comprises of	Too much detail and complexity	Could be captured in the overall overview diagram for a pipeline.

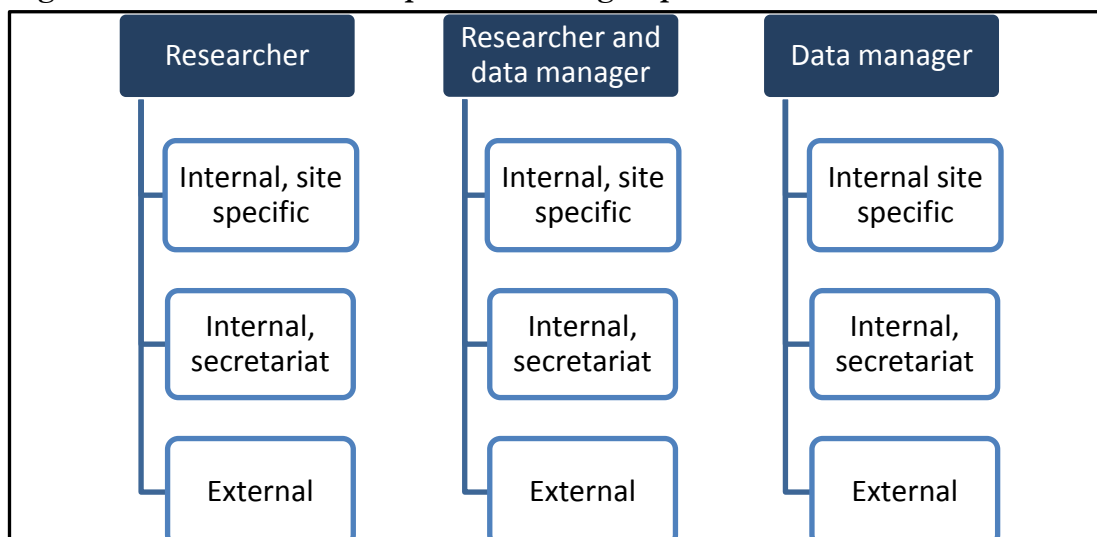
There was a perspective that particular attention should be paid to the navigation mechanism of the metadata browser. This is needed in order to manage detail and complexity of the metadata. Potential users will have different needs. Care will therefore need to be taken to make the metadata as complex and as detailed as it needs to be, layering the metadata from high level to more granular content, with granular content accessible to the user on demand.

7.3.4 Overarching aspects

All respondents expressed a general positive sentiment towards the approach being proposed for the ALPHA provenance metadata browser though the degree of interest in individual features varied among the participants. Besides the role of facilitating the communication of the data properties with third party users external to ALPHA, the respondents held the view that these metadata had a role to play in knowledge management and knowledge transfer in the high data staff turnover situations that partner institutions experience.

As expected, two broad user experiences have emerged from the results, the data manager experience and the researcher experience Figure 44. The two user groups have expressed different needs which will need to be explicitly catered for in the metadata browser.

Figure 44: Metadata browser potential user groups



While the researcher role generally requires high level documentation, the data managers have expressed need for more granular details. Five of the participants work in both data management and researcher roles (Table 19). The broad groups are further divided into potential users at the partner institutions, those working within the ALPHA secretariat and those external to ALPHA. Figure 44 provides a summary of the potential user groups alluded to in the interviews.

The respondents found the language highly specialised to the data integration and HDSS context and recommended revision into a more generic tone. Terms such as *ETL*, *business process*, *data pipeline*, and *illegal transitions* are among those which presented challenges to the participants.

The mock-ups used in the interviews were only on the metadata relating to the data harmonisation processes. The respondents suggested an integration of the data processing metadata with the metadata for the source data generated in the partner studies and those for the final harmonised datasets. The interviewees expressed the desire to ultimately have a documentation pathway linking a variable in the ALPHA harmonised dataset to the variable(s) in the questionnaires used to collect the primary data at the partner sites. In expressing these needs, they also acknowledged the potential difficulties involved. The stated challenges include the handling of the changes in data collection instruments/ data management practices or rules over the years and the handling of the differences between versions of the source data held at the partner sites. Partner source data exist in numerous versions/ incarnations. These include data collected in the field - as represented by the questionnaire, those stored in the databases after cleaning/ processing, those catalogued in the partner institution's data repository and the views extracted from databases to begin the harmonisation process. As a result, a documentation approach of the input data used in the harmonisation process would require flexibility to capture the various manifestations of the source data.

Participants involved in either project manager or principal investigator roles were interested in knowing the effort required from the partners to maintain/ update the envisaged provenance metadata. Knowing the effort required would help them in relating the proposal to demands on funds and human resources.

A trial period was recommended for the incremental versions of the metadata browser to allow users to evaluate if it will be performing the desired functions. This could be in the form of a dedicated day in a workshop setting allowing users both internal and external to ALPHA to test the software.

The use of DDI as the underlying standard for the provenance metadata was commended by participants who have some experience with the standard. They pointed out the interoperability capabilities that DDI metadata provides for data and tools sharing.

Some respondents expressed interest in the documentation of the linkages between data specifications as part of external validation of a specification or an intermediate dataset.

It was also pointed out that some aspects of the context needed for sensible analysis will still require communication between the ALPHA staff and the external users of the data. One example of such contexts was in relationship to ALPHA socio economic assessment variables on availability of electricity in a study site. In one site, prior to government programme introducing electricity, having electricity was a measure of SES status. When government introduced power, electricity ceased to be an indicator of higher economic status. This context require input from a data producer for a user to understand the context.

7.3.5 Feedback requiring structural changes to metadata

Four points coming up in the interviews require alterations to the metadata schema underpinning the ALPHA provenance metadata. These are:

The addition of an overall overview diagram linking datasets to tasks involved in producing a particular data specification. Respondents, especially those in data management roles, gave high scores to diagrams showing the association between a particular task and its input and output datasets. They also scored highly diagrams showing the association of a particular dataset to the tasks creating the dataset and those consuming it. However, rather than showing these associations for each task or dataset separately, an overall diagram of the associations for a data specification was generally preferred. Currently, the metadata doesn't show inputs and outputs with each intermediate step. Instead input and output datasets are presented for each high-level business process like the production of event data from staging data, the calculation of data quality metrics and so forth. Restructuring the metadata so that it follows datasets through the intermediate steps through which a business process is implemented has already been explored with the SAPRIN network. Here the metadata structure has been modified to support additional research and development.

The addition of documentation on the reasons for decisions made during data transformations, especially decisions to do with including or excluding individuals from a harmonised dataset. These reasons would be used to assess bias, for instance, owing to the systematic exclusions of individuals. In fact, DDI already supports this type of documentation but it wasn't included in the ALPHA "profile". It will be added so examples can be generated for future research and development.

The accommodation of hierarchies and other comparisons of concepts related to a task. The comparison would cater for differences in concepts definitions across data network partners. While definitions of concepts such as a birth and a death are clear cut, others such as migration, residency, HIV status, cohabitation etc are less consistently defined across ALPHA partners. The rules, assumptions and cut offs underpinning those definitions differ. The differences in these definitions have implications for data analysis and interpretation of results.

Consequently, they need to be explicitly stated in the structured metadata. In fact, DDI supports such concept comparison. This facility will be added to the ALPHA profile to support future research and development.

Accommodation of documentation of what a specification should and should not be used for – For example, a specification maybe suitable for survival analysis but not for fertility analysis due to the fact that birth histories are not collected for dead women in particular partner studies

7.4 Discussion

The question posed at the beginning of this chapter sought to elicit experts' views on the requirements for an ALPHA provenance metadata browser. The diversity in the scores per feature for most of the features and the differing preferences expressed in the rationales distinguish the needs of the two major user groups – data managers and researchers. While perspectives from researchers showed interest in high level metadata, those from data managers were inclined towards more granular metadata. In addition, the responses from the interviewees showed that the proposed features did not cater for network partners' specific contexts. This was indicated by suggestions to make definitions of concepts related to a task or step site-specific and to have site-specific statements of assumptions, rules and cut-offs applied to censoring and loss to follow-up for instance.

All but one of the interviewees who participated had at least a master's degree and at least two years of work experience in the fields of research and or data management. This indicates high levels of capability to know what is needed in a metadata browsing tool. In addition, it also gives an idea of the importance of the topic of discussion as highly skilled professionals were willing to commit time to participate in the interviews. To the best of the author's knowledge, this study represents the first attempt at systematically synthesising the perspectives of potential uses of harmonised HDSS data on the requirements for presenting, in human intelligible manner, standardised provenance metadata for these harmonised data. The mock-ups used metadata automatically mined from data transformations performed in the CiB. The elicitation study is a step towards extending the provenance documentation capabilities of the CiB which currently only provides tool specific documentation of data transformation processes. HDSS data and their management is complex (INDEPTH Network 2002; Benzler, Herbst, and Macleod 1998), standardised documentation of data transformations performed in creating secondary datasets can facilitate the communication of the nature of these data to would be users. The elicitation study contributes to the amelioration of the challenges faced by investigators trying to understand and accurately use

the ALPHA data. Efforts such as this study partially address appeals made in the data documentation literature (Vardigan, Heus, and Thomas 2008) for the development of software tools for presenting documentation of research datasets done using metadata standards.

Most interviewees said that they found the language to be very specialised to the areas of data transformations and integration and longitudinal demographic surveillance systems. They suggested the use of more generic terms and expressions and definitions of terms that needed to be used as they are. This will be addressed by the use of more accessible terms such as tasks instead of business processes, and phrases such as illegal transitions will be fully defined to facilitate comprehension.

The mock-ups used in this study captured metadata to do with the data transformation processes. Interviewees expressed the need to see an integrated provenance metadata browser solution combining partner source data documentation, metadata for the data transformations and those for the final harmonised datasets. This reinforces the view and goal in this thesis since chapter 4 addresses the documentation of partner source data and chapters 5 and 6 are focussed on the data harmonisation. The elicited requirements will be used together with the metadata developed in chapters 4 to 6 to provide a user-friendly platform for metadata catering for the entire scope of the ALPHA data lineage. This integrated view could be based on a dependency graph. Starting with an input variable a user could traverse the graph through one or more interim and output variables. The trace would do branching. Going backwards, one would start with an output variable and encounter first one or more interim variables and then one or more input variables.

The suggestions made for improving the proposed features are both technical and cultural. The technical suggestions are the ones needing changes in the metadata schema for the provenance metadata. Programmers will need to revisit the schema used to create the metadata presented in the mock-ups and adjust it accordingly to meet the suggestions from interviewees. Cultural changes required include paying more attention to metadata during data production than has been traditionally the case. As partners produce the harmonised datasets, they will need to provide the metadata not automatically captured by the metadata development software agents.

Taken together with the metadata developed in chapters 4 to 6, the elicited requirements provide the input needed to create the provenance metadata browsing platform. First, a requirements specification document will need to be created. This specification will then guide the development of the platform.

7.5 Data availability statement

The data collected and used in this study will be made available on the London School of Hygiene and Tropical Medicine institutional repository – the Data Compass. The transcripts will first be anonymised in line with the informed consent terms agreed to by the study participants. The associated documentation and data access conditions can be accessed on Data Compass through the following DOI: <https://doi.org/10.17037/DATA.00001522>.

8. CONCLUSION

SUMMARY OF FINDINGS, CONTRIBUTIONS, RECOMMENDATIONS AND FUTURE WORK

8.1 Introduction

The data documentation community has designed standards and models to promote common understanding of data and data production processes. Though they have potential to address ALPHA's documentation needs, these technologies are designed to cater for a wide user community which makes them too generic for direct use. For ALPHA to fully utilise them, an additional step is required, that of adapting them to fit the African demographic and HIV surveillance context. Today, exemplar recipes for adapting the generic standards and models in the documentation of HDSS data are missing. Consequently, this thesis aimed to provide recipes for customising available standards and models for the documentation of HDSS data and secondary datasets derived from them.

ALPHA data documentation can be sub-divided into three components – the metadata related to the input data collected by the partners, the documentation of the data harmonisation processes and the metadata for the harmonised datasets. The CiB technology used for the production of the ALPHA datasets already includes the last mentioned dimension – the documentation of harmonised data resulting from the ETLs, it does not address the first two though. This thesis sought to investigate the documentation of those first two dimensions of the ALPHA data provenance. In addition, it sought to gather requirements for an ALPHA provenance metadata browser software to guide developers of the software. This envisaged metadata browser would work as a platform for integrating the developed metadata and presenting them for searching and browsing in a human user-friendly fashion.

In the first stages of the project, the choices and considerations to be made in implementing metadata standards for the documentation of the primary data from the network partners were explored (chapter 4). Next, a business process model for African demographic and epidemiological surveillance was developed. This was done by analysing the tasks performed in the Pentaho ETLs, mapping them to the GLBPM steps and then specialising the identified GLBPM steps (Chapter 5). The specialisation involved relating the ETL tasks to the objects in the HDSS reference model and to concepts and relationships in the demographic and epidemiological surveillance domain. The lower level details of the ETL tasks were then mapped to the SDTL (Chapter 6) to bridge the gap between the business process models

and the implementation of the transformations in production systems. Statistical aggregates created in the quality assessment tasks of the ETLs were documented using SDMX (Chapter 6). Finally, an online requirements elicitation study was performed with data management and research experts to gather the requirements for a user-friendly software for browsing and searching the provenance metadata.

The next sections summarise the major findings, the research contributions and recommendations corresponding to each of the pieces of work done in this project. Finally, the study limitations and the potential future work are outlined.

8.2 Documentation of primary data from the network partners

8.2.1 Objective

- (1) To adopt and adapt the DDI metadata standard for the annotation of HDSS primary data using Kisesa HDSS, north western Tanzania as prototype

8.2.2 Summary of findings

Steps of the GLBPM which mapped to the activities in the Magu HDSS data life cycle were identified and used to determine the DDI metadata elements to use. A comparison of the use of the Nesstar Publisher and the Colectica Designer tools for documentation of typical HDSS data showed three key results. The first was that the amount of metadata entry required was comparable between the Codebook and Lifecycle strands of DDI. The second was that for both tools, the majority of the metadata had to be entered manually. The third was that DDI Codebook based tools had lower software costs and staff training needs than DDI Lifecycle ones. The software costs and training needs are a consequence of the well-known differences between the Codebook and Lifecycle versions of DDI. Lifecycle is more complex and comprehensive compared to Codebook (Data Documentation Initiative 2009). While the generic codebook based tools are free, mainly in the form of the IHSN tools funded by the World Bank (International Household Survey Network 2013), the most generic Lifecycle tool at the time of the investigation, the Colectica suite is commercial. It is rather the software costs and training demands that HDSS that are planning to implement metadata standards will need to consider not the amount of metadata entry. Other tools are open source (CentERdata 2019; Jensen 2012; Hebing 2015a), it appeared they required a lot more customisation compared to the Nesstar Publisher and Colectica Designer used in this study.

8.2.3 Research contributions

The foregoing findings add to the literature on population health research data documentation in general and on the documentation of primary datasets produced within HDSS settings in particular. To the best of the author's knowledge, this analysis is the first one to dig into the choices and considerations an HDSS needed to make in adapting metadata standards. While evidence of the use of the metadata standards for documenting data was there prior to this study (Ifakara Health Institute 2019; African Population and Health Research Center 2015; Africa Health Research Institute 2018), it was not clear what steps a typical HDSS needed to take in implementing them. These results will be of use to principal investigators leading HDSS studies, they will have a basis for deciding on the staffing and resource allocation for metadata development.

This study has also highlighted that the bulk of the metadata creation is manual. The tools used still required metadata to be entered manually. This is important information for planning purposes, with this information, a PI can budget accordingly and a data management lead can allocate staff to tasks for realistic periods of time based on this knowledge. On the other hand, data systems managers might also use the findings from this study to make decisions relating to the development of bespoke tools to improve automation of some metadata entry tasks.

8.2.4 Recommendations

The highlighted findings have informed some recommendations.

An HDSS will need to weigh the advantages that DDI lifecycle could bring to the management and preservation of the HDSS data versus the training needs and financial costs. The better resourced HDSS may need to endure the costs and staffing demands to promote the more comprehensive documentation of their data than what Codebook can provide.

Efforts to develop bespoke tools for metadata capturing will need to complement the existing tools and focus on improving areas such as automated metadata capture.

8.2.5 Future work

For ALPHA partners already using DDI-based Nesstar Publisher tools, further research might explore the development of a software agent for mining metadata relating to the variables that are the inputs for ALPHA data specifications from the partners' DDI instances. For partners not yet using DDI, the development of an ALPHA metadata profile for the input data used to create the harmonised data.

8.3 Documentation of data harmonisation processes

8.3.1 Objective

- (2) To develop a standards-based framework for the documentation of retrospective data harmonisation routines performed in ALPHA and similar networks

8.3.2 Summary of findings

All the tasks within the ALPHA ETL for spec 6.1 were mapped to steps in the GLBPM. While the mapping to the GLBPM communicated the essence of the tasks in an internationally understood manner, it was too basic and generic, hiding the idiosyncrasies and the complexity of the ETL. A specialisation of the mapped GLBPM steps resulted in a more concrete business process model, the ADESBPM, this specialised model described the ETL in terms and concepts familiar to experts working in the demographic and epidemiological surveillance domain. The analysis of the ETL and initial attempts to describe them using the proposed DDI 4 process model highlighted the shortcomings of the model. In its original form, the process model was more inclined towards survey data collection with inadequate provision for information objects, flow logic and patterns associated with data management (Greenfield, Kanjala, and Gregory 2019).

The ADESBPM provided a high level description of the data transformations. It provides a basis for prospective or retrospective business level documentation of ETL based on HDSS data. Regarding the lower level details, that is the variable level data transformations, the results showed that the SDTL catered for the description of the majority of the granular services performed in Pentaho to transform data. In addition, SDMX was successfully applied for documentation of the data quality metrics annotating the statistical values with the SDMX concepts, dimensions and keys.

8.3.3 Research contributions

The present study makes some noteworthy contributions to both literature and practice relating to research data documentation. It adds to the literature in the areas of data harmonisation, data provenance, metadata standards and application of business process models. The results contribute to the enhancement of the descriptive power of the DDI 4 process model. Prior to the field testing of the model in the ALPHA use case, it was inclined towards data collection.

This study is the first comprehensive investigation of the development of structured and tool-agnostic provenance metadata for HDSS harmonised datasets. It has expanded the metadata capabilities of the CiB. Prior to this investigation, the CiB technology only relied

on the data transformations description within Pentaho. Now, a framework for documenting Pentaho ETLs in a DDI compliant format is available. This is important because it provides common understanding of the data transformations to potential users not familiar with the Pentaho tool used to create the transformations. In addition, these structured provenance metadata enhance the potential for automated exchange and use of the data and metadata across processing and analysis platforms. Besides the benefit it brings to potential users of the data, it is also valuable for increasing transparency and standardisation within the network and the potential for reuse of procedures thus streamlining the data production processes among the partners. These metadata are crucial for institutional memory and knowledge transfer should site data managers change or if ALPHA was to switch from Pentaho to another ETL software. Most of the ALPHA data management problems have stemmed from staff changes.

Beyond the contributions at high level addressed in Chapter 5, the largely successful mapping of the granular details of the transformations to the SDTL and the description of data metrics using SDMX in Chapter 6 are both first attempts in the HDSS data harmonisation arena. Beyond the description afforded by the ADESBPM, this project has also provided structured metadata of the finer details of the data transformations and the statistical aggregates generated as part of the harmonisation. These more granular metadata serve the crucial role of capturing the details necessary for moving from a general understanding of the processes provided for by the business process models to implementation steps.

Overall, through the use of DDI, SDTL and SDMX and the process models, this study is showing that there is no one magic standard for the documentation of “after the fact” harmonised data such as the ALPHA datasets. Rather, it is through the use of a combination of the standards depending on the stage of the process that the appropriate level of detail of the documentation can be achieved.

8.3.4 Recommendations

These findings suggest several courses of action for various stakeholders involved/ interested in ALPHA data and their provenance. Producers of ALPHA data may find it useful to follow the approach proposed in communicating the provenance of their data to users. Principal investigators will need to communicate to funders the metadata development work that complements the development of the ETLs in order for funding to include metadata development. Data experts whose role is to perform the transformations may need to browse the metadata in cases where they are available in order to use them as models for their work.

They will need to put the tasks in an ETL into context, annotating the steps/ transformations with the theoretical underpinnings related to the transformation.

Preliminary results from this work have been presented at a conference (Kanjala et al. 2018) and at the 2019 ModernStats World Workshop (Greenfield, Kanjala, and Gregory 2019). Metadata standards and business process models user communities may find it useful to consider the approach used in this thesis in customising these within their own contexts.

8.3.5 Future work

Outside the scope of the thesis, some prototyping work of a software agent for automated harvesting of provenance metadata from Pentaho ETLs have been done (Greenfield, Kanjala, and Gregory 2019). The metadata mined by the agent have been complemented by the infusion metadata that were created in this thesis (Chapter 5). The prototyped software agent could be developed further into a production ready scale. So far, the infusion metadata were being entered directly into a DDI 4 XML document. To facilitate the entry of these metadata, further work of developing a user-friendly platform could enable those not familiar with XML to participate in this metadata entry task.

In the future, it will be important to extend the approach devised in the current study to the harmonisation of the disease or behaviour specific variables. Analysis of ETLs for the disease/ behaviour specific specs such as the HIV status specification or the sexual behaviour specification will most likely enrich the ADESBPM with objects and tasks that are not covered in the ETL for spec 6.1.

8.4 Provenance metadata browser software requirements

The foregoing summarised findings in essence represent an end to end standards-based metadata solution for ALPHA datasets capturing the context of the primary data from the partners (Chapter 4) and the details of the data transformations (Chapters 5 and 6). When integrated with the codebook metadata for the harmonised datasets already catered for within the CiB, this would provide an end to end structured provenance metadata for the ALPHA datasets. While the provenance metadata provided in DDI and SDMX, in the case of ALPHA in XML format, are machine friendly, they are not suitable for human users. Humans need an intelligible presentation of the metadata for querying and searching. The software to present the ALPHA provenance metadata in human friendly format is not available off shelf. In preparation for the designing and development of this software, the perspectives of the current and potential future users of the ALPHA data were elicited and analysed.

8.4.1 Objective

- (3) To gather and analyse the requirements for a user-friendly provenance metadata platform for ALPHA datasets

8.4.2 Summary of findings

Proposed features were generally accepted by respondents with comments to improve those features also provided. Two user groups emerged in the study, the data management group and the researchers group. Researchers had interest in high level documentation and not primarily interested in replicating the process but to get a business level understanding. Data managers required more granular details of the metadata and expressed interest in reproducing the data harmonisation processes. Interviewees suggested that the interface of the software will need to not overwhelm researchers with too much detail but making it possible for data managers to get the details when they need them. Some participants pointed out the need for a metadata capturing culture to be developed among the ALPHA network partners.

8.4.3 Research contributions

This research adds to our knowledge, the perspectives of data managers and researchers on what is required providing metadata to users of the ALPHA data. Developers now have a basis for beginning the development of the metadata browser. The results from the study emphasise the need for the metadata browser to accommodate the differences between partner studies in the way they define concepts and the rules and the assumptions they make regarding cut off points in defining events such as migration or loss to follow up. This will enable users to understand the differences in the data resulting from the differences in the assumptions, rules and cut-offs applied. The findings have also highlighted the need for partners to pay attention to the capturing of contextual metadata along the ETL processes.

8.4.4 Recommendations

Developers will need to review the evaluation of the proposed features and adjust the initial requirements specification as needed. This will enable them to target a fit for purpose software for ALPHA metadata browsing in their work.

The network partners will need to adopt a culture of capturing contextual metadata along the ETL life cycle recording assumptions, rules and reasons for decisions made in creating the ETLs. This is better done during the time of data processing. Retrospectively capturing these metadata is challenging due to the huge amount of data involved and the information loss as the time between the ETL development and its documentation increases.

8.4.5 Future work

The metadata browser: The most logical follow up to the requirements elicitation study will be to engage the expertise of developers to build the envisaged human user-friendly metadata browser. This will be a platform for browsing and searching the provenance metadata to facilitate the understanding and more accurate interpretation and use of the harmonised data.

In addition, this browser will need to integrate the three components of the ALPHA provenance metadata – the metadata for the primary HDSS data from the partners, the metadata on the harmonisation processes and those for the final harmonised datasets.

Another interesting future investigation related to this browser would be to explore the extension of its interactive capabilities to receiving and structuring feedback from users of the software. This can help to continually improve the metadata to meet the needs of the users better.

User guides and manuals: Manuals will need to be prepared outlining the steps for creating the metadata and describing the functionality of the software tools for (1) automated harvesting of metadata from DDI instances for primary data and from the ETL, (2) manual metadata entry forms, (3) metadata integration and (4) the metadata browsing. These will facilitate the use and maintenance of metadata infrastructure.

A curriculum: A curriculum at postgraduate level is needed for training a cadre of staff competent in the metadata development, system design, development, deployment and maintenance. This curriculum could also be used for training of HDSS personnel on the methods and software for use with the data documentation subsystem.

8.5 Study limitations

While this study has several potential contributions to theory and practice in the area of HDSS data management and preservation, it also has limitations.

The scope of this study is limited in terms of the number of HDSS studies and ALPHA Specifications investigated. It only considered two HDSS and only the core HDSS data. The full complement of ALPHA datasets comprises of eleven data specifications and the network has ten partners. Capturing the diversity of the ten partners' practices and data characteristics would better inform the decisions involved in implementing metadata standards in HDSS. This would be a flexible approach accommodating their funding availability and skill sets differences among the partners.

Disease and or behaviour specific specifications not considered in this project are important for enriching the ADESBPM with epidemiological aspects that are not present in the residency episodes data used in this thesis.

While the approaches and provenance metadata were developed and users' views on the requirements for metadata browsing software were elicited, the study did not go as far as developing the tools to facilitate the creation of the said metadata and the searching and browsing of the same. These tools needed in production ready form for this kind of documentation to be realised beyond prototypes. Without this accompanying tooling, the documentation is not feasible. The envisaged tools will need to automate the metadata development in order to minimise the workload on data producers. There are plans, outside the scope of the current study, to build and deploy these tools as outlined in the future work section.

The study only compared two tools for DDI documentation of HDSS primary data—Colectica Designer and Nesstar Publisher. It did not expand the scope of that comparison to check if amount of manual entry of metadata, costs and or training needs were going to be altered by using alternative tools.

Overall, the forgoing investigation has shown that the existing standards when combined can do the job, but it is time and labour intensive, requires a lot of knowledge and expertise to join it all up, these are all in short supply. Whilst that can be partially addressed by the development of training materials and by training people, there is a substantial gap in the market for tools to streamline and automate these processes. Longitudinal studies, and meta-analyses, where large volumes of sometimes poorly documented data are combined, would be the main beneficiaries but this would be useful wherever data are harmonised. Until documentation as described here is built into studies and is completely routine, data sharing will always be hampered.

9. REFERENCES

- Africa Centre for Population Health. 2015. 'Research Data Management Platform'. 2015. <http://www.africacentre.ac.za/index.php/data-research-management>.
- Africa Health Research Institute. 2018. 'AHRI Data Repository'. 2018. <https://data.africacentre.ac.za/index.php/auth/login/?destination=home>.
- African Population and Health Research Center. 2015. 'APHRC Data Portal'. African Population and Health Research Center. 2015. <http://microdataportal.aphrc.org/index.php/catalog>.
- Alistair Hamilton, Eden Brinkley, and Therese Lalor. 2012. 'Operationalizing Metadata Frameworks – An ABS Perspective'. In *Proceedings of the Fourth International Conference on Establishment Surveys*. Montréal, Canada. www.amstat.org/meetings/ices/2012/papers/302207.pdf.
- ALPHA. 2013. 'The ALPHA Network | London School of Hygiene & Tropical Medicine | LSHTM'. 2013. <http://www.lshtm.ac.uk/eph/dph/research/alpha/>.
- Alter, George. 2018. 'Continuous Capture of Metadata for Statistical Data (NSF ACI-1640575)'. presented at the IASSIST and CARTO 2018: Once Upon a Data Point: Sustaining our Data Storytellers, Montréal, Canada, May. https://www.dropbox.com/sh/f22xwmt5h2h63m/AACQZ_-xq568aovR1n3MNHIIa/B5_Alter.pdf?dl=0.
- ANDLA. 2019. 'African NCD Longitudinal Data Alliance'. 2019. <https://andla.org/>.
- Anne Thomson, Graham Elele, and Felix Schmieding. 2013. 'Independent Evaluation of the International Household Survey Network (IHSN) and Accelerated Data Program (ADP)'. Final Report. Oxford, United Kingdom: Oxford Policy Management.
- Antoniou, Grigoris, and Frank Van Harmelen. 2004. *A Semantic Web Primer*. MIT press.
- Arofan Gregory, and Pascal Heus. 2008. 'The Data Documentation Initiative (DDI)'. Wiesbaden, Germany, June.
- Asiki, Gershim, Georgina Murphy, Jessica Nakiyingi-Miir, Janet Seeley, Rebecca N Nsubuga, Alex Karabarinde, Laban Waswa, Sam Biraro, Ivan Kasamba, and Cristina Pomilla. 2013. 'The General Population Cohort in Rural South-Western Uganda: A Platform for Communicable and Non-Communicable Disease Studies'. *International Journal of Epidemiology* 42 (1): 129–41.
- Asiki, Gershim, Georges Reniers, Robert Newton, Kathy Baisley, Jessica Nakiyingi-Miir, Emma Slaymaker, Ivan Kasamba, Janet Seeley, Jim Todd, and Pontiano Kaleebu. 2016. 'Adult Life Expectancy Trends in the Era of Antiretroviral Treatment in Rural Uganda (1991–2012)'. *Aids* 30 (3): 487–93.
- Ausborn, Scot, Julia Rotondo, and Tim Mulcahy. 2014. 'Mapping the General Social Survey to the Generic Statistical Business Process Model: NORC's Experience'. *IASSIST QUARTERLY*, 21.
- Barkow, Ingo, William Block, Jay Greenfield, Arofan Gregory, Marcel Hebing, Larry Hoyle, and Wolfgang Zenk-möltgen. 2013. 'GENERIC LONGITUDINAL BUSINESS PROCESS MODEL DDI Working Paper Series – Longitudinal Best Generic Longitudinal Business Process Model'. *Business*, 1–26.
- Barkow, Ingo. 2016. 'The Challenges of Metadata Management in Computer-Based Surveys and Assessments'.
- Barkow, Ingo, William Block, Jay Greenfield, Arofan Gregory, Marcel Hebing, Larry Hoyle, and Wolfgang Zenk-möltgen. 2013. 'GENERIC LONGITUDINAL BUSINESS PROCESS MODEL DDI Working Paper Series – Longitudinal Best Generic Longitudinal Business Process Model'. *Business*, 1–26.
- Barkow, Ingo, William Block, Jay Greenfield, Arofan Gregory, Marcel Hebing, Larry Hoyle, and Wolfgang Zenk-Möltgen. 2013. 'The Generic Longitudinal Business Process

- Model'. 5. *Longitudinal Best Practices*. DDI Working Paper Series – Longitudinal Best Practices. Data Documentation Alliance.
- Barkow, Ingo, and David Schiller. 2013. 'Rogatus—a Planned Open Source Toolset to Cover the Whole Lifecycle'. presented at the North American DDI Conference.
- Beguy, Donatien, Patricia Elung'ata, Blessing Mberu, Clement Oduor, Marylene Wamukoya, Bonface Nganyi, and Alex Ezech. 2015. 'HDSS Profile: The Nairobi Urban Health and Demographic Surveillance System (NUHDSS)'. *International Journal of Epidemiology*, dyu251.
- Benzler, Justus, Kobus Herbst, and Bruce Macleod. 1998. 'A Data Model for Demographic Surveillance Systems 1'. *Science*.
- Bergeron, Julie, Dany Doiron, Yannick Marcon, Vincent Ferretti, and Isabel Fortier. 2018. 'Fostering Population-Based Cohort Data Discovery: The Maelstrom Research Cataloguing Toolkit'. *PLoS One* 13 (7): e0200926.
- Blank, Grant, and Karsten Boye Rasmussen. 2004. 'The Data Documentation Initiative: The Value and Significance of a Worldwide Standard'. *Social Science Computer Review* 22 (3): 307–18.
- Bocquier, Philippe, Carren Ginsburg, Kobus Herbst, Osman Sankoh, and Mark A Collinson. 2017. 'A Training Manual for Event History Data Management Using Health and Demographic Surveillance System Data'. *BMC Research Notes* 10 (1): 224.
- Bosch, Thomas, Olof Olsson, Benjamin Zapilko, Arofan Gregory, and Joachim Wackerow. 2013. 'DDI-RDF Discovery—A Discovery Model for Microdata'. *LASSIST Quarterly*, 17.
- Bosch-Capblanch, Xavier. 2011. 'Harmonisation of Variables Names Prior to Conducting Statistical Analyses with Multiple Datasets: An Automated Approach'. *BMC Medical Informatics and Decision Making* 11 (1).
- Braa, Jørn, and Sundeep Sahay. 2017. 'The DHIS2 Open Source Software Platform: Evolution over Time and Space'. *Global Health Informatics. Principles of EHealth and MHealth to Improve Quality of Care*. The MIT Press.
- Brancato, Giovanna, and Giorgia Simeoni. 2012. 'Istat Statistical Process Modelling and the Generic Statistical Business Process Model: A Comparison'. In .
- Brown, Adam, Jeremy Iverson, Dan Smith, and Sally Vermaaten. 2012. 'Powering Official Statistics at Statistics New Zealand with DDI-L and Colectica: A Case Study'. In . <http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi12/paper/view/42>.
- Bruijne, Marika de, and Alek Amin. 2009. 'Questasy: Online Survey Data Dissemination Using DDI 3'. *LASSIST Quarterly* 33 (Spring): 10–15.
- Byass, Peter, Clara Calvert, Jessica Miro-Nakiyingi, Tom Lutalo, Denna Michael, Amelia Crampin, Simon Gregson, Albert Takaruzza, Laura Robertson, and Kobus Herbst. 2013. 'InterVA-4 as a Public Health Tool for Measuring HIV/AIDS Mortality: A Validation Study from Five African Countries'. *Global Health Action* 6.
- C2Metadata. 2017. 'Structured Data Transform Language'. Structured Data Transform Language. 2017. <http://c2metadata.gitlab.io/sdtl-docs/>.
- Candela, Leonardo, Donatella Castelli, Paolo Manghi, and Alice Tani. 2015. 'Data Journals: A Survey'. *Journal of the Association for Information Science and Technology* 66 (9): 1747–62.
- CentERdata. 2019. 'Questasy | CentERdata.NL'. 2019. <https://www.centerdata.nl/en/projects-by-centerdata/questasy-0>.
- Chandramohan, Daniel, Kenji Shibuya, Philip Setel, Sandy Cairncross, Alan D Lopez, Christopher J L Murray, Basia Zaba, Robert W Snow, and Fred Binka. 2008. 'Should Data from Demographic Surveillance Systems Be Made More Widely Available to Researchers?' *PLoS Medicine* 5 (2): e57. <https://doi.org/10.1371/journal.pmed.0050057>.

- CLOSER. 2019. 'CLOSER Discovery - CLOSER - UCL Wiki'. 2019. <https://wiki.ucl.ac.uk/display/CLOS/CLOSER+Discovery>.
- Colectica. 2015a. 'Colectica — Colectica 5.1 Documentation'. 2015. <http://docs.colectica.com/>.
- . 2015b. 'Colectica Information Model — Colectica 5.1 Documentation'. 2015. <http://docs.colectica.com/introduction/information-model/>.
- . 2019a. 'Colectica — Colectica 5.5 Documentation'. 2019. <https://docs.colectica.com/>.
- . 2019b. 'Colectica - Colectica Customers'. 2019. <http://www.colectica.com/customers>.
- . 2019c. 'Colectica - Colectica for Excel'. 2019. <http://www.colectica.com/software/colecticaforexcel>.
- . 2019d. 'Colectica - Colectica Reader'. 2019. <http://www.colectica.com/software/reader>.
- Cooper, Rachel, Rebecca Hardy, Avan Aihie Sayer, Yoav Ben-Shlomo, Kate Birnie, Cyrus Cooper, Leone Craig, Ian J Deary, Panayotes Demakakos, and John Gallacher. 2011. 'Age and Gender Differences in Physical Capability Levels from Mid-Life Onwards: The Harmonisation and Meta-Analysis of Data from Eight UK Cohort Studies'. *PLoS One* 6 (11): e27899.
- Corti, Louise, and Arofan Gregory. 2011. 'CAQDAS Comparability. What about CAQDAS Data Exchange?' In . Vol. 12. FQS.
- Cotton, Franck, Richard Cyganiak, RTAM Grim, Daniel W Gillman, Yves Jaques, and Wendy Thomas. 2013. 'XKOS: An SKOS Extension for Statistical Classifications'. In . Citeseer.
- Crampin, Amelia C, Albert Dube, Sebastian Mboma, Alison Price, Menard Chihana, Andreas Jahn, Angela Baschieri, Anna Molesworth, Elnaeus Mwaiyeghele, and Keith Branson. 2012. 'Profile: The Karonga Health and Demographic Surveillance System'. *International Journal of Epidemiology* 41 (3): 676–85.
- Cui, Licong, Ningzhou Zeng, Matthew Kim, Remo Mueller, Emily R Hankosky, Susan Redline, and Guo-Qiang Zhang. 2018. 'X-Search: An Open Access Interface for Cross-Cohort Exploration of the National Sleep Research Resource'. *BMC Medical Informatics and Decision Making* 18 (1): 99–99. <https://doi.org/10.1186/s12911-018-0682-y>.
- Dan Gillman, and Arofan Gregory. 2015. 'Modernizing the Data Documentation Initiative (DDI-4)'. presented at the WICS, Geneva, Switzerland, May. <https://slideplayer.com/slide/5961883/>.
- Data Documentation Initiative. 2009. 'DDI 3.1 Schema and Documentation-Part I—Overview, Schema and Field Level Documentation'. Data Documentation Initiative Alliance.
- DataCite. 2019. 'DataCite Schema'. Website. DataCite Schema. February 2019. <https://schema.datacite.org/>.
- DCC. 2019. 'List of Metadata Standards | Digital Curation Centre'. DCC. 2019. <http://www.dcc.ac.uk/resources/metadata-standards/list?page=1>.
- DCMI. 2013. 'Dublin Core Metadata Element Set, Version 1.1'. 2013. <http://dublincore.org/documents/dces/>.
- DDI Alliance. 2014a. 'Data Documentation Initiative (DDI) Technical Specification Part II: User Guide'. <https://ddialliance.org/explore-documentation>.
- . 2014b. 'Structure of DDI 4 - DDI - Confluence'. DDI/Moving Forward Project (DDI4). November 2014. <https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/491707/Structure+of+DDI+4>.

- . 2015a. ‘Workflows — DDI 4.0 Dev Documentation’. DDI-Views (Q2 Release 2016) Documentation. 2015. <https://ddi-views.readthedocs.io/en/latest/Package/Workflows/index.html>.
- . 2015b. ‘DDI Lifecycle 3.2 (2014-02-05) [HTML Corrected 2014-05-15]’. DDI Lifecycle 3.2 (2014-02-05) [HTML Corrected 2014-05-15]. 5 February 2015. <http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/>.
- . 2018a. ‘Data Life Cycle | Data Documentation Initiative’. DDI. 2018. <https://www.ddialliance.org/taxonomy/term/170>.
- . 2018b. ‘DDI 3.2 XML Schema Documentation (2014-02-05)’. Explore Documentation. 2018. <http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/>.
- . 2018c. ‘DDI Tools | Data Documentation Initiative’. DDI. 2018. <http://www.ddialliance.org/resources/tools>.
- . 2018d. ‘Welcome to the Data Documentation Initiative | Data Documentation Initiative’. 2018. <https://www.ddialliance.org/>.
- . 2019. ‘Moving Forward Project (DDI4)’. 2019. <https://ddialliance.atlassian.net/wiki/spaces/DDI4/pages/491703/Moving+Forward+Project+DDI4>.
- DDIAlliance.org. 2013. ‘DDI - Data Documentation Initiative’. Welcome to the Data Documentation Initiative. 2013. <http://www.ddialliance.org/>.
- Desai, Meghna, Ann M Buff, Sammy Khagayi, Peter Byass, Nyaguara Amek, Annemieke van Eijk, Laurence Slutsker, John Vulule, Frank O Odhiambo, and Penelope A Phillips-Howard. 2014. ‘Age-Specific Malaria Mortality Rates in the KEMRI/CDC Health and Demographic Surveillance System in Western Kenya, 2003–2010’.
- Di Pasquale, Aurelio. 2018. ‘Improving Quality, Timeliness and Efficacy of Data Collection and Management in Population-Based Surveillance of Vital Events’.
- Digital Curation Centre. 2013. ‘Nesstar | Digital Curation Centre’. 12 June 2013. <http://www.dcc.ac.uk/resources/external/nesstar>.
- . 2019. ‘List of Metadata Standards’. List of Metadata Standards | Digital Curation Centre. 4 January 2019. <http://www.dcc.ac.uk/resources/metadata-standards/list>.
- Doiron, Dany, Yannick Marcon, Isabel Fortier, Paul Burton, and Vincent Ferretti. 2017. ‘Software Application Profile: Opal and Mica: Open-Source Software Solutions for Epidemiological Data Management, Harmonization and Dissemination’. *International Journal of Epidemiology*.
- Dunnet, G. 2007. ‘The BmTS: Creating a New Business Model for a National Statistical Office of the 21st Century’. In , 8–11.
- Dupriez, Olivier, and Geoffrey Greenwell. 2007. ‘Quick Reference Guide for Data Archivists’. *IHSN Paper*. <Http://Www.Ihsn.Org/Home/Download.Php>.
- Duval, Erik. 2001. ‘Metadata Standards: What, Who & Why’. *Journal of Universal Computer Science* 7 (7): 591–601.
- Eric Miller. 1998. ‘An Introduction to the Resource Description Framework’. *D-Lib Magazine*, 1998.
- Eurostat, Directorate B: Statistical Methodologies and Tools, and Unit B-5: Statistical Information Technologies. 2010. ‘Introduction to SDMX’. Student Book 1. SDMX Self-Learning Package. Eurostat. https://circabc.europa.eu/d/a/workspace/SpacesStore/d400cde2-f042-4e2c-8901-694ce3217507/01_SDMX_Introduction_student_book_2010.pdf.
- Federer, Lisa M., Christopher W. Belter, Douglas J. Joubert, Alicia Livinski, Ya-Ling Lu, Lissa N. Snyders, and Holly Thompson. 2018. ‘Data Sharing in PLOS ONE: An

- Analysis of Data Availability Statements'. Edited by Jelte M. Wicherts. *PLOS ONE* 13 (5): e0194768. <https://doi.org/10.1371/journal.pone.0194768>.
- Fortier, Isabel, Paul R Burton, Paula J Robson, Vincent Ferretti, Julian Little, Francois L'heureux, Mylène Deschênes, Bartha M Knoppers, Dany Doiron, and Joost C Keers. 2010. 'Quality, Quantity and Harmony: The DataSHaPER Approach to Integrating Data across Bioclinical Studies'. *International Journal of Epidemiology* 39 (5): 1383–93.
- Fortier, Isabel, Parminder Raina, Edwin R Van den Heuvel, Lauren E Griffith, Camille Craig, Matilda Saliba, Dany Doiron, Ronald P Stolk, Bartha M Knoppers, and Vincent Ferretti. 2017. 'Maelstrom Research Guidelines for Rigorous Retrospective Data Harmonization'. *International Journal of Epidemiology* 46 (1): 103–5.
- Geubbels, Eveline, Shamte Amri, Francis Levira, Joanna Schellenberg, Honorati Masanja, and Rose Nathan. 2015. 'Health & Demographic Surveillance System Profile: The Ifakara Rural and Urban Health and Demographic Surveillance System (Ifakara HDSS)'. *International Journal of Epidemiology* 44 (3): 848–61.
- Ghys, Peter D, Basia Zaba, and Maria Prins. 2007. 'Survival and Mortality of People Infected with HIV in Low and Middle Income Countries: Results from the Extended ALPHA Network.' *AIDS (London, England)* 21 Suppl 6 (November): S1–4. <https://doi.org/10.1097/01.aids.0000299404.99033.bf>.
- Gierl, Claude, and Jon Johnson. 2012. '70 Years of UK Birth Cohort Data into DDI Lifecycle?' In *EDDI12–4th Annual European DDI User Conference*. <http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi12/paper/view/38>.
- Gómez-Olivé, Francesc Xavier, Nicole Angotti, Brian Houle, Kerstin Klipstein-Grobusch, Chodziwadziwa Kabudula, Jane Menken, Jill Williams, Stephen Tollman, and Samuel J Clark. 2013. 'Prevalence of HIV among Those 15 and Older in Rural South Africa'. *AIDS Care*, no. ahead-of-print: 1–7.
- Granda, Peter, and Emily Blasczyk. 2016. 'XIII. Data Harmonization'. In *Cross-Cultural Survey Guidelines*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.
- Greenberg, Jane. 2005. 'Understanding Metadata and Metadata Schemes'. *Cataloging & Classification Quarterly* 40 (3–4): 17–36.
- Greenfield, Jay. 2018. 'The DDI Data Management View: Prospective and Retrospective Workflow Description'. In . Geneva, Switzerland. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2018/mtg1/DDI_GREENFILED_Paper.pdf.
- Greenfield, Jay, Chifundo Kanjala, and Arofan Gregory. 2019. 'Documenting Health and Population Research in DDI 4'. Geneva, Switzerland, June. https://statswiki.unece.org/download/attachments/254672934/MWW2019_DDI_KanjalaGregoryGreenfield_Presentation.pptx?version=3&modificationDate=1560813937355&api=v2.
- Gregory, Arofan, and Pascal Heus. 2007. 'DDI and SDMX: Complementary , Not Competing , Standards'.
- Gregory, Arofan, Heus Pascal, and Jostein Ryssevik. 2009. 'Metadata'. *SSRN Electronic Journal*, August. <https://doi.org/10.2139/ssrn.1447866>.
- Gregson, Simon, Owen Mugurungi, Jeffrey Eaton, Albert Takaruzza, Rebecca Rhead, Rufurwokuda Maswera, Junior Mutsvangwa, Justin Mayini, Morten Skovdal, and Robin Schaefer. 2017. 'Documenting and Explaining the HIV Decline in East Zimbabwe: The Manicaland General Population Cohort'. *BMJ Open* 7 (10): e015898.
- Hallett, Timothy B, Basia Zaba, Jim Todd, Ben Lopman, Wambura Mwita, Sam Biraro, Simon Gregson, J Ties Boerma, and Alpha Network. 2008. 'Estimating Incidence

- from Prevalence in Generalised HIV Epidemics: Methods and Validation'. *PLoS Med* 5 (4): e80.
- Hansen, S.E., Jeremy Iverson, Uwe Jensen, Hilde Orten, and Johanna Vompras. 2011. 'ENABLING LONGITUDINAL DATA COMPARISON USING DDI'. *Ddialliance.Org*, no. 2.
- Hebing, Marcel. 2015a. 'A Metadata-Driven Approach to Panel Data Management and Its Application in DDI on Rails'.
- . 2015b. 'A Metadata-Driven Approach to Panel Data Management and Its Application in DDI on Rails'. University of Bamberg. https://opus4.kobv.de/opus4-bamberg/frontdoor/deliver/index/docId/46614/file/hebingdissopusse_A3a.pdf.
- Herbst, Kobus, Sanjay Juvekar, Tathagata Bhattacharjee, Martin Bangha, Nidhi Patharia, Titus Tei, Brendan Gilbert, and Osman Sankoh. 2015. 'The INDEPTH Data Repository: An International Resource for Longitudinal Population and Health Data From Health and Demographic Surveillance Systems'. *Journal of Empirical Research on Human Research Ethics* 10 (3): 324–33.
- HL 7. 2019. 'Health Level Seven International - Homepage | HL7 International'. 2019. <https://www.hl7.org/index.cfm>.
- HL7. 2019. 'Index - FHIR v4.0.0'. HL7 FHIR. December 2019. <http://hl7.org/fhir/>.
- Hoyle, By Larry, Fortunato Castillo, Benjamin Clark, Denise Perpich, and Joachim Wackerow. 2011. 'METADATA FOR THE LONGITUDINAL DATA LIFE CYCLE Metadata for the Longitudinal Data Life Cycle'.
- Huff, Stanley M, Roberto A Rocha, Clement J McDonald, Georges JE De Moor, Tom Fiers, W Dean Bidgood Jr, Arden W Forrey, William G Francis, Wayne R Tracy, and Dennis Leavelle. 1998. 'Development of the Logical Observation Identifier Names and Codes (LOINC) Vocabulary'. *Journal of the American Medical Informatics Association* 5 (3): 276–92.
- Ifakara HDSS. 2010. 'Ifakara HDSS'. 2010. [http://www.indepth-network.org/Profiles/Ifakara HDSS.pdf](http://www.indepth-network.org/Profiles/Ifakara%20HDSS.pdf).
- Ifakara Health Institute. 2019. 'Central Data Catalog - Ifakara Health Institute'. 2019. <http://data.ihl.or.tz/index.php/catalog/>.
- IHSN. 2012. 'IHSN Technical Note on Metadata Standards - DRAFT'. International Household Survey Network. www.ihsn.org/sites/default/files/resources/DDI_SDMX_IHSN_DRAFT.pdf.
- . 2013. 'About | ADP'. 2013. <http://adp.ihsn.org/node/203>.
- Ihsn.org. 2013. 'Mission and Objectives | IHSN.' 2013. <http://www.ihsn.org/home/content/about/objectives>.
- INDEPTH Network. 2002. *Population, Health and Survival at INDEPTH Sites*. International Development Research Centre.
- . 2013. 'INDEPTH Data Repository - History'. 2013. <http://www.indepth-ishare.org/index.php/history>.
- Ingo Barkow. 2015. 'Data Management Module (DMM) A New Module for the Rogatus Survey Platform'. presented at the North American DDI Conference.
- Inter University Consortium of Political and Social Research. 2019. 'What Is a Codebook?' ICPSR Find and Analyse Data. 2019. <https://www.icpsr.umich.edu/icpsrweb/content/shared/ICPSR/faqs/what-is-a-codebook.html>.
- International Household Survey Network. 2013. 'Mission and Objectives | IHSN'. 2013. <http://www.ihsn.org/home/content/about/objectives>.
- . 2016. 'Microdata Cataloging Tool (NADA) | IHSN'. 2016. <http://www.ihsn.org/home/software/nada>.

- . 2018. 'Software | IHSN'. 2018. <http://www.surveynetwork.org/software>.
- Ionescu, Sanda. 2018. 'End-to-End Process and DDI'. presented at the IASSIST and CARTO 2018: Once upon a datapoint: Sustaining our data storytellers, Montréal, Canada, May. https://www.dropbox.com/sh/f22xwmt5h2h63m/AACDKWm6_5zv12hRnOICYLpea?dl=0&preview=B5_Ionescu.pdf.
- Irish, Seth, David Kyalo, Robert W. Snow, and Maureen Coetzee. 2019. 'Anopheles Species Present in Countries in Sub-Saharan Africa and Associated Islands'. Harvard Dataverse. <https://doi.org/10.7910/dvn/phgadl>.
- J Madise, Nyovani, Abdhalah K Ziraba, Joseph Inungu, Samoel A Khamadi, Alex Ezech, Eliya M Zulu, John Kebaso, Vincent Okoth, and Matilu Mwau. 2012. 'Are Slum Dwellers at Heightened Risk of HIV Infection than Other Urban Residents? Evidence from Population-Based HIV Prevalence Surveys in Kenya'. *Health & Place* 18 (5): 1144–52.
- Jensen, Jannik. 2012. 'DdiEditor'. In *EDDI12–4th Annual European DDI User Conference*.
- Joachim Wackerow, Larry Hoyle, and Thomas Bosch. 2014. 'Physical Data Description (PHDD)'. DDI Alliance. 2014. <http://rdf-vocabulary.ddialliance.org/phdd.html>.
- 'JSON'. 2019. Introducing JSON. 6 October 2019. <http://json.org/>.
- Kabudula, Chodziwadziwa W, Benjamin D Clark, Francesc Xavier Gómez-Olivé, Stephen Tollman, Jane Menken, and Georges Reniers. 2014. 'The Promise of Record Linkage for Assessing the Uptake of Health Services in Resource Constrained Settings: A Pilot Study from South Africa'. *BMC Medical Research Methodology* 14 (1): 71.
- Kahn, Kathleen, Mark A Collinson, F Xavier Gómez-Olivé, Obed Mokoena, Rhian Twine, Paul Mee, Sulaimon A Afolabi, Benjamin D Clark, Chodziwadziwa W Kabudula, and Audrey Khosa. 2012. 'Profile: Agincourt Health and Socio-Demographic Surveillance System'. *International Journal of Epidemiology* 41 (4): 988–1001.
- Kanjala, Chifundo, Jay Greenfield, David Beckles, and Basia Zaba. 2018. 'ALPHA Network Data Lineage'. Conference presented at the IASSIST and CARTO Conference, Montréal, Canada, May.
- King, Gary. 2007. 'An Introduction to the Dataverse Network as an Infrastructure for Data Sharing'. *Sociological Methods & Research* 36 (2): 173–99.
- Kishamawe, Coleman, Raphael Isingo, Baltazar Mtenga, Basia Zaba, Jim Todd, Benjamin Clark, John Chagalucha, and Mark Urassa. 2015. 'Health & Demographic Surveillance System Profile: The Magu Health and Demographic Surveillance System (Magu HDSS)'. *International Journal of Epidemiology*, dyv188.
- Lazar, Jonathan, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research Methods in Human-Computer Interaction*. Morgan Kaufmann.
- Locke, Joanne, and Alan Lowe. 2007. 'XBRL: An (Open) Source of Enlightenment or Disillusion?' *European Accounting Review* 16 (3): 585–623. <https://doi.org/10.1080/09638180701507163>.
- LOINC. 2019. 'Logical Observation Identifiers Names and Codes'. 2019. <https://loinc.org/>.
- Maelstrom Research. 2019. 'Home Page | Maelstrom Research'. Maelstrom. 2019. <https://www.maelstrom-research.org/>.
- Maher, D, S Biraro, V Hosegood, R Isingo, T Lutalo, P Mushati, B Ngwira, M Nyirenda, J Todd, and B Zaba. 2010. 'Translating Global Health Research Aims into Action: The Example of the ALPHA Network.' *Tropical Medicine & International Health : TM & IH* 15 (3): 321–28. <https://doi.org/10.1111/j.1365-3156.2009.02456.x>.
- Marker, Hans Jørgen, Wolfgang Zenk-Möltgen, Wendy Thomas, and Achim Wackerow. 2009. 'Workflows for Metadata Creation Regarding Recoding, Aggregation and Other Data Processing Activities (2009-03-21)'. Working Paper 4.

- Marston, Milly, Denna Michael, Alison Wringe, Raphael Isingo, Benjamin D Clark, Aswile Jonas, Julius Mngara, et al. 2012. 'The Impact of Antiretroviral Therapy on Adult Mortality in Rural Tanzania'. *Tropical Medicine & International Health* 17 (8): e58—e65. <https://doi.org/10.1111/j.1365-3156.2011.02924.x>.
- McCormick, Tyler H, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark. 2016. 'Probabilistic Cause-of-Death Assignment Using Verbal Autopsies'. *Journal of the American Statistical Association* 111 (515): 1036–49.
- McMahon, Christiana. 2017. 'The Evaluation and Harmonisation of Disparate Information Metamodels in Support of Epidemiological and Public Health Research'. London: University College London. discovery.ucl.ac.uk/10025205/1/McMahon_Final_Corrected_Thesis.pdf.
- Mercè Crosas. 2011. 'The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data'. *D-Lib Magazine*, 2011.
- Metadata Technology North America Inc. 2019. 'Metadata Technology North America'. 2019. <https://www.mtna.us/>.
- Mikkelsen, Lene, David E Phillips, Carla AbouZahr, Philip W Setel, Don De Savigny, Rafael Lozano, and Alan D Lopez. 2015. 'A Global Assessment of Civil Registration and Vital Statistics Systems: Monitoring Data Quality and Progress'. *The Lancet* 386 (10001): 1395–1406.
- Miller, K, and M Vardigan. 2005. 'How Initiative Benefits the Research Community-the Data Documentation Initiative'. In *First International Conference on E-Social Science, Manchester, UK*.
- Murtha Baca. 2008. *Introduction to Metadata*. Second Edition. Los Angeles, CA, USA: Getty Research Institute.
- Næss, Øyvind, Anne Johanne Sogaard, Egil Arnesen, Anne Cathrine Beckstrøm, Espen Bjertness, Anders Engeland, Peter F Hjort, Jostein Holmen, Per Magnus, and Inger Njølstad. 2007. 'Cohort Profile: Cohort of Norway (CONOR)'. *International Journal of Epidemiology* 37 (3): 481–85.
- Ndirangu, James, Marie-Louise Newell, Claire Thorne, and Ruth Bland. 2011. 'Treating HIV-Infected Mothers Reduces under 5 Years of Age Mortality Rates to Levels Seen in Children of HIV-Uninfected Mothers in Rural South Africa.' *Antiviral Therapy* 17 (1): 81–90.
- Nesstar. 2011. 'Nesstar Publisher v4 . 0 User Guide', no. September.
- Nielsen, Mogens, Jeremy Iverson, and Dan Smith. 2013. 'Standardized Quality Declarations with DDI, SDMX, and Colectica'. In .
- NISO. 2004. 'Understanding Metadata'. *National Information Standards Organization*. Vol. 20. Baltimore.
- Norwegian Centre for Research Data. 2016. 'Nesstar Users'. Nesstar - Publish Data on the Web. 2016. <http://www.nesstar.com/about/customers.html>.
- NVivo Qualitative Data Analysis Software (version 12). 2018. QSR International Pty Ltd.
- Odiambo, Frank O, Kayla F Laserson, Maquins Sewe, Mary J Hamel, Daniel R Feikin, Kubaje Adazu, Sheila Ogwang, David Obor, Nyaguara Amek, and Nabie Bayoh. 2012. 'Profile: The KEMRI/CDC Health and Demographic Surveillance System—Western Kenya'. *International Journal of Epidemiology* 41 (4): 977–87.
- O'Neill, Dara, Michaela Benzeval, Andy Boyd, Lisa Calderwood, Cyrus Cooper, Louise Corti, Elaine Dennison, et al. 2019. 'Data Resource Profile: Cohort and Longitudinal Studies Enhancement Resources (CLOSER)'. *International Journal of Epidemiology*, February. <https://doi.org/10.1093/ije/dyz004>.
- Ontotext. 2019. 'Introduction to the Semantic Web — GraphDB Free 8.10 Documentation'. GraphDB 8.10. June 2019.

- <http://graphdb.ontotext.com/documentation/free/introduction-to-semantic-web.html>.
- Ouma, P., E. A. Okiro, and R. W. Snow. 2018. 'Sub-Saharan Public Hospitals Geo-Coded Database'. *Harvard Dataverse* 1: 2018.
- Oxford University Press. 2019. 'Data Resource Profile Series | International Journal of Epidemiology | Oxford Academic'. 2019. https://academic.oup.com/ije/pages/Data_Resource_Profile_Series.
- Pauline Ward. 2016. 'Sources of Dataset Peer Review - Datashare - Wiki Service'. 2016. <https://www.wiki.ed.ac.uk/display/datashare/Sources+of+dataset+peer+review>.
- Pentaho Corporation. 2018. 'Pentaho Data Integration'. Pentaho Documentation. 10 October 2018. https://help.pentaho.com/Documentation/8.2/Products/Data_Integration.
- Pisani, Elizabeth, Peter Aaby, J Gabrielle Breugelmans, David Carr, Trish Groves, Michelle Helinski, Dorcas Kamuya, Steven Kern, Katherine Littler, and Vicki Marsh. 2016. 'Beyond Open Data: Realising the Health Benefits of Sharing Data'. *Bmj* 355: i5295.
- Pisani, Elizabeth, and Carla AbouZahr. 2010. 'Sharing Health Data: Good Intentions Are Not Enough'. *Bulletin of the World Health Organization* 88 (6): 462–466.
- Porter, Kholoud, and Basia Zaba. 2004. 'The Empirical Evidence for the Impact of HIV on Adult Mortality in the Developing World: Data from Serological Studies.' *AIDS* 18 Suppl 2 (suppl 2): S9–S17.
- Pressman, Roger S. 2010. *Software Engineering: A Practitioner's Approach*. 7th ed. New York: McGraw-Hill Higher Education.
- Radler, Barry, Jeremy Iverson, and Dan Smith. 2013. 'Applying DDI to a Longitudinal Study of Aging'. In *North American Data Documentation Initiative Conference (NADDI 2013)*, University of Kansas, Lawrence, Kansas.
- Rasmussen, Karsten Boye. 2014. 'Social Science Metadata and the Foundations of the DDI'. *LASSIST Quarterly* 37 (1): 28–28.
- Rasmussen, Karsten Boye, and Grant Blank. 2007a. 'The Data Documentation Initiative: A Preservation Standard for Research.' *Archival Science* 7 (1): 55–71.
- . 2007b. 'The Data Documentation Initiative: A Preservation Standard for Research.' *Archival Science* 7 (1): 55–71.
- Reniers, Georges, Marylene Wamukoya, Mark Urassa, Amek Nyaguara, Jessica Nakiyingi-Miiro, Tom Lutalo, Vicky Hosegood, Simon Gregson, Xavier Gómez-Olivé, and Eveline Geubbels. 2016. 'Data Resource Profile: Network for Analysing Longitudinal Population-Based HIV/AIDS Data on Africa (ALPHA Network)'. *International Journal of Epidemiology* 45 (1): 83–93.
- Robert W. Snow. 2017. 'A Geo-Coded Inventory of Anophelines in the Afrotropical Region South of the Sahara: 1898-2016'. *Harvard Dataverse*. <https://doi.org/10.7910/DVN/NQ6CUN>.
- Ryssevik, Jostein. 1999. 'PROVIDING GLOBAL ACCESS TO DISTRIBUTED DATA THROUGH METADATA'. In , 22–24. Geneva, Switzerland.
- Sankoh, Osman, and Peter Byass. 2012. 'The INDEPTH Network: Filling Vital Gaps in Global Epidemiology'. *International Journal of Epidemiology* 41 (3): 579–588.
- Sankoh, Osman, David Sharrow, Kobus Herbst, Chodziwadziwa Whiteson Kabudula, Nurul Alam, Shashi Kant, Henrik Ravn, Abbas Bhuiya, Le Thi Vui, and Timotheus Darikwa. 2014. 'The INDEPTH Standard Population for Low-and Middle-Income Countries, 2013'. *Global Health Action* 7.
- Schoenwaelder, Juergen, and Aiko Pras. 2003. 'On the Difference between Information Models and Data Models', 2003.
- SDMX. 2019. 'Validation and Transformation Language (VTL) | SDMX – Statistical Data and Metadata EXchange'. SDMX. 2019. https://sdmx.org/?page_id=5096.

- SDMX Technical Working Group. 2018a. 'Learning | SDMX – Statistical Data and Metadata EXchange'. 2018. https://sdmx.org/?page_id=2555.
- . 2018b. 'What Is SDMX? | SDMX – Statistical Data and Metadata EXchange'. SDMX. 2018. https://sdmx.org/?page_id=3425.
- . 2018c. 'VTL – Version 2.0 (Validation and Transformation Language) Part 1 - User Manual'. <https://sdmx.org/wp-content/uploads/VTL-2.0-User-Manual-20180416-final.pdf>.
- Setel, Philip W, Sarah B Macfarlane, Simon Szreter, Lene Mikkelsen, Prabhat Jha, Susan Stout, Carla AbouZahr, and Monitoring of Vital Events (MoVE) writing group. 2007. 'A Scandal of Invisibility: Making Everyone Count by Counting Everyone'. *The Lancet* 370 (9598): 1569–77.
- Slaymaker, Emma, Estelle McLean, Alison Wringe, Clara Calvert, Milly Marston, Georges Reniers, Chodziwadziwa Whiteson Kabudula, Amelia Crampin, Alison Price, and Denna Michael. 2017. 'The Network for Analysing Longitudinal Population-Based HIV/AIDS Data on Africa (ALPHA): Data on Mortality, by HIV Status and Stage on the HIV Care Continuum, among the General Population in Seven Longitudinal Studies between 1989 and 2014'. *Gates Open Research* 1.
- Snow, Robert W., Benn Sartorius, David Kyalo, Joseph Maina, Punam Amratia, Clara W. Mundia, and Philip Bejon. 2017. 'The Prevalence of Plasmodium Falciparum in Sub-Saharan Africa since 1900'.
- Stahl, Reinhold, and Patricia Staab. 2018. *Measuring the Data Universe*. Springer, Cham.
- StataCorp. 2003. *Stata Statistical Software* (version 8). College Station, TX: StataCorp LP. www.stata.com.
- . 2019. *Stata Statistical Software* (version 16). College Station, TX: StataCorp LP. www.stata.com.
- Streatfield, P Kim, Wasif A Khan, Abbas Bhuiya, Syed MA Hanifi, Nurul Alam, Eric Diboulo, Ali Sié, et al. 2014. 'Malaria Mortality in Africa and Asia: Evidence from INDEPTH Health and Demographic Surveillance System Sites'. *Global Health Action* 7: 10.3402/gha.v7.25369. <https://doi.org/10.3402/gha.v7.25369>.
- Tanser, Frank, Victoria Hosegood, Till Bärnighausen, Kobus Herbst, Makandwe Nyirenda, William Muhwava, Colin Newell, Johannes Viljoen, Tinofa Mutevedzi, and Marie-Louise Newell. 2008. 'Cohort Profile: Africa Centre Demographic Information System (ACDIS) and Population-Based HIV Survey'. *International Journal of Epidemiology* 37 (5): 956–62.
- Taylor, Marie. 2014. *Introduction to Javascript Object Notation: A to-the-Point Guide to JSON*. CreateSpace Independent Publishing Platform.
- The DHS Program. 2019. 'The DHS Program - Quality Information to Plan, Monitor and Improve Population, Health, and Nutrition Programs'. 2019. <https://dhsprogram.com/>.
- The World Bank Group. 2019. 'Microdata Library'. 2019. <https://microdata.worldbank.org/index.php/home>.
- Thérèse Lalor, and Steven Vale. 2013. 'Modernising the Production of Official Statistics' 4 (2): 72–79.
- Todd, Jim, Judith R Glynn, Milly Marston, Tom Lutalo, Sam Biraro, Wambura Mwita, Vinai Suriyanon, Ram Rangsins, Kenrad E Nelson, and Pam Sonnenberg. 2007. 'Time from HIV Seroconversion to Death: A Collaborative Analysis of Eight Studies in Six Low and Middle-Income Countries before Highly Active Antiretroviral Therapy'. *Aids* 21: S55–63.
- Turbelin, Clément, and Pierre-Yves Boëlle. 2013. 'Open Data in Public Health Surveillance Systems: A Case Study Using the French Sentinelles Network'. *International Journal of Medical Informatics* 82 (10): 1012–21. <https://doi.org/10.1016/j.ijmedinf.2013.06.009>.

- UNECE. 2018a. 'Generic Statistical Information Model GSim): Communication Paper for a General Statistical Audience (Version 1.1, December 2013)'. United Nations Economic Commission for Europe (UNECE). <https://statswiki.unece.org/display/gsim/GSIM+Communication+Paper>.
- . 2018b. 'GSBPM v5.0 - Generic Statistical Business Process Model - UNECE Statswiki'. August 2018. <https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0>.
- . 2018c. 'ABS Prices Processes Mapped to GSBPM - Generic Statistical Business Process Model - UNECE Statswiki'. 16 August 2018. <https://statswiki.unece.org/display/GSBPM/ABS+Prices+processes+mapped+to+GSBPM>.
- . 2019a. 'GSIM on a Page - GSIM on a Page - UNECE Statswiki'. 2019. <https://statswiki.unece.org/display/clickablegsim/GSIM+on+a+page>.
- . 2019b. 'High-Level Group for the Modernisation of Official Statistics - High-Level Group for the Modernisation of Official Statistics - UNECE Statswiki'. 2019. <https://statswiki.unece.org/display/hlgbas>.
- UNECE Secretariat. 2009. 'Generic Statistical Business Process Model: Version 4.0–April 2009'.
- UNICEF MICS. 2019. 'Multiple Indicator Cluster Survey (MICS)'. 2019. <http://mics.unicef.org/>.
- United Nations. Statistical Division. 2001. *Principles and Recommendations for a Vital Statistics System*. Vol. 836. United Nations Publications.
- Urassa, M, J T Boerma, R Isingo, J Ngalula, J Ng'weshemi, G Mwaluko, and B Zaba. 2001. 'The Impact of HIV/AIDS on Mortality and Household Mobility in Rural Tanzania.' *AIDS* 15 (15): 2017–2023.
- Van den Eynden, Veerle, Louise Corti, Matthew Woollard, Libby Bishop, and Laurence Horton. 2011. 'Managing and Sharing Data; a Best Practice Guide for Researchers'.
- Van Panhuis, Willem G, Proma Paul, Claudia Emerson, John Grefenstette, Richard Wilder, Abraham J Herbst, David Heymann, and Donald S Burke. 2014. 'A Systematic Review of Barriers to Data Sharing in Public Health'. *BMC Public Health* 14 (1): 1144.
- Vardigan, Mary, Pascal Heus, and Wendy Thomas. 2008. 'Data Documentation Initiative: Toward a Standard for the Social Sciences'. *International Journal of Digital Curation* 3 (1).
- Vet, Edwin de. 2013. 'Update on Questasy, a Data Dissemination Tool Based on DDI3'. In *EDDI13–5th Annual European DDI User Conference*. <http://www.eddi-conferences.eu/ocs/index.php/eddi/EDDI13/paper/view/71>.
- Visual Paradigm. n.d. 'What Is Unified Modeling Language (UML)?' Visual Paradigm. Accessed 28 August 2019. <https://www.visual-paradigm.com/guide/uml-unified-modeling-language/what-is-uml/>.
- W3schools.com. 2013. 'Introduction to HTML'. 2013. http://www.w3schools.com/html/html_intro.asp.
- . 2015. 'XML Introduction - What Is XML?' 2015. http://www.w3schools.com/xml/xml_what.asp.
- Walport, Mark, and Paul Brest. 2011. 'Sharing Research Data to Improve Public Health'. *The Lancet* 377 (9765): 537–39.
- Weisfeld, Matt. 2008. *The Object-Oriented Thought Process*. Pearson Education.
- Welch, Matthew John, and Mehmood Asghar. 2018. 'Using DDI to Build Open Source Solutions for Curation and Dissemination'. presented at the North American Data Documentation Initiative Conference (NADDI) 2018 Annual Conference (NADDI2018), Washington, DC, April 5. <https://zenodo.org/record/1217515>.
- Wellcome Trust. 2014. 'Enhancing Discoverability of Public Health and Epidemiology Research Data'. Wellcome Trust.

- <https://wellcome.ac.uk/sites/default/files/enhancing-discoverability-of-public-health-and-epidemiology-research-data-phrdf-jul14.pdf>.
- Weske, Mathias. 2007. *Business Process Management: Concepts, Languages, Architectures*. Berlin ; New York: Springer.
- Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, and Philip E Bourne. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3.
- William Block, Thomas Bosch, Bryan Fitzpatrick, Dan Gillman, Jay Greenfield, Arofan Gregory, Marcel Hebing, et al. 2012. 'Developing a Model-Driven DDI Specification'. Workshop Paper. DDI Lifecycle: Moving Forward. Wadern, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH. www.ddialliance.org/system/files/DevelopingaModel-DrivenDDISpecification2013_05_15.pdf.
- Winters, Kristi, and Sebastian Netscher. 2016. 'Proposed Standards for Variable Harmonization Documentation and Referencing: A Case Study Using QuickCharmStats 1.1'. *PloS One* 11 (2): e0147795.
- World Health Organization. 2016. *ICD 10 - International Statistical Classification of Diseases and Related Health Problems*. 5th ed. Vol. 1.
- Zaba, Basia, Clara Calvert, Milly Marston, Raphael Isingo, Jessica Nakiyingi-Müro, Tom Lutalo, Amelia Crampin, Laura Robertson, Kobus Herbst, and Marie-Louise Newell. 2013a. 'Effect of HIV Infection on Pregnancy-Related Mortality in Sub-Saharan Africa: Secondary Analyses of Pooled Community-Based Data from the Network for Analysing Longitudinal Population-Based HIV/AIDS Data on Africa (ALPHA)'. *The Lancet* 381 (9879): 1763–71.
- . 2013b. 'Effect of HIV Infection on Pregnancy-Related Mortality in Sub-Saharan Africa: Secondary Analyses of Pooled Community-Based Data from the Network for Analysing Longitudinal Population-Based HIV/AIDS Data on Africa (ALPHA)'. *The Lancet* 381 (9879): 1763–71.

APPENDIX A METADATA INFUSION FILE SCHEMA

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="infusion_6.1">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="credits"/>
        <xs:element ref="concepts"/>
        <xs:element ref="study"/>
        <xs:element ref="datapipeline"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="credits">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="title"/>
        <xs:element ref="abstract"/>
        <xs:element ref="sources"/>
        <xs:element ref="creator"/>
        <xs:element ref="contributors"/>
        <xs:element ref="resources"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="title" type="xs:string"/>
  <xs:element name="abstract" type="xs:string"/>
  <xs:element name="concepts">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="concept" minOccurs="0" maxOccurs="unbounded"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="concept">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="agency"/>
        <xs:element ref="id"/>
        <xs:element ref="version"/>
        <xs:element ref="name"/>
        <xs:element ref="definition"/>
        <xs:element ref="localid" minOccurs="0" maxOccurs="unbounded"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="agency" type="xs:string"/>
  <xs:element name="id" type="xs:string"/>
  <xs:element name="version" type="xs:string"/>
  <xs:element name="definition" type="xs:string"/>
  <xs:element name="localid" type="xs:string"/>
  <xs:element name="sources">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" ref="source"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="source">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="citationrole"/>
        <xs:element ref="uri"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

```



```

        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="citationrole" type="xs:string"/>
<xs:element name="creator" type="name"/>
<xs:element name="contributors">
    <xs:complexType>
        <xs:sequence>
            <xs:element maxOccurs="unbounded" ref="contributor"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="contributor">
    <xs:complexType>
        <xs:complexContent>
            <xs:extension base="name">
                <xs:sequence>
                    <xs:element ref="role"/>
                </xs:sequence>
            </xs:extension>
        </xs:complexContent>
    </xs:complexType>
</xs:element>
<xs:element name="role">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="description"/>
            <xs:element ref="value"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="value" type="xs:NCName"/>
<xs:element name="resources">
    <xs:complexType>
        <xs:sequence>
            <xs:element maxOccurs="unbounded" ref="resource"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="resource">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="description"/>
            <xs:element ref="uri"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="study">
    <xs:complexType>
        <xs:complexContent>
            <xs:extension base="name">
                <xs:sequence>
                    <xs:element ref="overview"/>
                    <xs:element ref="unittype"/>
                    <xs:element ref="population"/>
                    <xs:element ref="methodologyoverview"/>
                    <xs:element ref="designoverview"/>
                    <xs:element ref="algorithmoverview"/>
                    <xs:element ref="coverage"/>
                </xs:sequence>
            </xs:extension>
        </xs:complexContent>
    </xs:complexType>
</xs:element>
<xs:element name="unittype">
    <xs:complexType/>
</xs:element>
<xs:element name="population">

```



```

    <xs:complexType/>
  </xs:element>
  <xs:element name="methodologyoverview">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="overview"/>
        <xs:element ref="externalmaterials"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="designoverview">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="overview"/>
        <xs:element ref="externalmaterials"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="coverage">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="overview"/>
        <xs:element ref="temporal"/>
        <xs:element ref="spatial"/>
        <xs:element ref="topical"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="temporal">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="begins"/>
        <xs:element ref="ends"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="begins">
    <xs:complexType/>
  </xs:element>
  <xs:element name="ends">
    <xs:complexType/>
  </xs:element>
  <xs:element name="spatial">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" ref="spatialareacode"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="spatialareacode">
    <xs:complexType/>
  </xs:element>
  <xs:element name="topical">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" ref="keyword"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="keyword">
    <xs:complexType/>
  </xs:element>
  <xs:element name="datapipeline">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" ref="bp"/>
        <xs:element ref="attribution"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

```



```

    </xs:complexType>
  </xs:element>
  <xs:element name="bp">
    <xs:complexType>
      <xs:complexContent>
        <xs:extension base="name">
          <xs:sequence>
            <xs:element ref="alias" minOccurs="0" maxOccurs="1"/>
            <xs:element ref="purpose" minOccurs="0" maxOccurs="1"/>
            <xs:element ref="standardmodelused"/>
            <xs:element ref="algorithmoverview"/>
            <xs:element ref="preconditions"/>
            <xs:element ref="postconditions"/>
          </xs:sequence>
          <xs:attribute name="id" use="required" type="xs:integer"/>
        </xs:extension>
      </xs:complexContent>
    </xs:complexType>
  </xs:element>
  <xs:element name="alias" type="xs:string" />
  <xs:element name="purpose" type="xs:string" />
  <xs:element name="standardmodelused">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="uppermodel"/>
        <xs:element ref="lowermodel"/>
        <xs:element ref="curation"/>
        <xs:element minOccurs="0" ref="vocabulary"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="uppermodel">
    <xs:complexType>
      <xs:complexContent>
        <xs:extension base="name">
          <xs:sequence>
            <xs:element ref="step"/>
          </xs:sequence>
        </xs:extension>
      </xs:complexContent>
    </xs:complexType>
  </xs:element>
  <xs:element name="lowermodel">
    <xs:complexType>
      <xs:complexContent>
        <xs:extension base="name">
          <xs:sequence>
            <xs:element ref="step"/>
          </xs:sequence>
        </xs:extension>
      </xs:complexContent>
    </xs:complexType>
  </xs:element>
  <xs:element name="curation">
    <xs:complexType>
      <xs:complexContent>
        <xs:extension base="name">
          <xs:sequence>
            <xs:element ref="step"/>
          </xs:sequence>
        </xs:extension>
      </xs:complexContent>
    </xs:complexType>
  </xs:element>
  <xs:element name="vocabulary">
    <xs:complexType>
      <xs:complexContent>
        <xs:extension base="name">

```



```

        <xs:sequence>
            <xs:element ref="step"/>
        </xs:sequence>
    </xs:extension>
</xs:complexContent>
</xs:complexType>
</xs:element>
<xs:element name="preconditions">
    <xs:complexType>
        <xs:sequence>
            <xs:element minOccurs="0" maxOccurs="unbounded" ref="precondition"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="precondition">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="recordname"/>
            <xs:element ref="location"/>
            <xs:element ref="type"/>
            <xs:element ref="aggregate"/>
            <xs:element minOccurs="0" ref="datadescription"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="postconditions">
    <xs:complexType>
        <xs:sequence>
            <xs:element minOccurs="0" maxOccurs="unbounded" ref="postcondition"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="postcondition">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="recordname"/>
            <xs:element ref="location"/>
            <xs:element ref="type"/>
            <xs:element ref="aggregate"/>
            <xs:element minOccurs="0" maxOccurs="unbounded" ref="datadescription"/>
            <xs:element minOccurs="0" ref="dataexceptions"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="attribution">
    <xs:complexType>
        <xs:sequence>
            <xs:element maxOccurs="unbounded" ref="entity"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>
<xs:element name="entity">
    <xs:complexType>
        <xs:complexContent>
            <xs:extension base="name">
                <xs:sequence>
                    <xs:element ref="abbreviation"/>
                    <xs:element ref="description"/>
                </xs:sequence>
            </xs:extension>
        </xs:complexContent>
    </xs:complexType>
</xs:element>
<xs:element name="abbreviation" type="xs:NCName"/>
<xs:element name="uri" type="xs:anyURI"/>
<xs:complexType name="name">
    <xs:sequence>
        <xs:element ref="name"/>

```



```

    </xs:sequence>
</xs:complexType>
<xs:element name="name" type="xs:string"/>
<xs:element name="description" type="xs:string"/>
<xs:element name="overview">
  <xs:complexType>
    <xs:sequence>
      <xs:element minOccurs="0" maxOccurs="unbounded" ref="para"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="para" type="xs:string"/>
<xs:element name="externalmaterials">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="uniquename"/>
      <xs:element ref="description"/>
      <xs:element ref="uri"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="uniquename">
  <xs:complexType/>
</xs:element>
<xs:element name="algorithmoverview">
  <xs:complexType>
    <xs:choice>
      <xs:element minOccurs="0" maxOccurs="unbounded" ref="step"/>
      <xs:sequence>
        <xs:element ref="overview"/>
        <xs:element ref="externalmaterials"/>
      </xs:sequence>
    </xs:choice>
  </xs:complexType>
</xs:element>
<xs:element name="step">
  <xs:complexType mixed="true">
    <xs:attribute name="id" type="xs:NMTOKEN"/>
    <xs:attribute name="name"/>
  </xs:complexType>
</xs:element>
<xs:element name="recordname" type="xs:NCName"/>
<xs:element name="location" type="xs:string"/>
<xs:element name="type" type="xs:NCName"/>
<xs:element name="aggregate" type="xs:boolean"/>
<xs:element name="datadescription" >
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="sibling"/>
      <xs:element ref="uri"/>
      <xs:element ref="description"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="dataexceptions" >
  <xs:complexType>
    <xs:sequence>
      <xs:element name="dropped" type="xs:NCName" minOccurs="0" maxOccurs="unbounded"/>
      <xs:element ref="renamed" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="sibling" type="xs:NCName"/>
<xs:element name="renamed">
  <xs:complexType>
    <xs:sequence>
      <xs:element minOccurs="1" maxOccurs="1" ref="oldname"/>
      <xs:element minOccurs="1" maxOccurs="1" ref="newname"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

```



```
        </xs:sequence>
      </xs:complexType>
    </xs:element>

    <xs:element name="oldname" type="xs:NCName"/>
    <xs:element name="newname" type="xs:NCName"/>
  </xs:schema>
```


APPENDIX B INFUSION FILE

```
<?xml version="1.0" encoding="UTF-8"?>
<infusion_6.1>
  <credits>
    <title>A program generated metadata description of a Pentaho Data
Integration data processing pipeline</title>
    <abstract>An XSLT program was used to merge potentially three metadata
sources into a DDI4 compliant DataManagementView. One source is Pentaho XML
users generate when
      creating a data pipeline in the Pentaho authoring environment. The
second source is a metadata infusion that adds both shape and detail to the
Pentaho XML. This metadata
      infusion is in part authored by a data scientist. However, the
infusor is "greedy" and is also is able to draw in legacy metadata from
Nesstar Publisher. The XSLT program
      takes these potentially three metadata sources as input and
produces as output a DDI4 DataManagementView. The DataManagementView and its
DataPipeline is able to traverse
      either prospectively or retrospectively a business process model
once to describe the data lifecycle of a Study and many times to describe the
data lifecycle of a StudySeries. </abstract>
    <sources>
      <source>
        <citationrole>The CORE Master Job and the child job that it
references provide metadata Pehntaho uses at run time to orchestrate the study
data pipeline</citationrole>

<uri>https://www.dropbox.com/s/m88qfv4ry8r8hd/CORE%20Master%20Job.kjb?dl=0</u
ri>

      </source>
      <source>
        <citationrole>The core_bpm_mapper provides additional
metadata. It is a metadata infusion that facilitates pipeline understanding
and data discovery</citationrole>

<uri>https://www.dropbox.com/s/b7v8ep4wscgf0yf/core_bpm_mapper.xml?dl=0</uri>

      </source>
      <source>
        <citationrole>Depending on its availability in a given
project, The infusor can consume Nesstar Publisher DDI 2.x compliant
descriptions of unit and dimensional data</citationrole>
        <uri>http://www.nesstar.com</uri>
      </source>
    </sources>
    <creator>
      <name>datapipeline3.xsl</name>
    </creator>
    <contributors>
      <contributor>
        <name>Chifundo Kanjala</name>
        <role>
          <description>Metadata architect, ETL architect and
demographic and epidemiological surveillance domain expert</description>
          <value>Equal</value>
        </role>
      </contributor>
      <contributor>
        <name>Jay Greenfield</name>
        <role>
          <description>Metadata architect and data
scientist</description>
          <value>Equal</value>
        </role>
      </contributor>
      <contributor>
        <name>Tathagata Bhattacharjee</name>
```



```

        <role>
            <description>Pentaho data integration platform
architect</description>
            <value>Equal</value>
        </role>
    </contributor>
</contributors>
<resources>
    <resource>
        <description>Demographic and epidemiological surveillance
reference model</description>

<uri>https://www.dropbox.com/s/gmwug087g92wwgb/INDEPTH%20Monograph%20I%20Ch1-
7%20Introduction%2C%20Methods%20%26%20Life%20Tables.pdf?dl=0</uri>
    </resource>
    <resource>
        <description>The DataManagementView</description>

<uri>https://www.dropbox.com/s/x35kiaitw0krzep/The%20DataManagementView.docx?d
l=0</uri>
    </resource>
</resources>
</credits>
<concepts>
    <concept>
        <agency>ALPHA</agency>
        <id>1</id>
        <version>1</version>
        <name>health-and-demographic-surveillance-system</name>
        <definition>A Health and Demographic Surveillance System(HDSS) is
defined as a set of field and computing operations applied within a clearly
demarcated geographic area to handle
            the longitudinal follow-up of well-defined entities or primary
subjects (individuals, social units (e.g. households), and residential units
(physical locations))
            and their related demographic, socio-economic and health
outcomes.</definition>
        <localid></localid>
        <localid></localid>
    </concept>
    <concept>
        <agency>ALPHA</agency>
        <id>2</id>
        <version>1</version>
        <name>demographic-surveillance-area</name>
        <definition>The demographic surveillance area (DSA) is an area
with clearly and fairly permanent delineated boundaries, preferably
recognizable on the ground (for example, rivers,
            roads, and clearly demarcated administrative
boundaries).</definition>
        <localid></localid>
        <localid></localid>
    </concept>
    <concept>
        <agency>ALPHA</agency>
        <id>3</id>
        <version>1</version>
        <name>initial-census</name>
        <definition>Data collection done within the DSA to register and
define the target population. During the initial census an extensible system
of
            unique identifiers is assigned to the primary entities of the
HDSS. The initial census also works as a basis for the development of the HDSS
database</definition>
        <localid></localid>
        <localid></localid>
    </concept>
</concept>

```



```

        <agency>ALPHA</agency>
        <id>4</id>
        <version>1</version>
        <name>unique-identifier</name>
        <definition>A number or identification code used to uniquely
identify the primary entities of an HDSS. Unique identifiers are assigned at
baseline and they should be
        extensible to accommodate the addition of future primary
entities</definition>
        <localid></localid>
        <localid></localid>
    </concept>
    <concept>
        <agency>ALPHA</agency>
        <id>5</id>
        <version>1</version>
        <name>longitudinality</name>
        <definition>The longitudinal measurement of demographic and health
dynamics in the registered population by constantly updating a set of
prescribed attributes
        for the primary subjects during rounds of follow up visits to
the registered residents in the DSA</definition>
        <localid></localid>
        <localid></localid>
    </concept>
    <concept>
        <agency>ALPHA</agency>
        <id>6</id>
        <version>1</version>
        <name>primary-hdss-entities</name>
        <definition>Residential units, Social units and individuals
comprise the main entities of interest in an HDSS. </definition>
        <localid></localid>
        <localid></localid>
    </concept>
    <concept>
        <agency>ALPHA</agency>
        <id>7</id>
        <version>1</version>
        <name>residential-units</name>
        <definition>Residential units are the places in the DSA where
individuals live</definition>
        <localid></localid>
        <localid></localid>
    </concept>
    <concept>
        <agency>ALPHA</agency>
        <id>8</id>
        <version>1</version>
        <name>residency</name>
        <definition>The state of being physically present in a given
residential unit for a defined threshold of time. Residency is an essential
pre-requisite
        for enumeration of individuals at risk of demographic events
or disease exposure. Residency associates individuals with residential
units</definition>
        <localid></localid>
        <localid></localid>
    </concept>
    <concept>
        <agency>ALPHA</agency>
        <id>9</id>
        <version>1</version>
        <name>social-units</name>
        <definition>these are groups to which individuals in the DSA
belong</definition>
        <localid></localid>
        <localid></localid>

```



```

    </concept>
    <concept>
      <agency>ALPHA</agency>
      <id>10</id>
      <version>1</version>
      <name>membership</name>
      <definition>The state of belonging to a social group. Membership
associates individuals with social groups such as households</definition>
      <localid></localid>
      <localid></localid>
    </concept>
    <concept>
      <agency>ALPHA</agency>
      <id>11</id>
      <version>1</version>
      <name>individuals</name>
      <definition>Individuals are the subject of primary interest in an
HDSS. They are the people who are residents of a residential unit or members
of a social unit</definition>
      <localid></localid>
      <localid></localid>
    </concept>
    <concept>
      <agency>ALPHA</agency>
      <id>12</id>
      <version>1</version>
      <name>eligibility</name>
      <definition>Every HDSS defines a population under surveillance.
Individuals have places of residence and attachments to social units such as
households.
      First the residential units, the social groups and the
individuals need to be identified. A set of inclusion criteria to distinguish
eligible from
      ineligible entities needs to be defined. Residential and
social units are eligible if they are situated in the DSA. Individuals are
eligible if they are resident
      at eligible residential units or if they belong to eligible
social units</definition>
      <localid></localid>
      <localid></localid>
    </concept>
    <concept>
      <agency>ALPHA</agency>
      <id>13</id>
      <version>1</version>
      <name>hdss-core-events</name>
      <definition>Three core events alter the size of the registered
population in an HDSS. These are births, deaths and migration in consistency
with the fundamental demographic equation:
      
$$Pt1 = Pt0 + Bt0, t1 - Dt0, t1 + It0, t1 - Ot0, t1.$$

      Pt1 -Population
at time t1, Pt0 - Population at time t2, Bto, t1 - Births between time t0 and
time t1,
      Dto, t1 - Deaths between time t0 and time t1, Ito, t1 -
Immigrations between time t0 and time t1, Oto, t1 - Outmigration between time
t0 and time t1</definition>
      <localid></localid>
      <localid></localid>
    </concept>
    <concept>
      <agency>ALPHA</agency>
      <id>14</id>
      <version>1</version>
      <name>birth</name>
      <definition>Pregnancies and their outcomes are recorded for all
registered women in the HDSS regardless of the location at which the outcomes
occur. An HDSS aims to
      record all outcomes including miscarriages, induced abortions,
stillbirths and live births. All pregnancy outcomes are needed for fertility

```



```

estimation
    All live births are registered as individual members of the
HDSS</definition>
    <localid></localid>
    <localid></localid>
</concept>
<concept>
    <agency>ALPHA</agency>
    <id>15</id>
    <version>1</version>
    <name>death</name>
    <definition>All deaths to registered members of the HDSS are
recorded regardless of the place of occurrence of the death</definition>
    <localid></localid>
    <localid></localid>
</concept>
<concept>
    <agency>ALPHA</agency>
    <id>16</id>
    <version>1</version>
    <name>migration</name>
    <definition>The change of residence by a registered individual or
social group (e.g., a household). There are two types of migration that occur
among the registered population.
        These are internal and external migration. </definition>
    <localid></localid>
    <localid></localid>
</concept>
<concept>
    <agency>ALPHA</agency>
    <id>17</id>
    <version>1</version>
    <name>internal-migration </name>
    <definition>This involves residence changes from one residential
unit to another in the same DSA. While internal migration does not alter the
size of the registered population,
        it is essential to record internal migration to avoid double
counting of individuals and to also correctly apportion exposure to social and
physical environment.
        Decisions have to be made for deciding if migration has
occurred based on the duration since the event.. </definition>
    <localid></localid>
    <localid></localid>
</concept>
<concept>
    <agency>ALPHA</agency>
    <id>18</id>
    <version>1</version>
    <name>external-migration </name>
    <definition>This involves residence changes between a residential
unit in a DSA and one that's outside it.</definition>
    <localid></localid>
    <localid></localid>
</concept>
<concept>
    <agency>ALPHA</agency>
    <id>19</id>
    <version>1</version>
    <name>external-immigration </name>
    <definition>This is when an individual or social unit migrates
into the DSA from a location outside the DSA</definition>
    <localid></localid>
    <localid></localid>
</concept>
<concept>
    <agency>ALPHA</agency>
    <id>20</id>
    <version>1</version>

```



```

        <name>internal-immigration </name>
        <definition>Relocating into a residential unit within the DSA
after exiting another also within the DSA</definition>
        <localid></localid>
        <localid></localid>
    </concept>
    <concept>
        <agency>ALPHA</agency>
        <id>21</id>
        <version>1</version>
        <name>internal-outmigration </name>
        <definition>Exiting a residential unit in the DSA in order to join
another also within the DSA </definition>
        <localid></localid>
        <localid></localid>
    </concept>
    <concept>
        <agency>ALPHA</agency>
        <id>22</id>
        <version>1</version>
        <name>external-outmigration </name>
        <definition>When an individual relocates from a residential unit
in the DSA to a place outside the DSA </definition>
        <localid></localid>
        <localid></localid>
    </concept>
    <concept>
        <agency>ALPHA</agency>
        <id>23</id>
        <version>1</version>
        <name>episodes</name>
        <definition>Episodes are the logical complement to events. They
are meaningful and identifiable segments of time started and ended by
events.</definition>
        <localid></localid>
        <localid></localid>
    </concept>
    <!-- id is the identifier given by the agency that specified the
concept -->
    <!-- the agency may have created multiple versions of a concept -->
    <!-- I will search on the name parsing the text we want to tag -->
    <!-- Except if there are one or more localids: in this event I will
search on the localids -->
</concepts>
<study>
    <name>Health and Demographic Surveillance System (HDSS)</name>
    <overview>
        <para>
            The data that are used to create the ALPHA specifications come
from health-and-demographic-surveillance-system (HDSS) studies being conducted
in
            Eastern and Southern Africa where HIV is a major public health
problem. In these settings, national vital registration and population-based
health
            information are scarce. HDSS are an attempt to address this
dearth of data.
        </para>
        <para>
            Development of structured documentation for ALPHA
specifications inevitably refers to the concepts that are core to HDSS. In
addition, it also refers to objects, relationships and
            attributes from the HDSS reference data model.
        </para>
    </overview>
    <!-- UnitType is the most general in the hierarchy of UnitType,
Universe, and Population.
    It is a description of the basic characteristic for a general
set of Units.

```



```

        A Universe is a set of entities defined by a more narrow
specification than that of an underlying UnitType.
        A Population further narrows the specification to a specific
time and geography -->
        <unittype></unittype>
        <!--Set of specific units (people, entities, objects, events) with
specification of time and geography -->
        <population></population>
        <methodologyoverview>
            <overview>
                <para>
                    The ADESBPM extends both the UNECE Generic Statistical
Business Process Model (GSBPM) and DDI's Generic Longitudinal
Business Process Model (GLBPM).
                </para>
                <para>
                    The extensions make the subject matter of each business
process or step more domain specific by introducing as applicable and
                    chronicling the entities of the Health and Demographic
Surveillance System (HDSS) reference data model. In line with the HDSS,
                    ADESBPM tells a story about the production of demographic
and epidemiological events and episodes.
                </para>
            </overview>
            <!-- repeat external materials as needed -->
            <externalmaterials>
                <uniquename></uniquename>
                <description></description>
                <uri></uri>
            </externalmaterials>
        </methodologyoverview>
        <designoverview>
            <overview>
                <para>
                    The design of the ADESBPM is proceeding bottom up based on
two use cases. The first use case is a demographic surveillance one. INDEPTH
                    has defined a set of business processes and
transformations using the Pentaho Integration platform to perform demographic
surveillance. ALPHA
                    in turn has leveraged INDEPTH. It is in the process of
defining a set of business processes and transformations using the Pentaho
Integration platform
                    to perform successively first demographic surveillance
and then epidemiological surveillance.
                </para>
                <para>
                    There is the possibility that INDEPTH demographic
surveillance will have to be tweaked to support ALPHA epidemiological
surveillance. This is
                    what we are in the process of discovering.
                </para>
            </overview>
            <!-- repeat external materials as needed -->
            <externalmaterials>
                <uniquename></uniquename>
                <description></description>
                <uri>test</uri>
            </externalmaterials>
        </designoverview>
        <algorithmoverview>
            <overview></overview>
            <!-- repeat external materials as needed -->
            <externalmaterials>
                <uniquename></uniquename>
                <description></description>
                <uri></uri>
            </externalmaterials>
        </algorithmoverview>

```



```

    <coverage>
      <overview></overview>
      <temporal>
        <begins></begins>
        <ends></ends>
      </temporal>
      <spatial>
        <spatialareacode></spatialareacode>
        <spatialareacode></spatialareacode>
      </spatial>
      <topical>
        <keyword></keyword>
        <keyword></keyword>
      </topical>
    </coverage>
  </study>
  <datapipeline>
    <bp id="1">
      <name>01 Site Specific ETL for 6.1</name>
      <alias>6.1/Nairobi/Nairobi 00 Generating Staging Tables</alias>
      <purpose>Creates staging tables from member centre specific data.
The staging tables are then transformed further to create the ALPHA
specification 6.1</purpose>
      <standardmodelused>
        <uppermodel>
          <name>Generic Longitudinal Business Process Model</name>
          <step>5.1 Integrate data</step>
        </uppermodel>
        <lowermodel>
          <name>African Demographic and Epidemiological Surveillance
Business Process Model (ADESBPM)</name>
          <step>Transform centre's operational data into relevant
entities of the HDSS reference data model</step>
          <!-- <step>Create event-specific and event-related staging
tables</step> -->
        </lowermodel>
        <curation>
          <name>Open Archival Information System Reference
Model</name>
          <step>Submission Information Package</step>
        </curation>
        <!-- repeat as needed vocabulary entries -->
        <!-- A vocabulary name is the name of a controlled vocabulary
or ontology -->
        <!-- A vocabulary step is the name of an "element" or concept
in the controlled vocabulary together with its its path -->
        <vocabulary>
          <name></name>
          <step></step>
        </vocabulary>
      </standardmodelused>
      <!-- algorithmoverview steps take names as needed -->
      <algorithmoverview>
        <step id="1.1" name=""></step>
        <step id="1.2" name=""></step>
      </algorithmoverview>
      <!-- preconditions here are site specific I am thinking -->
      <preconditions/>
      <postconditions>
      </postconditions>
    </bp>
    <bp id="2">
      <name>02 Core ETL for Raw 6.1</name>
      <alias>6.1/CORE/001 CORE Produce Raw 61 Dataset</alias>
      <purpose>Creates ALPHA specification 6.1 in event format from
staging tables created in the site specific ETL business process (01 Site
Specific ETL for 6.1)</purpose>
      <standardmodelused>

```



```

        <uppermodel>
            <name>Generic Longitudinal Business Process Model</name>
            <step>5.1 Integrate data</step>
        </uppermodel>
        <lowermodel>
            <name>African Demographic and Epidemiological Surveillance
Business Process Model (ADESBPM)</name>
            <step>Transform relevant entities of the HDSS reference
data model into harmonised data (ALPHA Spec 6.1) </step>
        </lowermodel>
        <curation>
            <name>Open Archival Information System Reference
Model</name>
            <step>Submission Information Package</step>
        </curation>
    </standardmodelused>
    <algorithmoverview>
        <step id="2.1" name="Generate anonymised IDs">Generate
anonymised unique-identifiers</step>
        <step id="2.2" name="Map original unique-identifiers to
anonymised IDs">Create a mapping between original and anonymised IDs</step>
        <step id="2.3" name="Store the mapping between original and
anonymised IDs">Store the IDs mapping information where it can be accessed
internally in the future</step>
        <step id="2.4" name="Create Raw Spec 6.1 from staging
data">Create Raw Spec 6.1 from staging data</step>
    </algorithmoverview>
    <preconditions>
        <!--<precondition>
            <recordname>Staging_Raw_6.1</recordname>
            <location>01 Generate ALPHA 6.1 Data Specs.ktr</location>
            <type>CubeOutput</type>
            <aggregate>false</aggregate>
            <datadescription>An excel spreadsheet</datadescription>
        </precondition-->
    </preconditions>
    <postconditions>
        <postcondition>
            <recordname>Indv_ID_Anonymise_Map</recordname>
            <location>6.1/CORE/001 CORE Produce Raw 61
Dataset.kjb</location>
            <type>SQL</type>
            <aggregate>false</aggregate>
            <datadescription>
                <sibling>Indv_ID_Anonymise_Map</sibling>
                <uri>6.1/CORE/01A CORE Anonymise Individual
Id.ktr</uri>
                <description>Stata Output</description>
            </datadescription>
        <!--
            <!--\- example data exception -\->
        <dataexceptions>
            <dropped>ccc</dropped>
            <dropped>eee</dropped>
            <renamed>
                <oldname>aaa</oldname>
                <newname>bbb</newname>
            </renamed>
            <renamed>
                <oldname>ccc</oldname>
                <newname>ddd</newname>
            </renamed>
        </dataexceptions-->
    </postcondition>
    <postcondition>
        <recordname>Indv_ID_Anonymise_Map</recordname>
        <location>6.1/CORE/01A CORE Anonymise Individual
Id.ktr</location>
        <type>StataOutput</type>

```



```

        <aggregate>false</aggregate>
        <!-- example data description -->
        <datadescription>
            <sibling>Indv_ID_Anonymise_Map</sibling>
            <uri>6.1/CORE/001 CORE Produce Raw 61
Dataset.kjb</uri>
            <description>SQL</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>HH_ID_Anonymise_Map</recordname>
        <location>6.1/CORE/001 CORE Produce Raw 61
Dataset.kjb</location>
        <type>SQL</type>
        <aggregate>false</aggregate>
        <datadescription>
            <sibling>HH_ID_Anonymise_Map</sibling>
            <uri>6.1/CORE/01B CORE Anonymise Household
Id.ktr</uri>
            <description>Stata Output</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>HH_ID_Anonymise_Map</recordname>
        <location>6.1/CORE/01B CORE Anonymise Household
Id.ktr</location>
        <type>StataOutput</type>
        <aggregate>false</aggregate>
        <datadescription>
            <sibling>HH_ID_Anonymise_Map</sibling>
            <uri>6.1/CORE/001 CORE Produce Raw 61
Dataset.kjb</uri>
            <description>SQL</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>Raw_61_Event_Format</recordname>
        <location>6.1/CORE/001 CORE Produce Raw 61
Dataset.kjb</location>
        <type>SQL</type>
        <aggregate>false</aggregate>
        <datadescription>
            <sibling>Raw_61_Event_Format</sibling>
            <uri>6.1/CORE/01C CORE Generate Raw 6.1.ktr</uri>
            <description>Stata Output</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>Raw_61_Event_Format</recordname>
        <location>6.1/CORE/01C CORE Generate Raw
6.1.ktr</location>
        <type>StataOutput</type>
        <aggregate>false</aggregate>
        <datadescription>
            <sibling>Raw_61_Event_Format</sibling>
            <uri>6.1/CORE/001 CORE Produce Raw 61
Dataset.kjb</uri>
            <description>SQL</description>
        </datadescription>
    </postcondition>
    </postconditions>
</bp>
<bp id="3">
    <name>03 Core ETL Raw 6.1 Dataset Quality Metrics</name>
    <alias>6.1/CORE/002 CORE Data Quality Metrics</alias>
    <purpose>Assesses the quality of the data in the raw specification
created in business process, 02 Core ETL for Raw 6.1, on the basis of a set of
quality metrics</purpose>

```



```

        <standardmodelused>
            <uppermodel>
                <name>Generic Longitudinal Business Process Model</name>
                <step>5.3 Explore, validate and clean data</step>
            </uppermodel>
            <lowermodel>
                <name>African Demographic and Epidemiological Surveillance
Business Process Model (ADESBPM)</name>
                <step>Validate sex, dob, events order and events
dates</step>
            </lowermodel>
            <curation>
                <name>Open Archival Information System Reference
Model</name>
                <step>Archival Information Package</step>
            </curation>
        </standardmodelused>
        <algorithmoverview>
            <step id="3.1" name="Compile a list of Quality
Metrics">Compile a list of quality metrics relevant to the data
specification</step>
            <step id="3.2" name="Create events consistency matrix">Create
events consistency matrix showing the logical ordering of event
sequences</step>
            <step id="3.3" name="Compile residency starting
events">Identify in the data, events that start a residency episode (birth,
external-immigration, enumeration,
becoming eligible for a study, found after being lost to
follow-up)</step>
            <step id="3.4" name="Compile residency ending events">Identify
in the data, events that end a residency episode (external-outmigration,
death, became ineligible for study,
lost to follow-up, internal-outmigration, present in the
study (right censored)) </step>
            <step id="3.5" name="Compile legal and illegal start
events">Review the identified start events and distinguish between legal and
illegal ones</step>
            <step id="3.6" name="Compile legal and illegal end
events">Review the identified end events and distinguish between legal and
illegal ones</step>
            <step id="3.7" name="Compile legal and illegal
transitions">Review all transitions between two events and distinguish between
legal and illegal ones</step>
            <step id="3.8" name="Compile illegal, missing or unknown
sex">Compile illegal, missing or unknown sex</step>
            <step id="3.9" name="Compile illegal, missing or runknown Date
of Birth (DOB)">Compile illegal, missing or unknown DOB</step>
            <step id="3.10" name="Compile quality metrics">Calculate
numbers of legal and illegal start events, end events, event transitions, sex
values, out of range DOBs
and missing sex and DOBs</step>
        </algorithmoverview>
        <preconditions>
            <precondition>
                <recordname>Raw_61_Event_Format</recordname>
                <location>6.1/CORE/001 CORE Produce Raw 61
Dataset.kjb</location>
                <type>SQL</type>
                <aggregate>>false</aggregate>
            </precondition>
        </preconditions>
        <postconditions>
            <postcondition>
                <recordname>StartingEvents</recordname>
                <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>
                <type>SQL</type>
                <aggregate>>false</aggregate>
            </postcondition>
        </postconditions>
    </algorithm>

```



```

        <datadescription>
            <sibling>StartingEvents</sibling>
            <uri>6.1/CORE/02A CORE Illegal Start Events.ktr</uri>
            <description>Excel Output</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>StartingEvents</recordname>
        <location>6.1/CORE/02A CORE Illegal Start
Events.ktr</location>
        <type>ExcelOutput</type>
        <aggregate>false</aggregate>
        <datadescription>
            <sibling>StartingEvents</sibling>
            <uri>6.1/CORE/002 CORE Data Quality Metrics.kjb</uri>
            <description>SQL</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>StartQualityMetrics</recordname>
        <location>6.1/CORE/02A CORE Illegal Start
Events.ktr</location>
        <type>TypeExitExcelWriterStep</type>
        <aggregate>false</aggregate>
        <datadescription>
            <sibling>QualityMetrics</sibling>
            <uri>6.1/CORE/002 CORE Data Quality Metrics.kjb</uri>
            <description>SQL, Start events quality metrics added
to the QualityMetrics table in Staging database</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>QualityMetrics</recordname>
        <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>
        <type>SQL</type>
        <aggregate>false</aggregate>
        <datadescription>
            <sibling>StartQualityMetrics</sibling>
            <uri>6.1/CORE/02A CORE Illegal Start Events.ktr</uri>
            <description>Microsoft Excel Writer</description>
        </datadescription>
        <datadescription>
            <sibling>TransitionQualityMetrics</sibling>
            <uri>6.1/CORE/02B CORE Illegal Transitions.ktr</uri>
            <description>Microsoft Excel Writer</description>
        </datadescription>
        <datadescription>
            <sibling>SexValueQualityMetrics</sibling>
            <uri>6.1/CORE/02D CORE Unknown Sex.ktr</uri>
            <description>Microsoft Excel Writer</description>
        </datadescription>
        <datadescription>
            <sibling>EndingQualityMetrics</sibling>
            <uri>6.1/CORE/02C CORE Illegal End Events.ktr</uri>
            <description>Microsoft Excel Writer</description>
        </datadescription>
        <datadescription>
            <sibling>DoBQualityMetrics</sibling>
            <uri>6.1/CORE/02E CORE Illegal or Missing
DoB.ktr</uri>
            <description>DoBQualityMetrics</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>IllegalTransitions</recordname>
        <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>

```



```

        <type>SQL</type>
        <aggregate>>false</aggregate>
        <datadescription>
            <sibling>IllegalTransitions</sibling>
            <uri>6.1/CORE/02B CORE Illegal Transitions.ktr</uri>
            <description>Excel Output</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>IllegalTransitions</recordname>
        <location>6.1/CORE/02B CORE Illegal
Transitions.ktr</location>
        <type>ExcelOutput</type>
        <aggregate>>false</aggregate>
        <datadescription>
            <sibling>IllegalTransitions</sibling>
            <uri>6.1/CORE/002 CORE Data Quality Metrics.kjb</uri>
            <description>SQL</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>TransitionsCrosstab</recordname>
        <location>6.1/CORE/02B CORE Illegal
Transitions.ktr</location>
        <type>ExcelOutput</type>
        <aggregate>>false</aggregate>
        <datadescription>
            <sibling>EventCrossTab</sibling>
            <uri>6.1/CORE/002 CORE Data Quality Metrics.kjb</uri>
            <description>SQL</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>EventCrossTab</recordname>
        <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
        <datadescription>
            <sibling>TransitionsCrosstab</sibling>
            <uri>6.1/CORE/02B CORE Illegal Transitions.ktr</uri>
            <description>Excel Output 2</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>TransitionQualityMetrics</recordname>
        <location>6.1/CORE/02B CORE Illegal
Transitions.ktr</location>
        <type>TypeExitExcelWriterStep</type>
        <aggregate>>false</aggregate>
        <datadescription>
            <sibling>QualityMetrics</sibling>
            <uri>6.1/CORE/002 CORE Data Quality Metrics.kjb</uri>
            <description>SQL, Transition events quality metrics
added to the QualityMetrics table in Staging database</description>
        </datadescription>
    </postcondition>
    <postcondition>
        <recordname>EndingEvents</recordname>
        <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
        <datadescription>
            <sibling>EndingEvents</sibling>
            <uri>6.1/CORE/02C CORE Illegal End Events.ktr</uri>
            <description>Excel Output</description>
        </datadescription>

```



```

        </postcondition>
    </postcondition>
    <recordname>EndingEvents</recordname>
    <location>6.1/CORE/02C CORE Illegal End
Events.ktr</location>
    <type>ExcelOutput</type>
    <aggregate>false</aggregate>
    <datadescription>
        <sibling>EndingEvents</sibling>
        <uri>6.1/CORE/002 CORE Data Quality Metrics.kjb</uri>
        <description>SQL</description>
    </datadescription>
</postcondition>
<postcondition>
    <recordname>EndingQualityMetrics</recordname>
    <location>6.1/CORE/02C CORE Illegal End
Events.ktr</location>
    <type>TypeExitExcelWriterStep</type>
    <aggregate>false</aggregate>
    <datadescription>
        <sibling>QualityMetrics</sibling>
        <uri>6.1/CORE/002 CORE Data Quality Metrics.kjb</uri>
        <description>SQL, End events quality metrics added to
the QualityMetrics table in Staging database</description>
    </datadescription>
</postcondition>
<postcondition>
    <recordname>SexValues</recordname>
    <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>
    <type>SQL</type>
    <aggregate>false</aggregate>
    <datadescription>
        <sibling>SexValues</sibling>
        <uri>6.1/CORE/02D CORE Unknown Sex.ktr</uri>
        <description>Excel Output</description>
    </datadescription>
</postcondition>
<postcondition>
    <recordname>SexValues</recordname>
    <location>6.1/CORE/02D CORE Unknown Sex.ktr</location>
    <type>ExcelOutput</type>
    <aggregate>false</aggregate>
    <datadescription>
        <sibling>SexValues</sibling>
        <uri>6.1/CORE/002 CORE Data Quality Metrics.kjb</uri>
        <description>SQL</description>
    </datadescription>
</postcondition>
<postcondition>
    <recordname>SexValueQualityMetrics</recordname>
    <location>6.1/CORE/02D CORE Unknown Sex.ktr</location>
    <type>TypeExitExcelWriterStep</type>
    <aggregate>false</aggregate>
    <datadescription>
        <sibling>QualityMetrics</sibling>
        <uri>6.1/CORE/002 CORE Data Quality Metrics.kjb</uri>
        <description>SQL, Sex values quality metrics added to
the QualityMetrics table in Staging database</description>
    </datadescription>
</postcondition>
<postcondition>
    <recordname>DoBValues</recordname>
    <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>
    <type>SQL</type>
    <aggregate>false</aggregate>
    <datadescription>

```



```

        <sibling>DoBValues</sibling>
        <uri>6.1/CORE/02E CORE Illegal or Missing
DoB.ktr</uri>
        <description>Excel Output</description>
    </datadescription>
</postcondition>
<postcondition>
    <recordname>DoBValues</recordname>
    <location>6.1/CORE/02E CORE Illegal or Missing
DoB.ktr</location>
    <type>ExcelOutput</type>
    <aggregate>false</aggregate>
    <datadescription>
        <sibling>DoBValues</sibling>
        <uri>6.1/CORE/002 CORE Data Quality Metrics.kjb</uri>
        <description>SQL</description>
    </datadescription>
</postcondition>
<postcondition>
    <recordname>DoBQualityMetrics</recordname>
    <location>6.1/CORE/02E CORE Illegal or Missing
DoB.ktr</location>
    <type>TypeExitExcelWriterStep</type>
    <aggregate>false</aggregate>
    <datadescription>
        <sibling>QualityMetrics</sibling>
        <uri>6.1/CORE/002 CORE Data Quality Metrics.kjb</uri>
        <description>SQL, DoB quality metrics added to the
QualityMetrics table in Staging database</description>
    </datadescription>
</postcondition>
</postconditions>
</bp>
<bp id="4">
    <name>04 Core ETL to Clean 6.1 Dataset</name>
    <alias>6.1/CORE/003 CORE Data Cleaning</alias>
    <purpose>Applies cleaning procedures to correct some
inconsistencies identified in the quality assessment business process (03 Core
ETL Raw 6.1 Dataset Quality Metrics)
    This business process does not clean all the errors
identified, those requiring the attention of the ALPHA member centre are
compiled in preparation for sending
    to the member centre</purpose>
    <standardmodelused>
        <uppermodel>
            <name>Generic Longitudinal Business Process Model</name>
            <step>5.3 Explore, validate and clean data</step>
        </uppermodel>
        <lowermodel>
            <name>African Demographic and Epidemiological Surveillance
Business Process Model (ADESBPM)</name>
            <step>Clean Event Dates and Event ordering</step>
        </lowermodel>
        <curation>
            <name>Open Archival Information System Reference
Model</name>
            <step>Archival Information Package</step>
        </curation>
    </standardmodelused>
    <algorithmoverview>
        <step id="4.1" name="Check if first event is a legal
first">Check if the first event to be ever recorded for each individual is
enumeration, birth or external-immigration</step>
        <step id="4.2" name="Change internal-immigration (ENT) to
external-immigration (IMG) for first events">If first event is an internal-
immigration change it to an
            external immigration</step>
        <step id="4.3" name="Classify all other first events as

```



```

illegal">Classify all first events other than enumeration, birth or external-
immigration as illegal first events</step>
    <step id="4.4" name="Check if its a legal first event">Check
if the marked as first event is a birth, an enumeration or an immigration from
outside DSA</step>
    <step id="4.5" name="Drop individuals with illegal start
events"></step>
    <step id="4.6" name="Check if last event is a legal
last">Check if last events are external-outmigration, death, present in study
site</step>
    <step id="4.7" name="Change internal-outmigration (EXT) to
external-outmigration (OMG) for last event">If last event is an internal-
outmigration change it to
        an external outmigration</step>
    <step id="4.8" name="Classify all other last events as
illegal">Classify all last events other than external-outmigration, death,
present in study site as illegal last events</step>
    <step id="4.9" name="Drop individuals with illegal end
events"></step>
    <step id="4.10" name="Identify consecutive events and their
dates">Identify current and next event and their dates</step>
    <step id="4.11" name="Identify the event following each birth
event">Check if a birth event is followed by a birth, an enumeration,
external-immigration or internal-immigration</step>
    <step id="4.12" name="Identify the event following each death
event">Check if a death event is followed by an event other than a NULL</step>
    <step id="4.13" name="Compile event pairs violating
consistency matrix transitions">Review all other transitions in the data and
record violations of consistency matrix</step>
    <step id="4.14" name="Drop individuals with illegal
transitions"></step>
    <step id="4.15" name="Drop individuals with unknown sex or
DOB"></step>
</algorithmoverview>
<preconditions>
    <precondition>
        <recordname>Raw_61_Event_Format</recordname>
        <location>6.1/CORE/001 CORE Produce Raw 61
Dataset.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </precondition>
</preconditions>
<postconditions>
    <postcondition>
        <recordname>IllegalStartEvents</recordname>
        <location>6.1/CORE/003 CORE Data Cleaning.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>tmpMicroData</recordname>
        <location>6.1/CORE/003 CORE Data Cleaning.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>DeletedIndividualEvents</recordname>
        <location>6.1/CORE/003 CORE Data Cleaning.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>MicroDataStartCleaned</recordname>
        <location>6.1/CORE/003 CORE Data Cleaning.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>

```



```

        <postcondition>
            <recordname>IllegalEndEvents</recordname>
            <location>6.1/CORE/003 CORE Data Cleaning.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>MicroDataEndCleaned</recordname>
            <location>6.1/CORE/003 CORE Data Cleaning.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>MicroDataTransitionsCleaned</recordname>
            <location>6.1/CORE/003 CORE Data Cleaning.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>Cleaned_61_Event_Format</recordname>
            <location>6.1/CORE/003 CORE Data Cleaning.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>MicroDataCleanedSex</recordname>
            <location>6.1/CORE/003 CORE Data Cleaning.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </postcondition>
    </postconditions>
</bp>
<bp id="5">
    <name>05 Core ETL Clean 6.1 Dataset Quality Metrics</name>
    <alias>6.1/CORE/002 CORE Data Quality Metrics</alias>
    <purpose>Reruns data quality metrics first executed in Step 3
after events are cleaned in Step 4</purpose>
    <standardmodelused>
        <uppermodel>
            <name>Generic Longitudinal Business Process Model</name>
            <step>5.3 Explore, validate and clean data</step>
        </uppermodel>
        <lowermodel>
            <name>African Demographic and Epidemiological Surveillance
Business Process Model (ADESBPM)</name>
            <step>Validate sex, dob, events order and events
dates</step>
        </lowermodel>
        <curation>
            <name>Open Archival Information System Reference
Model</name>
            <step>Archival Information Package</step>
        </curation>
    </standardmodelused>
    <algorithmoverview>
        <step id="5.1" name="Compile a list of Quality
Metrics">Compile a list of quality metrics relevant to the data
specification</step>
        <step id="5.2" name="Create events consistency matrix">Create
events consistency matrix showing the logical ordering of event
sequences</step>
        <step id="5.3" name="Compile residency starting
events">Identify in the data events that start a residency episode (birth,
external-immigration, enumeration,
            becoming eligible for a study, found after being lost to
follow-up)</step>
        <step id="5.4" name="Compile residency ending events">Identify
in the data events that end a residency episode (external outmigration, death,

```



```

        became ineligible for study, lost to follow-up, internal-
outmigration, present in the study (right censored)) </step>
        <step id="5.5" name="Compile legal and illegal start
events">Review the identified start events and distinguish between legal and
illegal ones</step>
        <step id="5.6" name="Compile legal and illegal end
events">Review the identified end events and distinguish between legal and
illegal ones</step>
        <step id="5.7" name="Compile legal and illegal
transitions">Review all transitions between two events and distinguish between
legal and illegal ones</step>
        <step id="5.8" name="Compile illegal, missing or unknown
sex">Compile illegal, missing or unknown sex</step>
        <step id="5.9" name="Compile illegal, missing or runknown
dob">Compile illegal, missing or unknown dob</step>
        <step id="5.10" name="Compile quality metrics">Calculate
numbers of legal and illegal start events, end events, event transitions, sex
values, out of range DOBs
        and missing sex and DOBs</step>
</algorithmoverview>
<preconditions>
    <precondition>
        <recordname>MicroDataCleanedSex</recordname>
        <location>6.1/CORE/003 CORE Data Cleaning.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </precondition>
</preconditions>
<postconditions>
    <postcondition>
        <recordname>StartingEvents</recordname>
        <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>StartQualityMetrics</recordname>
        <location>6.1/CORE/02A CORE Illegal Start
Events.ktr</location>
        <type>TypeExitExcelWriterStep</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>IllegalTransitions</recordname>
        <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>TransitionsCrosstab</recordname>
        <location>6.1/CORE/02B CORE Illegal
Transitions.ktr</location>
        <type>ExcelOutput</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>TransitionQualityMetrics</recordname>
        <location>6.1/CORE/02B CORE Illegal
Transitions.ktr</location>
        <type>TypeExitExcelWriterStep</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>EndingEvents</recordname>
        <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>

```



```

        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>EndingQualityMetrics</recordname>
        <location>6.1/CORE/02C CORE Illegal End
Events.ktr</location>
        <type>TypeExitExcelWriterStep</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>SexValues</recordname>
        <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>SexValueQualityMetrics</recordname>
        <location>6.1/CORE/02D CORE Unknown Sex.ktr</location>
        <type>TypeExitExcelWriterStep</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>DoBValues</recordname>
        <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>DoBQualityMetrics</recordname>
        <location>6.1/CORE/02E CORE Illegal or Missing
DoB.ktr</location>
        <type>TypeExitExcelWriterStep</type>
        <aggregate>>false</aggregate>
    </postcondition>
</postconditions>
</bp>
<bp id="6">
    <name>06 CORE ETL to Anonymise Dataset</name>
    <alias>6.1/CORE/004 CORE Data Anonymisation</alias>
    <purpose>Randomises the individual and household identifiers to
anonymise the data</purpose>
    <standardmodelused>
        <uppermodel>
            <name>Generic Longitudinal Business Process Model</name>
            <step>5.8 Anonymise data</step>
        </uppermodel>
        <lowermodel>
            <name>African Demographic and Epidemiological Surveillance
Business Process Model (ADESBPM)</name>
            <step>Anonymise individuals' IDs, Physical Locations IDs
and Mothers' IDs</step>
        </lowermodel>
        <curation>
            <name>Open Archival Information System Reference
Model</name>
            <step>Archival Information Package</step>
        </curation>
    </standardmodelused>
    <algorithmoverview>
        <step id="6.1" name="Bring together original unique-
identifiers and anonymised IDs">Bring together original and anonymised IDs in
the cleaned spec 6.1</step>
        <step id="6.2" name="Remove original IDs from the cleaned spec
6.1">Create cleaned spec 6.1 with only anonymised IDs</step>
        <step id="6.3" name="Preserve ID Mappings">Preserve an

```



```

internal mapping of original IDs to the anonymised IDs</step>
  </algorithmoverview>
  <preconditions>
    <precondition>
      <recordname>Cleaned_61_Event_Format</recordname>
      <location>6.1/CORE/003 CORE Data Cleaning.kjb</location>
      <type>SQL</type>
      <aggregate>>false</aggregate>
    </precondition>
    <precondition>
      <recordname>Indv_ID_Anonymise_Map</recordname>
      <location>6.1/CORE/001 CORE Produce Raw 61
Dataset.kjb</location>
      <type>SQL</type>
      <aggregate>>false</aggregate>
    </precondition>
    <precondition>
      <recordname>HH_ID_Anonymise_Map</recordname>
      <location>6.1/CORE/001 CORE Produce Raw 61
Dataset.kjb</location>
      <type>SQL</type>
      <aggregate>>false</aggregate>
    </precondition>
  </preconditions>
  <postconditions>
    <postcondition>
      <recordname>Anonymised_61_Event_Format</recordname>
      <location>6.1/CORE/004 CORE Data
Anonymisation.kjb</location>
      <type>SQL</type>
      <aggregate>>false</aggregate>
    </postcondition>
  </postconditions>
</bp>
<bp id="7">
  <name>07 CORE Consolidate Start and End Events</name>
  <alias>6.1/CORE/005 CORE Consolidate Start and End Events</alias>
  <purpose>Quantifies proportion of records that are duplicates in
terms of unique-identifier, event and event date, cleans the duplicates and
drops individuals
  with single events</purpose>
  <standardmodelused>
    <uppermodel>
      <name>Generic Longitudinal Business Process Model</name>
      <step>5.3 Explore, validate and clean data</step>
    </uppermodel>
    <lowermodel>
      <name>African Demographic and Epidemiological Surveillance
Business Process Model (ADESBPM)</name>
      <step></step>
    </lowermodel>
    <curation>
      <name>Open Archival Information System Reference
Model</name>
      <step>Archival Information Package</step>
    </curation>
  </standardmodelused>
  <algorithmoverview>
    <step id="7.1" name="Dates of first events">Identify the date
for each individual's first event to be ever recorded</step>
    <step id="7.2" name="Dates of last events">Identify the date
for each individual's last event to be ever recorded</step>
    <step id="7.3" name="Identify and compile duplicates">Identify
duplicates in terms of ID, event and event date</step>
    <step id="7.4" name="Remove duplicates">Remove duplicate
record in terms of ID, event and event date identified in 7.3</step>
    <step id="7.5" name="Identify individuals with single
events">Identify individuals with total events amounting to 1 or less</step>

```



```

        <step id="7.6" name="Drop individuals with single events">Drop
individuals with total events amounting to 1 or less</step>
        <step id="7.7" name="Adjust events numbering">Adjust events
numbering to account for dropped events</step>
        <step id="7.8" name="Quantify proportion of records that are
duplicates">Quantify percentage of events that are duplicates as a quality
metric</step>
    </algorithmoverview>
    <preconditions>
        <precondition>
            <recordname>Anonymised_61_Event_Format</recordname>
            <location>6.1/CORE/004 CORE Data
Anonymisation.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </precondition>
    </preconditions>
    <postconditions>
        <postcondition>
            <recordname>StartEvents</recordname>
            <location>6.1/CORE/05A CORE Determine End Events and
Dates.ktr</location>
            <type>TypeExitExcelWriterStep</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>EndEvents</recordname>
            <location>6.1/CORE/05A CORE Determine End Events and
Dates.ktr</location>
            <type>TypeExitExcelWriterStep</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>Summary</recordname>
            <location>6.1/CORE/05A CORE Determine End Events and
Dates.ktr</location>
            <type>TextFileOutput</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>S01_Duplicates</recordname>
            <location>6.1/CORE/005 CORE Consolidate Start and End
Events.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>S01_Temp</recordname>
            <location>6.1/CORE/005 CORE Consolidate Start and End
Events.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>S01_Dropped</recordname>
            <location>6.1/CORE/005 CORE Consolidate Start and End
Events.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>S01_Temp2</recordname>
            <location>6.1/CORE/005 CORE Consolidate Start and End
Events.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>

```



```

        <recordname>S01</recordname>
        <location>6.1/CORE/005 CORE Consolidate Start and End
Events.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>QualityMetrics</recordname>
        <location>6.1/CORE/05D CORE Generate duplicate events
quality metric.ktr</location>
        <type>TableOutput</type>
        <aggregate>>false</aggregate>
    </postcondition>
    </postconditions>
</bp>
<bp id="8">
    <name>08 CORE Verify Temporal Integrity</name>
    <alias>6.1/CORE/006 CORE Verify Temporal Integrity</alias>
    <purpose>Assesses the ordering, in time, of dates for consecutive/
successive events, compiles those with illogical timing, quantifies their
proportion and
        drops individuals with wrongly timed successive events
</purpose>
    <standardmodelused>
        <uppermodel>
            <name>Generic Longitudinal Business Process Model</name>
            <step>5.3 Explore, validate and clean data</step>
        </uppermodel>
        <lowermodel>
            <name>African Demographic and Epidemiological Surveillance
Business Process Model (ADESBPM)</name>
            <step>Validate and clean event histories</step>
        </lowermodel>
        <curation>
            <name>Open Archival Information System Reference
Model</name>
            <step>Archival Information Package</step>
        </curation>
    </standardmodelused>
    <algorithmoverview>
        <step id="8.1" name="Identify an individual's dates of
consecutive events">For each individual in the data, determine the dates of
consecutive events event date
            and next event date</step>
        <step id="8.2" name="Use a future date if next event date is
NULL">Replace NULL next event dates with a date in the future</step>
        <step id="8.3" name="Assess temporal integrity of consecutive
events dates">Check if event date is less than next event date and record
violations</step>
        <step id="8.4" name="Quantify proportion event dates violating
temporal integrity">Quantify proportion event dates violating temporal
integrity as a quality metric</step>
        <step id="8.5" name="Drop individuals with temporal integrity
violations">Drop individuals with temporal integrity violations</step>
    </algorithmoverview>
    <preconditions>
        <precondition>
            <recordname>S01</recordname>
            <location>6.1/CORE/005 CORE Consolidate Start and End
Events.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </precondition>
    </preconditions>
    <postconditions>
        <postcondition>
            <recordname>S02_Violations</recordname>
            <location>6.1/CORE/006 CORE Verify Temporal

```



```

Integrity.kjb</location>
    <type>SQL</type>
    <aggregate>>false</aggregate>
</postcondition>
<postcondition>
    <recordname>QualityMetrics</recordname>
    <location>6.1/CORE/06A CORE Verify Temporal
Integrity.ktr</location>
    <type>TableOutput</type>
    <aggregate>>false</aggregate>
</postcondition>
<postcondition>
    <recordname>S02_DeletedTemporalViolations</recordname>
    <location>6.1/CORE/006 CORE Verify Temporal
Integrity.kjb</location>
    <type>SQL</type>
    <aggregate>>false</aggregate>
</postcondition>
<postcondition>
    <recordname>S02</recordname>
    <location>6.1/CORE/006 CORE Verify Temporal
Integrity.kjb</location>
    <type>SQL</type>
    <aggregate>>false</aggregate>
</postcondition>
</postconditions>
</bp>
<bp id="9">
    <name>09 CORE Update Event Timing</name>
    <alias>6.1/CORE/007 CORE Update Event Timing</alias>
    <purpose>Assesses and corrects migration event sequences. A
movement out of the study area is defined as an external-outmigration (OMG) if
the time between the
        external-outmigration and the subsequent external-immigration
(IMG) is above a defined period of time (threshold) - e.g. six
months.</purpose>
    <standardmodelused>
        <uppermodel>
            <name>Generic Longitudinal Business Process Model</name>
            <step>5.3 Explore, validate and clean data</step>
        </uppermodel>
        <lowermodel>
            <name>African Demographic and Epidemiological Surveillance
Business Process Model (ADESBPM)</name>
            <step>Smooth event histories</step>
        </lowermodel>
        <curation>
            <name>Open Archival Information System Reference
Model</name>
            <step>Archival Information Package</step>
        </curation>
    </standardmodelused>
    <algorithmoverview>
        <step id="9.1" name="Identify an individual's consecutive
events and their dates">For each individual in the data, determine previous
and current event and their
            corresponding dates</step>
        <step id="9.2" name="Compile OMG-IMG pairs">From consecutive
event pairs identified in 9.1, compile pairs where an external-outmigration is
followed by an external-immigration</step>
        <step id="9.3" name="Check if date difference between OMG and
IMG is below threshold">Identify and compile OMG-IMG pairs with date
differences below a recommended threshold</step>
        <step id="9.4" name="Change IMG to ENT for OMG-IMG date
difference below threshold">Change external-immigration to internal-
immigration if the OMG-IMG date differences
            is below a recommended threshold</step>
        <step id="9.5" name="Quantify proportion OMG-IMG pairs with

```



```

date differences below threshold">Quantify proportion OMG-IMG pairs with date
differences below a recommended
    threshold as a quality metric</step>
    <step id="9.6" name="Change OMG to EXT for OMG-IMG pair date
difference below threshold">Change external-outmigration to internal-
outmigration if the OMG-IMG date differences
    is below a recommended threshold</step>
    <step id="9.7" name="Compile EXT-ENT pairs">From consecutive
event pairs identified in 9.1, compile pairs where an internal-outmigration is
followed by an internal-immigration</step>
    <step id="9.8" name="Check if date difference between EXT and
ENT is above threshold">Identify and compile EXT-ENT pairs with date
differences above a recommended threshold</step>
    <step id="9.9" name="Change ENT to IMG for EXT-ENT pair date
difference above threshold">Change internal-immigration to external-
immigration if the EXT-ENT pair date differences
    is above a recommended threshold</step>
    <step id="9.10" name="Quantify proportion EXT-ENT pairs with
date differences above threshold">Quantify proportion EXT-ENT pairs with date
differences above a recommended
    threshold as a quality metric</step>
    <step id="9.11" name="Change EXT to IMG for EXT-ENT pair date
difference above threshold">Change internal-outmigration to external-
outmigration if the EXT-ENT pair date
    difference is above a recommended threshold</step>
    <step id="9.12" name="Assign times to event dates">Add
recommended times to the event dates dependent on the event type to maintain
temporal integrity and logical
    event sequences</step>
</algorithmoverview>
<preconditions>
    <precondition>
        <recordname>S02</recordname>
        <location>6.1/CORE/006 CORE Verify Temporal
Integrity.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </precondition>
</preconditions>
<postconditions>
    <postcondition>
        <recordname>S03_Temp</recordname>
        <location>6.1/CORE/007 CORE Update Event
Timing.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>S03_Recode_OMG</recordname>
        <location>6.1/CORE/007 CORE Update Event
Timing.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>QualityMetrics</recordname>
        <location>6.1/CORE/07A CORE Consolidate OMG-IMG
pairs.ktr</location>
        <type>TableOutput</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>S03_Temp2</recordname>
        <location>6.1/CORE/007 CORE Update Event
Timing.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </postcondition>

```



```

        <postcondition>
            <recordname>S03_Temp</recordname>
            <location>6.1/CORE/07C CORE Consolidate EXT-ENT
pairs.ktr</location>
            <type>TableOutput</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>S03_Recode_EXT</recordname>
            <location>6.1/CORE/007 CORE Update Event
Timing.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>QualityMetrics</recordname>
            <location>6.1/CORE/07C CORE Consolidate EXT-ENT
pairs.ktr</location>
            <type>TableOutput</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>S03_Temp2</recordname>
            <location>6.1/CORE/07D CORE Recode EXT events to
OMG.ktr</location>
            <type>TableOutput</type>
            <aggregate>>false</aggregate>
        </postcondition>
        <postcondition>
            <recordname>S03</recordname>
            <location>6.1/CORE/007 CORE Update Event
Timing.kjb</location>
            <type>SQL</type>
            <aggregate>>false</aggregate>
        </postcondition>
    </postconditions>
</bp>
<bp id="10">
    <name>10 CORE Produce Final Core MicroData Files</name>
    <alias>6.1/CORE/008 CORE Produce Final Core MicroData
Files</alias>
    <purpose>Produces the final dataset</purpose>
    <standardmodelused>
        <uppermodel>
            <name>Generic Longitudinal Business Process Model</name>
            <step>5.9 Finalize data outputs</step>
        </uppermodel>
        <lowermodel>
            <name>African Demographic and Epidemiological Surveillance
Business Process Model (ADESBPM)</name>
            <step>Create calendar of events</step>
        </lowermodel>
        <curation>
            <name>Open Archival Information System Reference
Model</name>
            <step>Archival Information Package</step>
        </curation>
    </standardmodelused>
    <algorithmoverview>
        <step id="10.1" name="Output events format data">Output data
in the form of a calendar of events</step>
        <step id="10.2" name="Identify residency episode start
events">Identify events that start a residency episode in the events format
data</step>
        <step id="10.3" name="Identify residency episode end
events">Identify events that end a residency episode in the events format
data</step>
        <step id="10.4" name="Bring together episode start and end

```



```

events for each episode">Bring together in one record corresponding events
that
        start and end each of the residency episodes in the
data</step>
        <step id="10.5" name="Compile unmatched start and end
events">Compile start and end events in one record but belonging to different
individuals</step>
        <step id="10.6" name="Compile illegal start events">Compile
illegal start events</step>
        <step id="10.7" name="Compile illegal end events">Compile
illegal end events</step>
        <step id="10.8" name="Output episodes format data">Output data
in the form of residency episodes</step>
        <step id="10.9" name="Generate MD5 checksum for events format
data">Generate and store an MD5 fingerprint for the events format data</step>
        <step id="10.10" name="Generate MD5 checksum for episodes
format data">Generate and store an MD5 fingerprint for the residency episodes
format data</step>
</algorithmoverview>
<preconditions>
<precondition>
        <recordname>S03</recordname>
        <location>6.1/CORE/007 CORE Update Event
Timing.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
</precondition>
<precondition>
        <recordname>Final_61_Episode_Format</recordname>
        <location>6.1/CORE/08C CORE Generate MD5 Fingerprint for
Final Microdata.ktr</location>
        <type>LoadFileInput</type>
        <aggregate>>false</aggregate>
</precondition>
<precondition>
        <recordname>Final_61_Event_Format</recordname>
        <location>6.1/CORE/08C CORE Generate MD5 Fingerprint for
Final Microdata.ktr</location>
        <type>LoadFileInput</type>
        <aggregate>>false</aggregate>
</precondition>
</preconditions>
<postconditions>
<postcondition>
        <recordname>Final_61_Event_Format</recordname>
        <location>6.1/CORE/008 CORE Produce Final Core MicroData
Files.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
</postcondition>
<postcondition>
        <recordname>Final_61_Episode_Format</recordname>
        <location>6.1/CORE/008 CORE Produce Final Core MicroData
Files.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
</postcondition>
<postcondition>
        <recordname>MD5-DataFingerPrint</recordname>
        <location>6.1/CORE/08C CORE Generate MD5 Fingerprint for
Final Microdata.ktr</location>
        <type>TextFileOutput</type>
        <aggregate>>false</aggregate>
</postcondition>
</postconditions>
</bp>
<bp id="11">
        <name>11 CORE Prepare Data Quality Feedback</name>

```



```

        <alias>6.1/CORE/009 CORE Prepare Data Quality Feedback</alias>
        <purpose>Compiles data quality assessment report to be shared with
the member centre</purpose>
        <standardmodelused>
            <uppermodel>
                <name>Generic Longitudinal Business Process Model</name>
                <step></step>
            </uppermodel>
            <lowermodel>
                <name>African Demographic and Epidemiological Surveillance
Business Process Model (ADESBPM)</name>
                <step></step>
            </lowermodel>
            <curation>
                <name>Open Archival Information System Reference
Model</name>
                <step>Dissemination Information Package</step>
            </curation>
        </standardmodelused>
        <algorithmoverview>
            <step id="11.1" name="Output duplicate records">Output
duplicate records in terms of unique-identifier, event and event date</step>
            <step id="11.2" name="Output core violations">Output core
violations (illegal start, end events, illegal transitions, missing sex and
missing DOBs</step>
            <step id="11.3" name="Output dropped events">Output dropped
events</step>
            <step id="11.4" name="Output temporal violations">Output
records with temporal violations</step>
            <step id="11.5" name="Output offending migration
events">Output offending migration events</step>
            <step id="11.6" name="Produce quality metrics report">Produce
a summary report of the quality metrics providing statistics to give the
magnitude of the errors</step>
        </algorithmoverview>
        <preconditions>
            <precondition>
                <recordname>S01_Duplicates</recordname>
                <location>6.1/CORE/005 CORE Consolidate Start and End
Events.kjb</location>
                <type>SQL</type>
                <aggregate>>false</aggregate>
            </precondition>
            <precondition>
                <recordname>S01_Dropped</recordname>
                <location>6.1/CORE/005 CORE Consolidate Start and End
Events.kjb</location>
                <type>SQL</type>
                <aggregate>>false</aggregate>
            </precondition>
            <precondition>
                <recordname>S02_Violations</recordname>
                <location>6.1/CORE/006 CORE Verify Temporal
Integrity.kjb</location>
                <type>SQL</type>
                <aggregate>>false</aggregate>
            </precondition>
            <precondition>
                <recordname>S02_DeletedTemporalViolations</recordname>
                <location>6.1/CORE/006 CORE Verify Temporal
Integrity.kjb</location>
                <type>SQL</type>
                <aggregate>>false</aggregate>
            </precondition>
            <precondition>
                <recordname>S03_Recode_EXT</recordname>
                <location>6.1/CORE/007 CORE Update Event
Timing.kjb</location>

```



```

        <type>SQL</type>
        <aggregate>>false</aggregate>
    </precondition>
    <precondition>
        <recordname>S03_Recode_OMG</recordname>
        <location>6.1/CORE/007 CORE Update Event
Timing.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </precondition>
    <precondition>
        <recordname>QualityMetrics</recordname>
        <location>6.1/CORE/002 CORE Data Quality
Metrics.kjb</location>
        <type>SQL</type>
        <aggregate>>false</aggregate>
    </precondition>
</preconditions>
<postconditions>
    <postcondition>
        <recordname>DuplicateEvents</recordname>
        <location>6.1/CORE/09A CORE Output
Duplicates.ktr</location>
        <type>StataOutput</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>DroppedEvents</recordname>
        <location>6.1/CORE/09B CORE Dropped Events.ktr</location>
        <type>StataOutput</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>Violationss</recordname>
        <location>6.1/CORE/09C CORE Violations.ktr</location>
        <type>StataOutput</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>DeletedTemporalViolations</recordname>
        <location>6.1/CORE/09D CORE Deleted
TemporalViolations.ktr</location>
        <type>StataOutput</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>Recode_EXT</recordname>
        <location>6.1/CORE/09E CORE Recoded EXT
events.ktr</location>
        <type>StataOutput</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>S03_Recode_OMG</recordname>
        <location>6.1/CORE/09F CORE Recoded OMG
events.ktr</location>
        <type>StataOutput</type>
        <aggregate>>false</aggregate>
    </postcondition>
    <postcondition>
        <recordname>QualityMetrics</recordname>
        <location>6.1/CORE/09G CORE Output Quality
Metrics.ktr</location>
        <type>TypeExitExcelWriterStep</type>
        <aggregate>>false</aggregate>
    </postcondition>
</postconditions>
</bp>

```



```

    <attribution>
      <entity>
        <name>Open Archival Information System Reference Model</name>
        <abbreviation>OAIS</abbreviation>
        <description></description>
      </entity>
      <entity>
        <name>African Demographic and Epidemiological Surveillance
Business Process Model (ADESBPM)</name>
        <abbreviation>HRS2</abbreviation>
        <description>The Household Registration System 2 maintains a
consistent record of significant demographic events that occur to a population
in a fixed geographic region.
        HRS2 has been extended to support epidemiological events
too by the ALPHA (Analyzing Longitudinal Population-based HIV/AIDS data on
Africa) Network. </description>
      </entity>
      <entity>
        <name>Generic Longitudinal Business Process Model</name>
        <abbreviation>GLBPM</abbreviation>
        <description></description>
      </entity>
    </attribution>
  </datapipeline>
</infusion_6.1>

```


Centre in a Box data documentation (CiBDoS) software requirements elicitation study

BACKGROUND INFORMATION

1. WHAT IS ALPHA?

The ALPHA network is an innovative secondary data analysis program aimed at improving our understanding of the HIV epidemiology. ALPHA is coordinated by its secretariat in the Department of Population Health (DPH) under the Faculty of Epidemiology and Population Health at the London School of Hygiene and Tropical Medicine. It comprises of 10 autonomous research institutions sharing similar interests in HIV Epidemiology. Each institution has its own research agenda and data management system. All partners pre-date the network formation. They all have population/community-based longitudinal demographic and HIV surveillance data.

ALPHA leverages the benefits of data pooling - Better statistical power gained by bringing together data from a number of research institutions and a wider perspective not possible to achieve with one research institution.

2. ALPHA data and “modus operandi”

ALPHA assembles datasets on various topics related to demographic and HIV surveillance. These data are referred to as ALPHA data specifications or data specs and are described here⁵. The ALPHA data specs have a well-defined structure to which each partner of the network has to transform their data. ALPHA is organised around data analysis and HIV research capacity strengthening workshops. At the workshops, partners bring their data and are involved in data analysis training addressing research questions of interest for the particular workshop.

⁵ <http://alpha.lshtm.ac.uk/metadata/>

3. Data harmonisation in ALPHA

ALPHA is working on a project to produce a sharable set of harmonised data that combines both population-based and clinic data from the partner studies with funding from the Wellcome Trust.

Whilst community-based cohorts and demographic surveillance systems provide a rich source of data, use of the data is often limited because successful analysis requires detailed knowledge of the study's contemporary and historical procedures and of data management practices. To date the ALPHA Network has successfully extracted and harmonised 10 standard data tables from the partner studies. However, these data are still complex and require considerable prior knowledge to use effectively, which in practice means the data can only be used in collaboration with one of the ALPHA staff.

The main project combines three sets of activities:

- (1) Using industry standard data integration methods, and a bespoke data appliance Centre in a Box - CiB (Herbst et al. 2015) to develop a robust process for deriving the ALPHA datasets.
- (2) Integration of the existing ALPHA clinical dataset with data contributed to the leDEA Network (which links HIV clinical cohorts).
- (3) High-quality documentation of both the data and the processes used to derive the data.

The proposed study relates to the third set of activities in the main ALPHA project outlined earlier. Work done so far includes development of the software agent for harvesting the process metadata within CiB and formatting it in line with international metadata standards. The utility of the harvested metadata lies in the availability of software tools for browsing, searching and constructing data lineages relating to the ALPHA datasets. In order to build such tools, software developers need domain experts' perspectives on the desired functionality of those tools to guide their work. This study seeks to gather, analyse and synthesise these domain experts' functional requirements.

4. Mock-ups

Included in the information pack is a set of mock-up diagrams showing features that the developers have proposed as a starting point for discussion. Please note that these mock-ups are not a reflection of what the software interface will look like, they only

show the features of the system. Please have a look at these before the interview as the interview questions will seek to elicit your views about the proposed features.

5. Terms used in mock-ups

The overall process implemented in Pentaho (Pentaho Corporation 2018) for creating an ALPHA data spec is called a **data pipeline** for that data spec. **Business processes** are Pentaho sub-jobs within the data pipeline for a data spec. Each business process has an **overview, purpose** and some **business steps**. Each Business step has a **description**, related demographic and epidemiological **concepts** and associated **files**. Each business step also has **input data stores** and **output data stores**. Each data store is linked to the business process that create the data store and the business process that uses the data store.

6. Why ALPHA network interviewees?

ALPHA researchers, as the producers of the harmonised data, will provide the viewpoint of users who are familiar with the specifics of the data harmonisation process.

7. Why CLOSER project interviewees?

CLOSER staff will provide the viewpoint of archivists and data scientists who are familiar with international metadata standards and with data harmonisation (they have successfully conducted an ongoing data harmonisation project involving eight UK birth cohorts).

Between these two groups of users, we feel that we will be able to identify the requirements of both internal and external users.

8. Requirements overview

A requirement is a statement that identifies a necessary attribute, capability, characteristic, or quality of a system in order for it to have value and utility to a stakeholder.

9. Types of Requirements

A requirement can be:

- A **Business Goal**: a state or target that the organisation intends to achieve or maintain with the system.

- An **Objective**: a quantitatively measurable and specific state or target that the organisation intends to achieve or maintain with the system.
- A **System Goal**: a state or target that you intend to achieve or maintain by using the system.
- A **Capability Constraint**: a restriction on how the system achieves your goal.
- A **Quality of Service Constraint**: a quality restriction on the behaviour of the system.
- A **Business Policy**: a directive from the organisation that defines what can be done and what must not be done, and may indicate or set limits on how it should be done.
- A **Business Rule**: a directive from the organisation that provides specific and discrete governance or guidance to implement Business Policies.

Examples	Templates
To view input datasets used in a data transformation To see the association between output datasets and a process step	To <a goal you want to achieve by using The system>.
To improve usability of ALPHA harmonised datasets.	To <a goal the organisation should achieve from the system in operation>.
ALPHA and external researchers should be able to access high level description of data transformations by using the CiB documentation system	<subject> should [not] be able to <action> (by using the system).
All business processes must have a human readable algorithm overview	By / Within / Per annum <a measurable criteria to know if the organisation's goal is achieved>.
The system must provide various access levels for different user groups as determined by ALPHA network scientists and data producers	<subject> must / should [not] <action> [If/while <condition>].

10. References

Herbst, K., Juvekar, S., Bhattacharjee, T., Bangha, M., Patharia, N., Tei, T., ... Sankoh, O. (2015). The INDEPTH Data Repository: An International Resource for Longitudinal Population and Health Data From Health and Demographic Surveillance Systems. *Journal of Empirical Research on Human Research Ethics*, 10(3), 324–333.

Pentaho Corporation. (2018, October 10). Pentaho Data Integration. Retrieved February 19, 2019, from https://help.pentaho.com/Documentation/8.2/Products/Data_Integration

Centre in a Box data documentation (CiBDoS) software requirements elicitation study

PARTICIPANT INFORMATION SHEET

My name is Chifundo Kanjala and I am a PhD student in the Department of Population Health, London School of Hygiene and Tropical Medicine. We are carrying out a study to elicit functional requirements for a software system for browsing, searching and constructing data lineages relating to the ALPHA datasets.

You are being invited to take part in a research study. Before you decide to take part, it is important for you to understand why the research is being done and what it will involve. I will read information to you about this study. Please ask me if there is anything that is not clear or if you would like more information.

WHAT ARE WE TRYING TO LEARN WITH THIS RESEARCH STUDY?

We would like to understand what the business requirements are for a data documentation software for ALPHA datasets from the perspectives of domain experts in ALPHA and the CLOSER project.

The full utility of structured data documentation is realised when tools are available to browse, search and explore those metadata. The functionalities of such a tool for ALPHA datasets documentation is currently not fully understood. It is important to understand these requirements from the perspective the potential users and experts in the area of research data harmonisation and dissemination. By interviewing domain experts, we can gather and analyse their views. We hope that the findings from this study will help improve our understanding of the desired functional requirements for the said tools.

WHY ARE WE ASKING YOU TO PARTICIPATE?

We are asking you to give us your perspectives on what the features of a data documentation software for ALPHA datasets should be. We are also seeking to hear your opinions on mock-ups showing some of the main features that the research team has come up with as we are beginning to work on the project.

WHAT HAPPENS IF I DON'T WANT TO PARTICIPATE IN THE STUDY?

You are free to refuse to participate in this study, or to withdraw your participation at any time. Refusal to participate or withdrawal will not affect you in any way.

WHAT WILL MY PARTICIPATION IN THIS STUDY INVOLVE?

If you choose to participate in this study, we will ask you for up one hour of your time for a recorded skype interview. Prior to the interview, we will also ask you to spare time to read through 4 pages of background materials and to annotate the material with notes and questions that you might have from the content. The annotated background materials will be requested for prior to the interview. Your notes will be used together with your responses during analysis, and your questions will be addressed during the interview.

ARE THERE ANY RISKS INVOLVED WITH PARTICIPATING IN THIS STUDY?

There are no direct risks from participating.

ARE THERE ANY BENEFITS INVOLVED WITH PARTICIPATING IN THIS STUDY?

It is hoped that the software resulting from this requirement elicitation exercise will be useful to producers and users of the ALPHA datasets.

WILL I BE ALLOWED TO WITHDRAW FROM THE STUDY IF I CHANGE MY MIND?

Taking part in this study is voluntary. Should you wish to withdraw from the study at any point or not to answer any of the questions you are free to do so. It will not affect you in any way.

WHO WILL SEE THE INFORMATION THAT IS COLLECTED?

Personal identifiers will be removed from the questionnaire before analysis, and all data will be stored in a way that only authorised people can access it. Your personal information will not be revealed in any published information.

HOW WILL THE INFORMATION I GIVE IN THE STUDY BE KEPT PRIVATE / WHO WILL SEE MY INFORMATION?

All your information will be kept confidential. Information will be stored in password protected computers. To protect your privacy, we will use a code number to identify you and all information about you. We will keep records securely locked/ password protected. Your name, or any other facts that might point to you, will not appear when we present this study or publish its results. Your data may be shared with other researchers only in securely anonymised form.

WHO TO CONTACT IF YOU WANT MORE INFORMATION, OR IF YOU HAVE A PROBLEM?

If you want more information before deciding to take part, or have questions at any time, please contact: Prof Jim Todd Email jim.todd@lshtm.ac.uk or Dr Jay Greenfield Email nightcleaner@gmail.com or Dr Emma Slaymaker Email: emma.slaymaker@lshtm.ac.uk

APPENDIX E Consent form

CONSENT FORM FOR PARTICIPANT



Title of Project: ALPHA Centre in a Box data provenance documentation subsystem requirements elicitation study

Name of PI/Researcher responsible for project: Chifundo Kanjala

Statement	Please initial or thumbprint* each box
I confirm that I have read and understood the information sheet dated.....(version.....) for the above named study. I have had the opportunity to consider the information, ask questions and have these answered satisfactorily.	
I understand that my consent is voluntary and that I am free to withdraw this consent at any time without giving any reason and without being affected in any way.	
I understand that relevant sections of my data collected during the study may be looked at by authorised individuals from [London School of Hygiene and Tropical Medicine and named consultants from the USA], where it is relevant to my taking part in this research. I give permission for these individuals to have access to these records.	
I understand that data about/from me may be shared via a public data repository or by sharing directly with other researchers, and that I will not be identifiable from this information	
I agree to me taking part in the above named study.	

Printed name of participant	Signature of participant (or thumbprint/mark if unable to sign)	Date
Printed name of person obtaining consent	Signature of person obtaining consent	Date

A copy of this informed consent document has been provided to the participant.

Centre Number:
Study Number:
Participant Identification Number:

[Informed Consent for Participant and Representative for Incapacitated adults_23.09.2016_v2]

APPENDIX F Questionnaire guide

Centre in a Box data documentation (CiBDoS) software requirements elicitation study

QUESTIONNAIRE

Please complete the following information about the interviewee.

Name	
Date	March, 2019
Position	
Organisation	
What stakeholder does the interviewee represent	

The following pages have are six mock-ups showing various proposed features of the CiB data documentation Subsystem. The interview will discuss the mock-ups. The purpose of the mock-ups is to show what metadata could be viewed through the system. The appearance of the mock-ups is not intended to be representative of the interface of the mature development. Feedback from the interviewee need to focus on the concepts not interface.

Question 1.

Figure 1:Mock-up 1 – the Data pipeline, its constituent business processes and the details of a business process

<p>ALPHA Data Pipeline Spec. 6.1 Business Processes:</p> <p>[02] Core ETL for Raw Input Overview Purpose Steps: [2.1] Generate Anonymized IDs [2.2] Map original IDs to Anonymized IDs [2.3] Store ID Mapping [2.4] Create 6.1 Data from Raw Data</p> <p>[03] 6.1 Data Quality Metrics Overview Purpose Steps: [3.1] Compile Quality Metrics [3.2] Compile Residency Starting Events [3.3] Compile Residency Ending Events [3.4] Compile Legal and Illegal Starting Events</p>	<p>02 Core ETL for Raw Input - Purpose</p> <p>Creates staging tables from member-centre-specific data. The staging tables are then transformed further to create the ALPHA specification 6.1.</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Write the interviewee's rating of the features from 0 (not important to them) to 5 (very important to them) in Table 1 against each feature.

Circle -1 for requirements that the user **actively does not want** in the system.

Table 21:

Requirements	Increasing importance						
1.1 Exposing of business processes	-1	0	1	2	3	4	5
1.1.1 Overview	-1	0	1	2	3	4	5
1.1.2 Purpose	-1	0	1	2	3	4	5
1.1.3 Business steps	-1	0	1	2	3	4	5
2.1 Exposure of business process component details	-1	0	1	2	3	4	5

The interviewee will be asked to explain their response choices for the presented features

Question 2.

Figure 2:

<p>ALPHA Data Pipeline Spec. 6.1 <i>Business Processes:</i></p> <p>[02] Core ETL for Raw Input Overview Purpose Steps: [2.1] Generate Anonymized IDs [2.2] Map original IDs to Anonymized IDs [2.3] Store ID Mapping [2.4] Create 6.1 Data from Raw Data</p> <p>[03] 6.1 Data Quality Metrics Overview Purpose Steps: [3.1] Compile Quality Metrics [3.2] Compile Residency Starting Events [3.3] Compile Residency Ending Events [3.4] Compile Legal and Illegal Starting Events</p>	<p>[3.2] Compile Residency Starting Events</p> <p>Description: Identify in the data, events that start a residency episode (birth, external-immigration, enumeration, becoming eligible for a study, found after being lost to follow-up).</p> <p>Concepts: This algorithm step references the following study concepts: residency birth migration</p> <p>Related Files: CompResSt.do (STATA Executable File) ResStart.sps (SPSS Executable File)</p>
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 3:

<p>ALPHA Data Pipeline Spec. 6.1 <i>Business Processes:</i></p> <p>[02] Core ETL for Raw Input Overview Purpose Steps: [2.1] Generate Anonymized IDs [2.2] Map original IDs to Anonymized IDs [2.3] Store ID Mapping [2.4] Create 6.1 Data from Raw Data</p> <p>[03] 6.1 Data Quality Metrics Overview Purpose Steps: [3.1] Compile Quality Metrics [3.2] Compile Residency Starting Events [3.3] Compile Residency Ending Events [3.4] Compile Legal and Illegal Starting Events</p>	<p>[3.2] Compile Residency Starting Events</p> <p>Description: Identify in the data, events that start a residency episode (birth, external-immigration, enumeration, becoming eligible for a study, found after being lost to follow-up).</p> <p>Concepts: This algorithm step references the following study concepts: residency birth migration</p> <div data-bbox="873 1394 1230 1642"> <p>The change of residence by a registered individual or social group (e.g., a household). There are two types of migration that occur among the registered population. These are internal and external migration.</p> </div>
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Write the interviewee's rating of the features from 0 (not important to them) to 5 (very important to them) in Table 2 against each feature.

Circle -1 for requirements that the user **actively does not want** in the system

Table 22

Requirements	Increasing importance						
1.1 Viewing list of algorithm steps	-1	0	1	2	3	4	5
1.2 Viewing algorithm step description	-1	0	1	2	3	4	5
1.3 Viewing demographic surveillance concepts associated with the algorithm step	-1	0	1	2	3	4	5
1.4 Viewing definitions of demographic and epidemiological surveillance concepts associated with the algorithm step	-1	0	1	2	3	4	5
1.5 Viewing setup scripts associated with the algorithm step	-1	0	1	2	3	4	5

The interviewee will be asked to explain their response choices for the presented features

Question 3.

Figure 4:



Write the interviewee’s rating of the features from 0 (not important to them) to 5 (very important to them) in Table 3 against each feature.

Circle -1 for requirements that the user actively does not want in the system

Table 3

Requirements	Increasing importance						
	-1	0	1	2	3	4	5
1.1 Viewing association of business process algorithms to input and output datasets							

The interviewee will be asked to explain their response choices for the presented features

Question 4.

Figure 5:

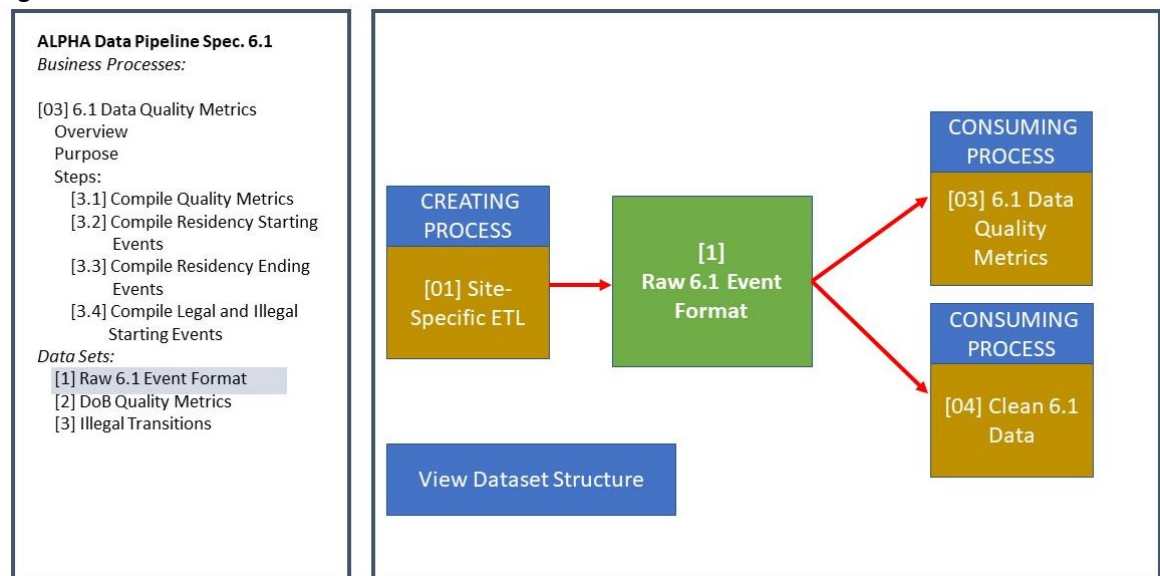
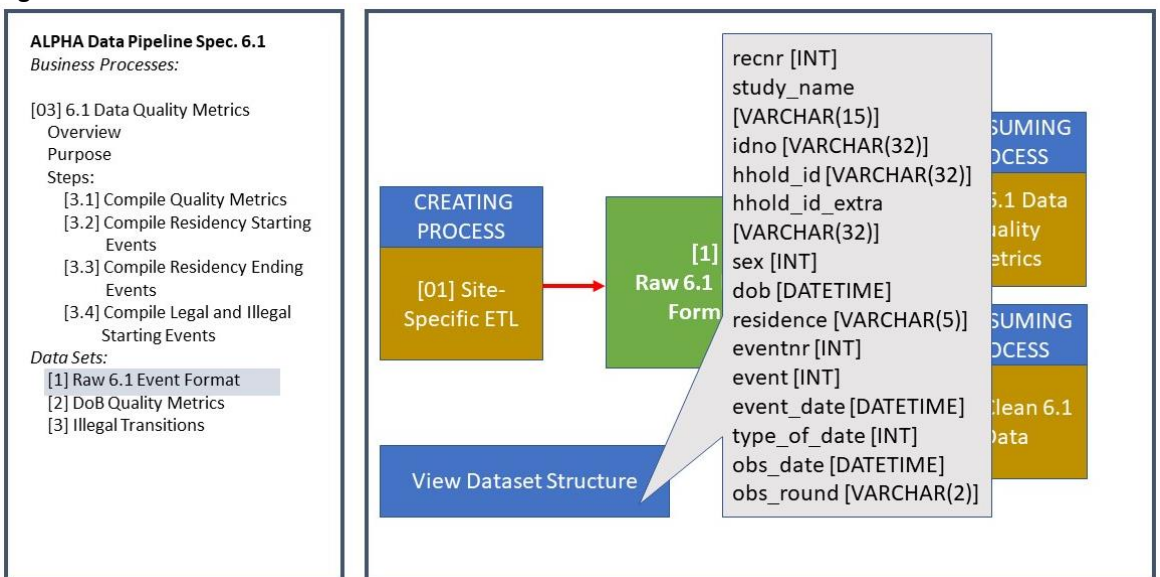


Figure 6:



Write the interviewee's rating of the features from 0 (not important to them) to 5 (very important to them) in Table 4 against each feature.
Circle -1 for requirements that the user **actively does not want** in the system

Table 23

Requirements		Increasing importance					
2.1 Viewing Dataset structure (constituent variables and their types) – Figure 6	-1	0	1	2	3	4	5
2.2 Viewing association of datasets and the corresponding creating and consuming processes	-1	0	1	2	3	4	5

The interviewee will be asked to explain their response choices for the presented features

Question 5.

Please **write the interviewee's requirements for CiB data documentation Subsystem** in the space below following the template provided.

Then, let the interviewee **rank the requirements** based on their importance to them in the right-hand column (1 being the most important).

Finally, **write requirements that the user actively does not want**, and **put an X** in the right-hand column.

Requirement	Rank/X

The interviewee will be asked to explain the rationale for the features and ranking

--

Question 6.

How much does the interviewee care about CiB data documentation Subsystem?

Please **circle** the appropriate answer.

Not at all	A little	So so	A lot
------------	----------	-------	-------

Please write any other comments the interviewee might have below.

--