ALPHA NETWORK

Analysing Longitudinal Population-based HIV/AIDS data on Africa

ALPHA network data lineage

Documentation of "after the fact" harmonised African longitudinal community-based demographic and HIV surveillance data

Chifundo Kanjala, Jay Greenfield, David Beckles and Basia Zaba

Improving health worldwide

www.lshtm.ac.uk

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Overview

- ALPHA network
  - Background
  - Data management
- Motivation
- Developing a business process model for ALPHA data management
- Going the last mile: Mapping Pentaho transformations to Structured Data Transform Language (SDTL)
- Where we are we now?
- Where to next?

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

Rakai Health Sciences Program
Improved Health Through Research

MRC | MRC / UVRI Uganda Research Unit on AIDS

APHRC African Population and Health Research Center

KEMRI.org
in search of better health

ALPHA NETWORK

Analysing Longitudinal Population-based HIV/AIDS data on Africa

NATIONAL INSTITUTE FOR MEDICAL RESEARCH

ihi IFAKARA HEALTH INSTITUTE
research | training | services

MEIRU
Malawi Epidemiology and Intervention Research Unit

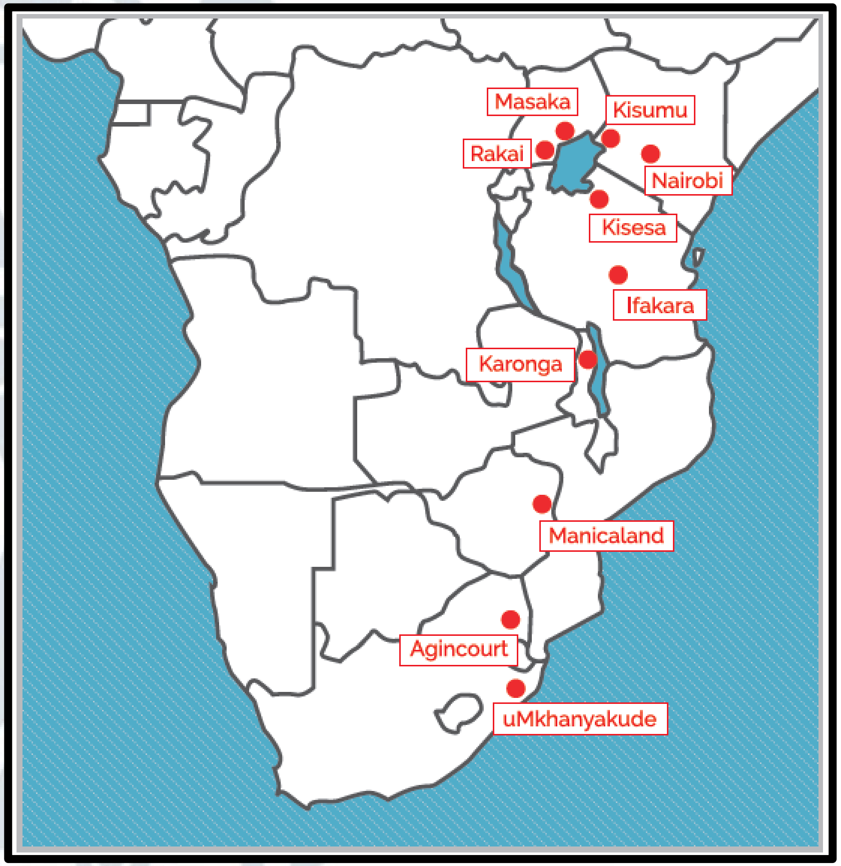
MANICALAND
HIV/STD PREVENTION PROJECT

AGINCOURT

saMRC
advancinglife

AHRI AFRICA HEALTH RESEARCH INSTITUTE

ALPHA partner studies and aims



Independent institutions

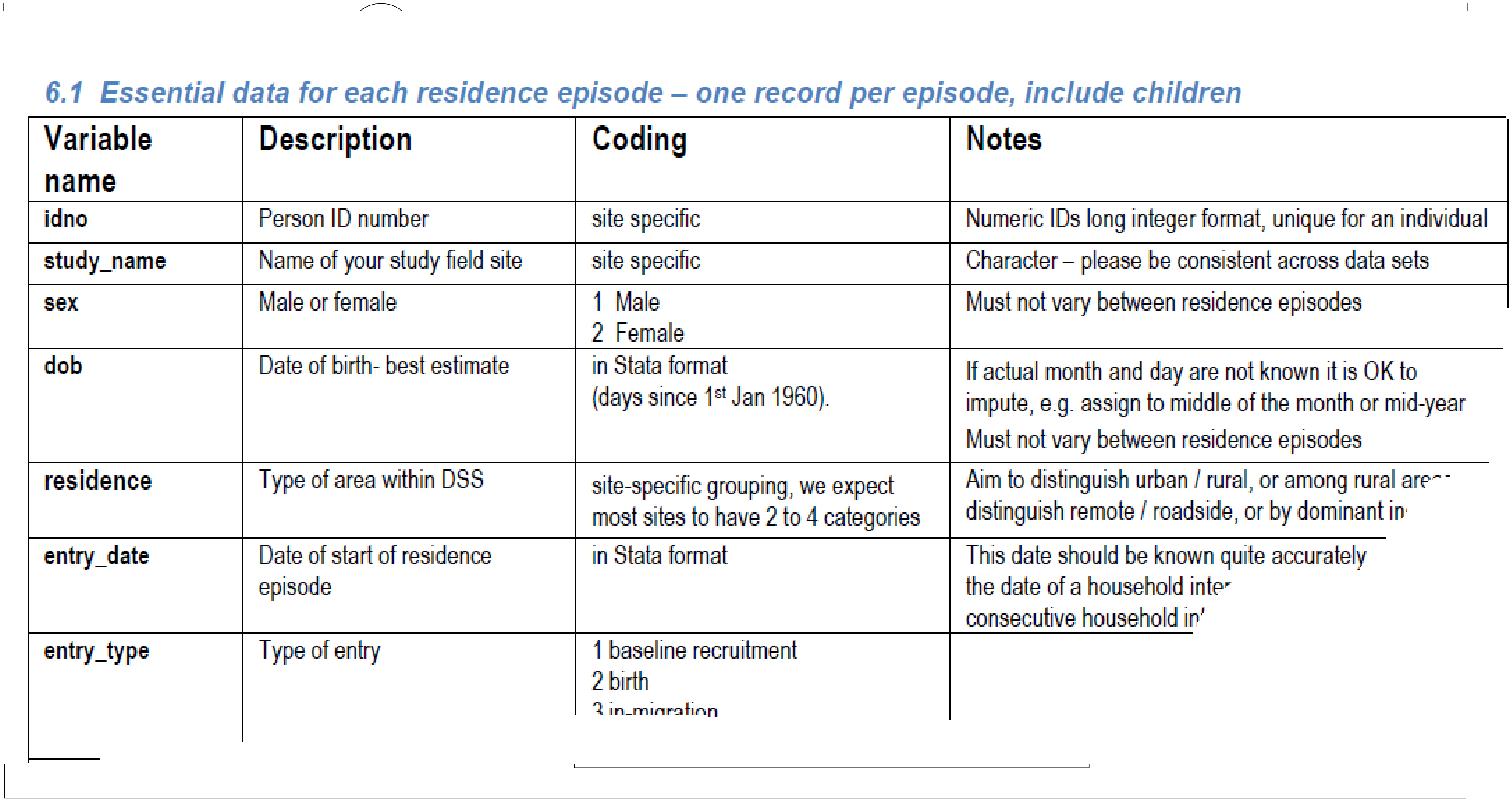
- Located in six high prevalence countries of Eastern and Southern Africa
- Managed by ten independent African research institutions
- Surveillance studies pre-date the network formation
- Facilitated by LSHTM secretariat

ALPHA network aims

- **Analyse community-based HIV surveillance data**
- **Pool data to strengthen analytical conclusions**
- **Present analyses to health policy makers**
- **Build data analysis and data management capability of partner institutions**

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

How are ALPHA datasets prepared?

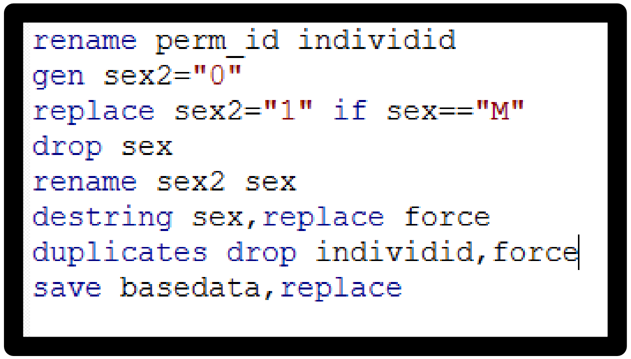


6.1 Essential data for each residence episode – one record per episode, include children

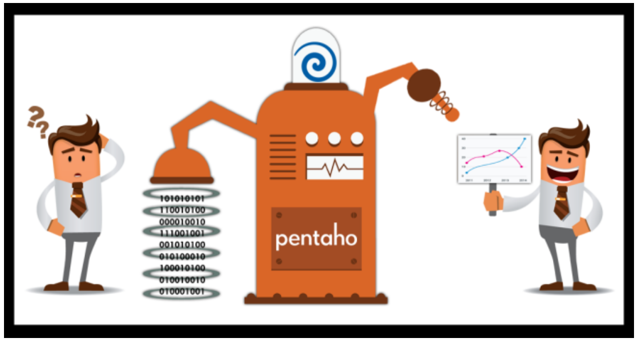
Variable name	Description	Coding	Notes
idno	Person ID number	site specific	Numeric IDs long integer format, unique for an individual
study_name	Name of your study field site	site specific	Character – please be consistent across data sets
sex	Male or female	1 Male 2 Female	Must not vary between residence episodes
dob	Date of birth- best estimate	in Stata format (days since 1st Jan 1960).	If actual month and day are not known it is OK to impute, e.g. assign to middle of the month or mid-year Must not vary between residence episodes
residence	Type of area within DSS	site-specific grouping, we expect most sites to have 2 to 4 categories	Aim to distinguish urban / rural, or among rural are distinguish remote / roadside, or by dominant in
entry_date	Date of start of residence episode	in Stata format	This date should be known quite accurately the date of a household inte consecutive household in
entry_type	Type of entry	1 baseline recruitment 2 birth 3 in-migration	

ALPHA data management over the years…

- Traditionally done in Stata
  - Complex transformations of longitudinal population data, repeated cross sectional surveys and health records
  - Different data managers/ researchers producing the ALPHA data
  - Source data for different sites are organised differently
  - Do-files show how something was done NOT why
  - New staff often start from scratch as they do not understand previously done work
- Now Migrating to Pentaho Data Integration (PDI) funded by Wellcome Trust



```
rename perm_id individid
gen sex2="0"
replace sex2="1" if sex=="M"
drop sex
rename sex2 sex
destring sex,replace force
duplicates drop individid,force
save basedata,replace
```



Motivation

Why is ALPHA considering structured documentation of its data management processes?

Primary motivation

- Need to share data beyond ALPHA

Other drivers

- Efficient multi-site data management and exchange among ALPHA members
- Standardisation will Improve the use of existing data
- Funders’ policies requiring data sharing
- Pooling and sharing data brings credit through data citation.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Issues…

- There is limited domain-specific metadata in Pentaho Data Integration (PDI)
- Need for domain specific framework to guide ALPHA data producers to design and document ETL processes
- There is a metadata specificity gap between business process models and ALPHA ETL implementations.
    - Need for a more concrete documentation of lower level data transformation details

Aims…

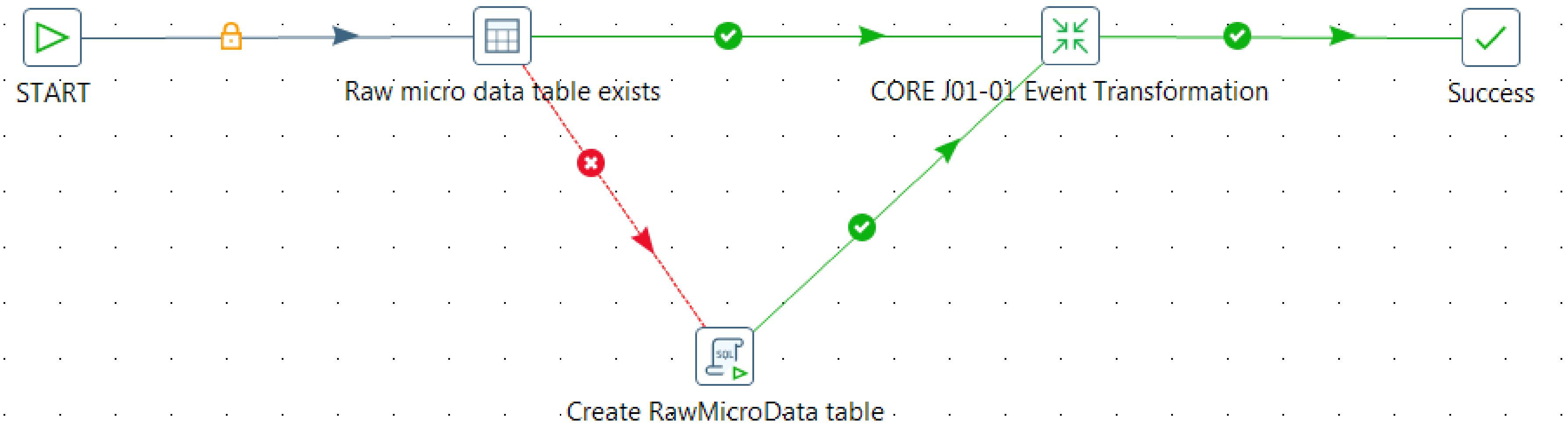
- Enhance transparency maintainability and reuse of ALPHA ETL routines by:
    - Developing a domain-specific business process model that specialises GLBPM
    - Expressing transformations done in PDI using generic data transformation language (SDTL). SDTL is mapped to DDI

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Developing a business process model for ALPHA data management

African population-based Demographic and Epidemiological Surveillance Business Process Model (ADESBPM)

Building ADESBPM: Pentaho data integration



START
Raw micro data table exists
CORE J01-01 Event Transformation
Success
Create RawMicroData table

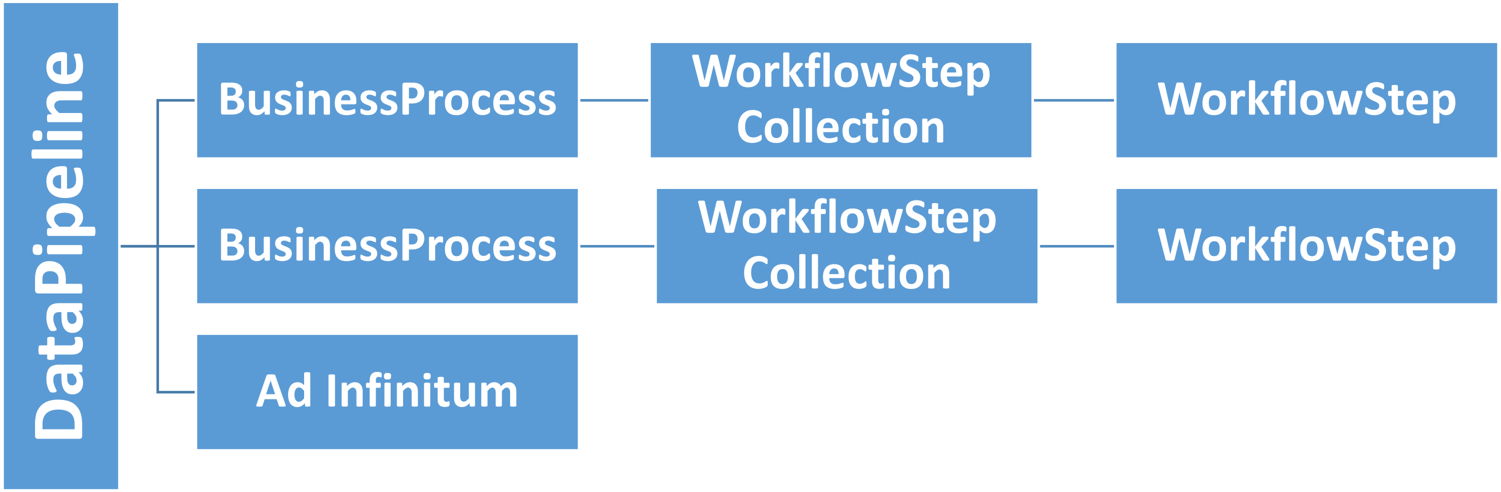
Building ADESBPM: DDI 4

- DDI 4 is built on earlier versions of DDI, the Generic Statistical Business Process Model (GSBPM) and the Generic Statistical Information Model (GSIM)
- DDI 4 prototype comes in several ***use cases***
- Use cases stand on their own and are called ***views*** in DDI 4
- The relevant view for ALPHA ETL is the DataManagementView
- The DataManagementView aims to account for the ingestion and production of new data types (registry data, health data, big data, spell data, event data, etc.) and both legacy and new data management services that give shape to these data types in the course of the data lifecycle…

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

Building ADESBPM: DDI 4 DataManagementView

- At one level DataManagementView has a DataPipeline which comprises a set of business processes traversing a business process model
- Each business step in the pipeline decomposes into a collection of WorkflowStep in DDI 4. Thus, a BusinessProcess is a collection of WorkflowSteps.
- BusinessProcesses and WorkflowStep map to GLBPM steps and sub-steps respectively



Building ADESBPM: PDI & DDI 4 Information models

ETL Information Model	DDI4 Information Model
• **Upper model**	• Upper Model
○ **Job**	○ DataPipeline
▪ **Hops**	▪ List
▪ **Transformation**	▪ BusinessProcesses
• **Lower model**	• Lower Model
○ **Transformation**	○ BusinessProcess
○ **Hops**	▪ WorkflowStepSequence
▪ **Steps**	▪ WorkflowSteps
• **Input**	• MetadataDrivenAction
• **Output**	• ComputationAction
• **Transform**	
• **Joins**	
• **Flow**	
• **Scripting**	
• **More…**	

Building ADESBPM

- Use PDI ETLs to define a set of business activities required to produce ALPHA data
  - Map jobs to GLBPM steps
  - Specialise the GLBPM steps to demographic and epidemiological surveillance concepts, entities and relationships
- GLBPM steps create reusable metadata using a generalised vocabulary.
  - Facilitates communication with users outside ALPHA.
  - Intended for a wide audience thus too broad for ALPHA
- ADESBPM steps create reusable metadata using domain-specific vocabulary
  - Provide a more concrete guide for ALPHA data managers and researchers during design and documentation of ALPHA ETLs

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

Going the last mile…

Mapping Pentaho transformations to Structured Data Transform Language (SDTL)

LONDON
SCHOOL *of*
HYGIENE
&TROPICAL
MEDICINE

Going the last mile: Structured Data Transform Language (SDTL)

- There is a specificity gap between GLBPM (and ADESBPM) and the ALPHA data production systems
  - We are using DDI 4 to bridge that gap
- PDI transformation steps details provide DDI 4 WorkflowStep descriptions.
- Map PDI transformation steps to SDTL
- SDTL maps to DDI 3/4

- SDTL – a model for describing data transformations
- Developed in the C^{2}metadata project
- http://c2metadata.gitlab.io/sdtl-docs/ & http://c2metadata.org/
- The language aims to document data transformations carried out in common statistical packages (SPSS, Stata, R, SAS etc)

Composite type	Properties	Type	Cardinality
ExpressionBase	Name	String	0..1
	TypeName	String	1..1
TransformBase	Command	String	0..1
	SourceInformation	SourceInformation	0..1
Comment	CommentText	String	0..1
Compute	Variable	String	0..1
	VariableRange	VariableRangeExpression	0..1
	Expression	ExpressionBase	0..1
	Condition	ExpressionBase	0..1
Delete	Variables	String	0..n
	VariableRange	VariableRangeExpression	0..n
FunctionCallExpression	Function	String	1..1
	ProprietaryOperator	String	0..1
	IsProprietary	Boolean	0..1
	Arguments	ExpressionBase	0..n
Recode	RecodedVariables	RecodeVariable	0..n
	RecodedVariableRange	VariableRangeExpression	0..1
	Rules	RecodeRule	0..n
RecodeRule	FromValue	string	0..n
	FromValueRange	ValueRange	0..n
	SpecialFromValue	string	0..1
	To	string	0..1
	SpecialToValue	string	0..1
	Label	string	0..1
RecodeVariable	Source	string	0..1
	Target	string	0..1
ReshapeLong	MakeItems	String	0..n
	IndexValues	Int	0..n
	IndexVarName	String	0..1
ReshapeWide	KeepItems	string	0..n
	IdVar	string	0..1
	IndexVar	string	0..1

The last mile: WorkflowStep decomposition

- Identify the SDTL equivalents of the PDI steps used in ALPHA ETLs
- Compile a list of PDI transformation steps that are not accounted for in SDTL (vice versa is unlikely…)
- Suggest extensions to SDTL if needed
- This will provide a way to document lower levels of PDI transformations in DDI

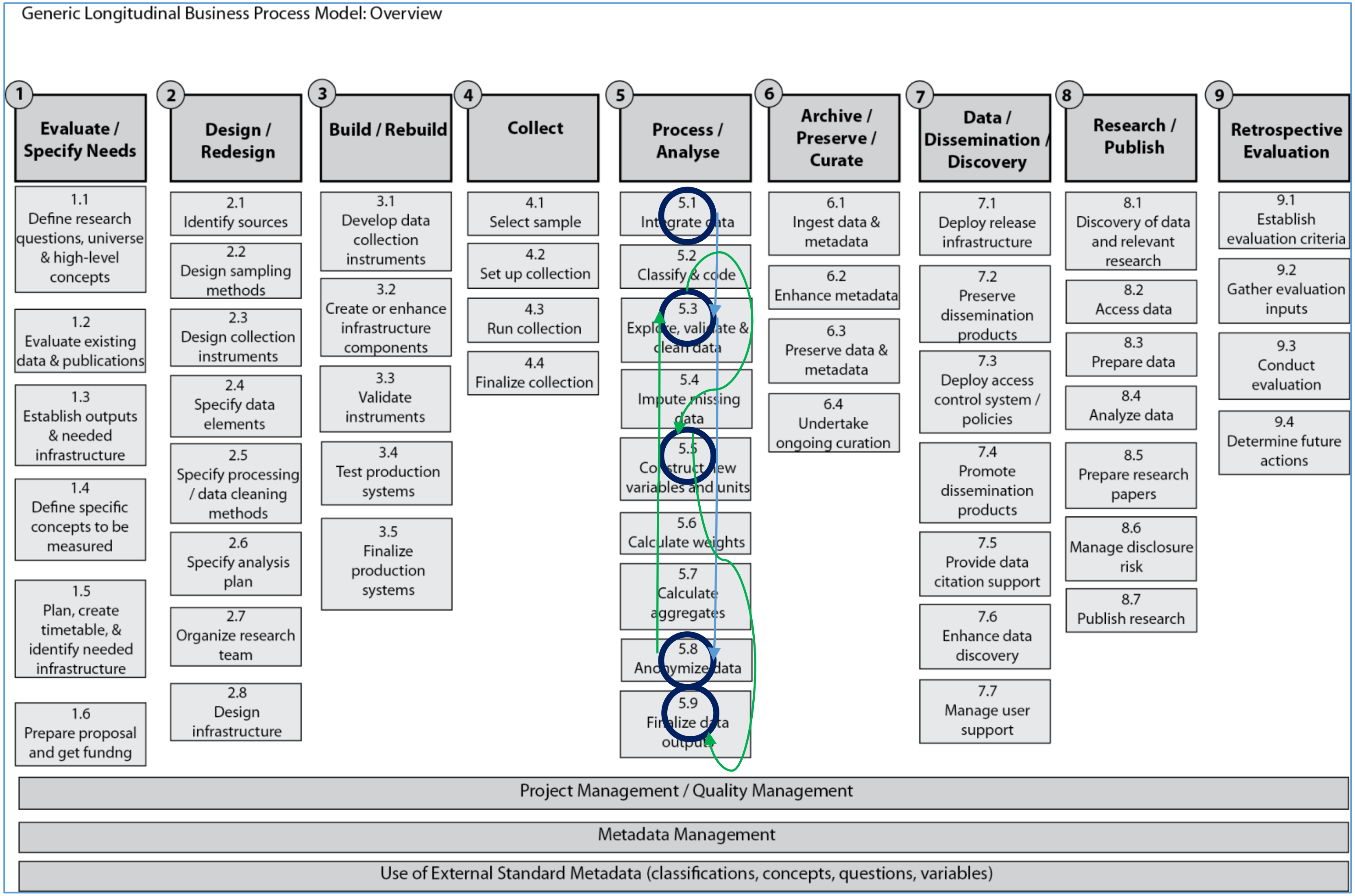
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Where are we now?

Initial results on developing ADESBPM and mapping of Pentaho transformation steps to SDTL

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Where are we now? ADESBPM



Generic Longitudinal Business Process Model: Overview

1 Evaluate / Specify Needs

- 1.1 Define research questions, universe & high-level concepts
- 1.2 Evaluate existing data & publications
- 1.3 Establish outputs & needed infrastructure
- 1.4 Define specific concepts to be measured
- 1.5 Plan, create timetable, & identify needed infrastructure
- 1.6 Prepare proposal and get fundng

2 Design / Redesign

- 2.1 Identify sources
- 2.2 Design sampling methods
- 2.3 Design collection instruments
- 2.4 Specify data elements
- 2.5 Specify processing / data cleaning methods
- 2.6 Specify analysis plan
- 2.7 Organize research team
- 2.8 Design infrastructure

3 Build / Rebuild

- 3.1 Develop data collection instruments
- 3.2 Create or enhance infrastructure components
- 3.3 Validate instruments
- 3.4 Test production systems
- 3.5 Finalize production systems

4 Collect

- 4.1 Select sample
- 4.2 Set up collection
- 4.3 Run collection
- 4.4 Finalize collection

5 Process / Analyse

- 5.1 Integrate data
- 5.2 Classify & code
- 5.3 Explore, validate & clean data
- 5.4 Impute missing data
- 5.5 Construct new variables and units
- 5.6 Calculate weights
- 5.7 Calculate aggregates
- 5.8 Anonymize data
- 5.9 Finalize data outputs

6 Archive / Preserve / Curate

- 6.1 Ingest data & metadata
- 6.2 Enhance metadata
- 6.3 Preserve data & metadata
- 6.4 Undertake ongoing curation

7 Data / Dissemination / Discovery

- 7.1 Deploy release infrastructure
- 7.2 Preserve dissemination products
- 7.3 Deploy access control system / policies
- 7.4 Promote dissemination products
- 7.5 Provide data citation support
- 7.6 Enhance data discovery
- 7.7 Manage user support

8 Research / Publish

- 8.1 Discovery of data and relevant research
- 8.2 Access data
- 8.3 Prepare data
- 8.4 Analyze data
- 8.5 Prepare research papers
- 8.6 Manage disclosure risk
- 8.7 Publish research

9 Retrospective Evaluation

- 9.1 Establish evaluation criteria
- 9.2 Gather evaluation inputs
- 9.3 Conduct evaluation
- 9.4 Determine future actions

Project Management / Quality Management

Metadata Management

Use of External Standard Metadata (classifications, concepts, questions, variables)

Where are we now? ADESBPM

GLBPM	ADESBPM	AlgorithmOverview
5.1 Integrate data	Transform operational data into relevant entities of the HDSS reference data model	1. Create individual table 2. Create events table 3. Create delivery detail table 4. Create delivery-mother-child link table 5. Create cause of death table 6. Populate individual table 7. etc
	Transform HDSS reference data model entities into harmonised data	1. Create raw harmonised data table 2. Merge events table to individuals table 3. Merge deliveries information to 1 and 2 4. Merge cause of death data 5. Etc
5.3 Explore, validate and clean data	Validate sex, dob, events order and events dates	1. Create QualityMatrics table 2. Create events consistency matrix table 3. Create illegal transitions table 4. Create starting events table 5. Etc 6. Compile illegal events ordering

Where are we now? Pentaho to SDTL mapping

- We have started identifying SDTL equivalents of some PDI steps used in ALPHA
- Some PDI steps could not be mapped to SDTL

SDTL Composite type	**Pentaho step**
Compute	calculator, formula
FunctionCallExpression	Sort,
Load	input
Recode	Value mapper
ReshapeLong	Row Normaliser
ReshapeWide	Row denormaliser
Save	output
Select	Select
?	String operations
?	Add value fields changing sequence

Where to next?

- ADESBPM: Will review ALPHA PDI ETLs and

add detail to ADESBPM based on this review

- Provide ADESBPM template that the ALPHA network management can use as a framework for designing and documenting ALPHA ETLs across member institutions
- Complete Pentaho to SDTL mapping to find SDTL equivalents for all transformation steps in ALPHA PDI ETLs
- Compile transformation steps not in SDTL to work as basis of extending SDTL on ETL platforms
- Develop DDI tools on top of the ADESBPM and SDTL that present the data lineage of PDI data products in a graphical format

Acknowledgements

- ALPHA network
- Data Documentation Initiative Alliance
- Research participants

Maraba
Tatenda
tack
Ahsante
Merci
Aitäh
Tá
kealeboga
wabeja
obrigado
Tsikomo
Danke
gracias
Tak
dankje
धन्यवाद
谢谢
Dankie
tawonga
Sikomo
shukrān
enkosi
Thankyou
Ndolivhuwa
Zikomo