Open-access for existing LMIC demographic surveillance data using DDI

by Chifundo Kanjala[1], Jim Todd[1], David Beckles[2], Tito Castillo[3], Gareth Knight[1], Baltazar Mtenga[4], Mark Urassa[4], and Basia Zaba[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.

Keywords

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

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. 2013; 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 2007; 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 1) 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 2015), 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.

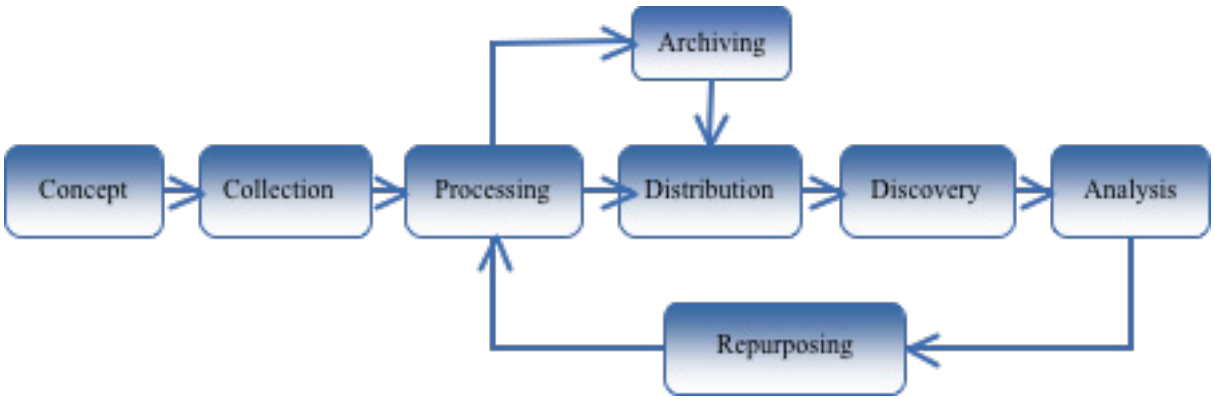

Figure 1: Data life cycle: DDI Lifecycle conceptual model

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 2. 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 (Barkow et al. 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.

Study settings and methods

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

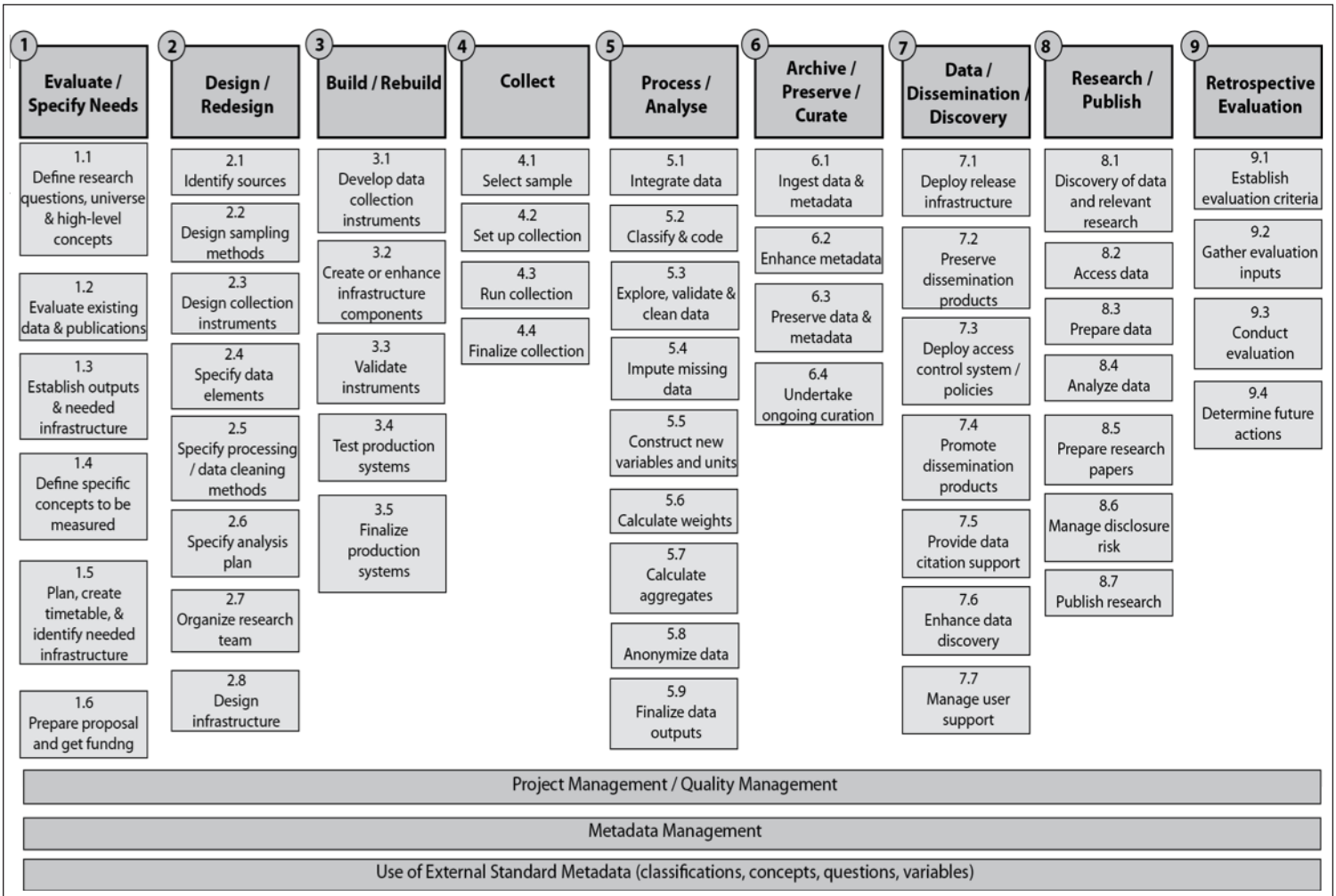

Figure 2: Generic Longitudinal Business Process Model

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.

Study methods

The data production process involved in the implementation of a typical HDSS data collection round in Kisesa is illustrated in Figure 3. 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 (Barkow et al. 2013). We then used the existing mapping from GLBPM to DDI Lifecycle (Barkow et al. 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.

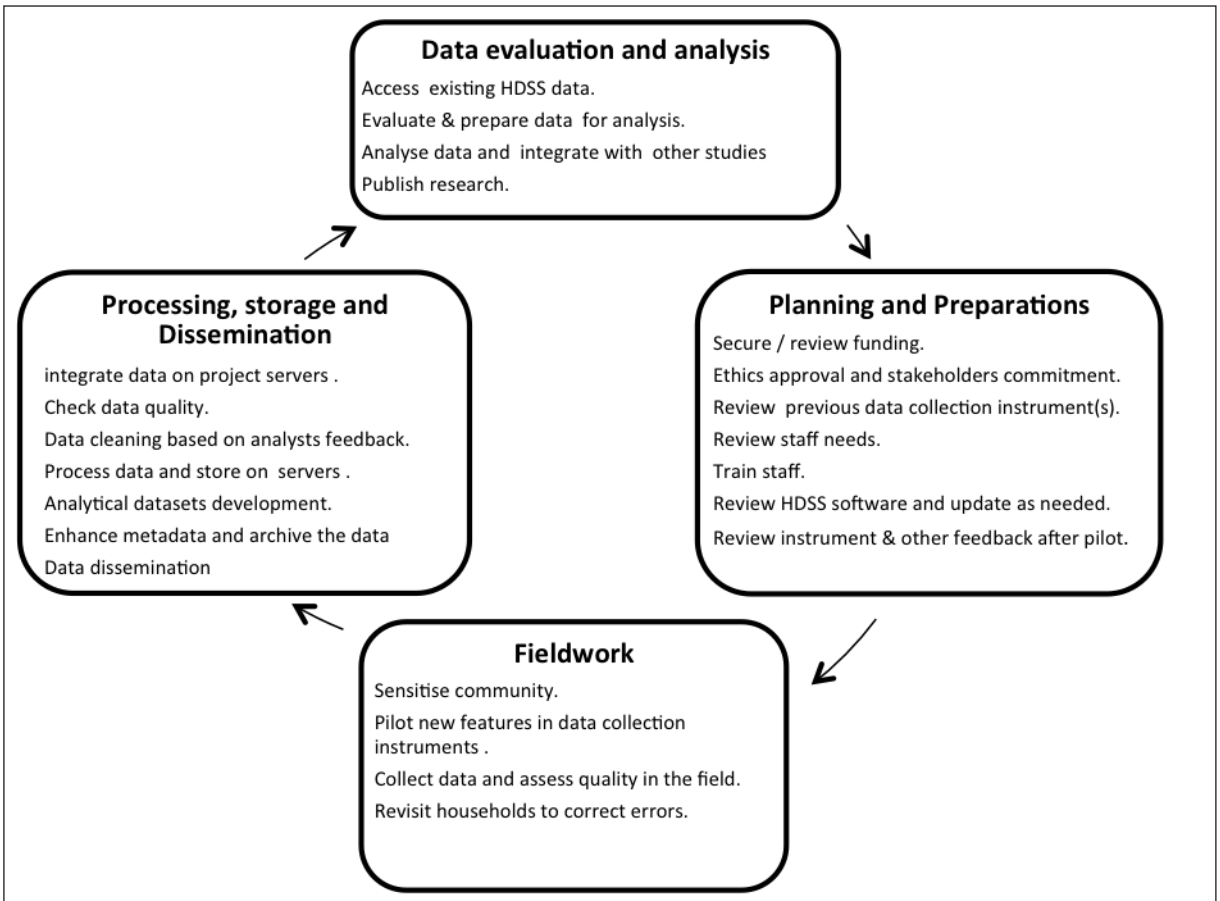

Figure 3: Steps in the Kisesa study HDSS data collection round

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.

Results

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 4. The GLBPM steps are shown in square brackets.

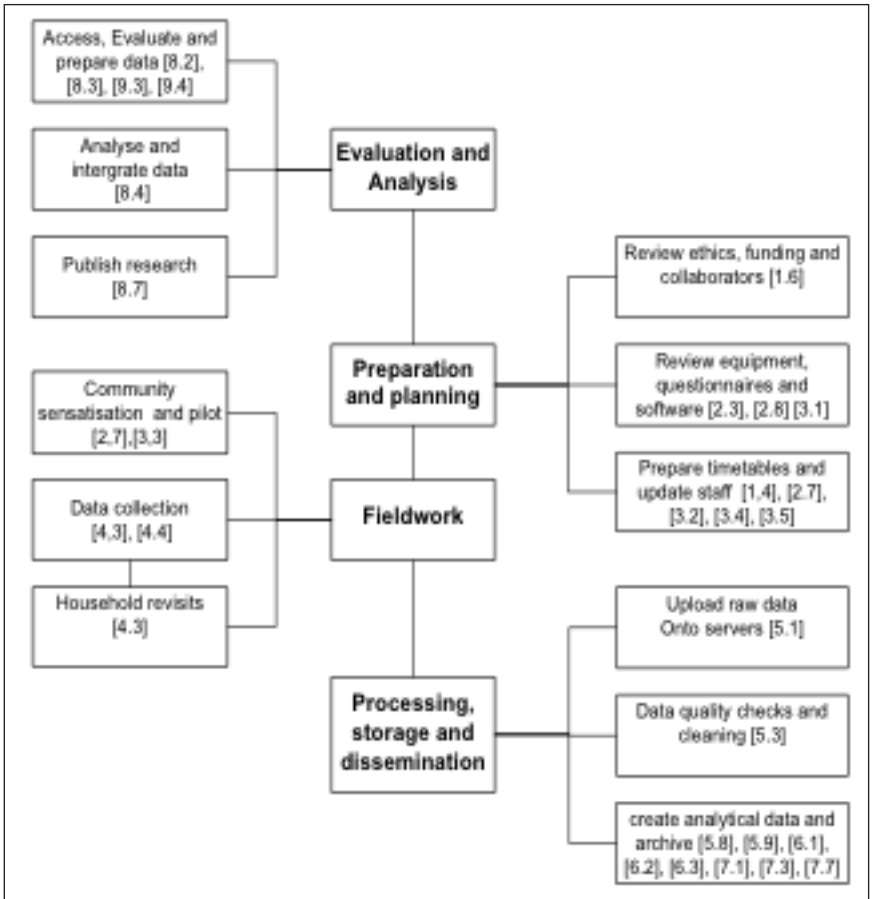

Figure 4: Kisesa study HDSS Documentation in Nesstar Publisher

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.

Implementation of DDI in Nesstar Publisher and Colectica Designer

Table 1 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 1: 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[5]	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. 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 2. 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.

	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

Table 2: Nesstar Publisher (NP) and Colectica Designer (Colectica) documentation

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 2 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 3. (page 28).

	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

Table 3: Training materials and software costs

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 website[6] 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. 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 3.
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.

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.

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 2015), 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.

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.

References

Africa Centre for Population Health. 2015. 'Research Data Management Platform'. *http://www.africacentre.ac.za/index.php/data-research-management*.

African Population and Health Research Center. 2015. 'Central Data Catalog'. *http://aphrc.org/catalog/microdata/index.php/catalog*.

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.

Beguy, Donatien, Patricia Elung'ata, Blessing Mberu, Clement Oduor, Marylene Wamukoya, Bonface Nganyi, and Alex Ezeh. 2015. 'HDSS Profile: The Nairobi Urban Health and Demographic Surveillance System (NUHDSS)'. International Journal of Epidemiology, dyu251.

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*.

Colectica. 2015a. 'Colectica — Colectica 5.1 Documentation'. *http://docs.colectica.com/*.

———. 2015b. 'Colectica Information Model — Colectica 5.1 Documentation'. *http://docs.colectica.com/introduction/information-model/*.

de Bruijne, Marika, and Alek Amin. 2009. 'Questasy: Online Survey Data Dissemination Using DDI 3'. IASSIST Quarterly 33 (Spring): 10–15.

de Vet, Edwin. 2013. 'Update on Questasy, a Data Dissemation 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*.

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'.

Digital Curation Centre. 2013. 'Nesstar | Digital Curation Centre'. June 12. *http://www.dcc.ac.uk/resources/external/nesstar*.

Dupriez, Olivier, and Geoffrey Greenwell. 2007. 'Quick Reference Guide for Data Archivists'. *http://www.ihsn.org/home/node/544*.

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. *doi: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*.

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.

Hebing, Marcel. 2015. 'A Metadata-Driven Approach to Panel Data Management and Its Application in DDI on Rails'.

INDEPTH Network. 2002. Population, Health and Survival at INDEPTH Sites. International Development Research Centre.

International Household Survey Network. 2013. 'Mission and Objectives | IHSN'. *http://www.ihsn.org/home/content/about/objectives*.

———. 2016. 'Microdata Cataloging Tool (NADA) | IHSN'. *http://www.ihsn.org/home/software/nada*.

Jensen, Jannik. 2012. 'DdiEditor'. In EDDI12–4th Annual European DDI User Conference.

Kishamawe, Coleman, Raphael Isingo, Baltazar Mtenga, Basia Zaba, Jim Todd, Benjamin Clark, John Changalucha, and Mark Urassa. 2015. 'Health & Demographic Surveillance System Profile: The Magu Health and Demographic Surveillance System (Magu HDSS)'. International Journal of Epidemiology, dyv188.

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. *doi:10.1111/j.1365-3156.2011.02924.x*.

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.

Nielsen, Mogens, Jeremy Iverson, and Dan Smith. 2013. 'Standardized Quality Declarations with DDI, SDMX, and Colectica'. In .

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.

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, and Grant Blank. 2007. 'The Data Documentation Initiative: A Preservation Standard for Research.' Archival Science 7 (1): 55–71.

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.

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. *doi: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.

Todd, Jim, Judith R Glynn, Milly Marston, Tom Lutalo, Sam Biraro, Wambura Mwita, Vinai Suriyanon, Ram Rangsin, 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.

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.

W3schools.com. 2015. 'XML Introduction - What Is XML?' *http://www.w3schools.com/xml/xml_whatis.asp*.

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*.

Zaba, Basia, Clara Calvert, Milly Marston, Raphael Isingo, Jessica Nakiyingi-Miiro, Tom Lutalo, Amelia Crampin, Laura Robertson, Kobus Herbst, and Marie-Louise Newell. 2013. '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.

Notes

1. London School of Hygiene and Tropical Medicine
2. Independent IT Consultant, UK
3. University College London Hospitals NHS Foundation Trust
4. National Institute for Medical Research, Mwanza Tanzania
5. 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
6. *http://www.ddialliance.org/*