Research transparency and data sharing in ALPHA

Chifundo Kanjala, Jay Greenfield, Arofan Gregory, Emma Slaymaker and Jim Todd

Abstract

Metadata, usually defined as data about data, are often incomplete in Health and Demographic Surveillance System (HDSS) harmonised datasets. Inspired by the work in the iSHARE project which resulted in the most comprehensive solution for harmonisation and curation of HDSS data to date, the Centre in a Box (CiB), we sought to extend its data provenance documentation capabilities. Hitherto, these were tool-specific thus inflexible for cross platform management and sharing. We investigated the provision of user-friendly access to data harmonisation metadata in a network of HDSS studies by applying a semi-automated documentation approach and running a requirements elicitation study with data managers and researchers. A business process model specialised to HDSS context captured the high level details of the HDSS data transformation routines. Proposed features for a metadata browser were well received by interviewees. These findings have implications for data documentation standards development and HDSS in data management automation.

Introduction

Consider the diagrams in Figure 1 and Figure 2. In Figure 1 we see sample code from the Stata software (StataCorp, 2019) for performing various data transformations. These include creating a new variable, recoding the values of a variable (birthyear), and reshaping data from one format called wide format ("short and fat") into long ("tall and skinny") format.

Figure 1: Stata data transformation code

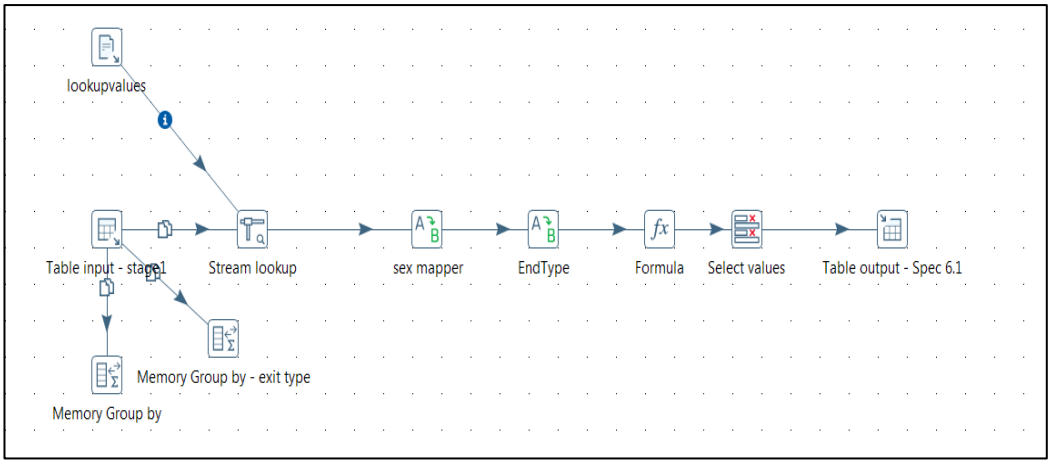
```
use "`root'/serodata.dta", clear

keep hhid memberid birthyear cd4_1 cd4_2
preserve
gen birthyears_5 = birthyear
recode birthyears_5 1989/1993=1 1994/1998=2 1999/2003=3 2004/2008=4

reshape long cd4, i(id round) j(visit)
drop if cd4==.
rename cd4 cd4count
rename birthyears_5 birthyeargroup
```

On the other hand, Figure 2 shows another set of data transformations performed in a platform different from Stata, Pentaho data integration (Pentaho Corporation, 2018). Here, a variable is being recoded from one set of values to another, some observations are being selected from the data in memory to create an output dataset among other transformations. Figure 2 is providing a graphical expression of the transformations while in Figure 1, Stata code is provided. Such data transformations are applied to Health and Demographic Surveillance System (HDSS) data in harmonisation projects such as the network for Analysing Longitudinal Population-based HIV/AIDS data on Africa (ALPHA) (Reniers et al., 2016). Structured documentation of these transformations compliant with international best practice is generally missing today.

Figure 2: Exemplar Pentaho data transformation



The absence of such provenance metadata poses data management and usability challenges. The human users find it hard to understand and correctly interpret the harmonised data. Software agents also struggle to exchange and exploit the data in automated ways. An investigation of options for providing this documentation may help to ameliorate the problems.

Research problem

Provenance or lineage of HDSS harmonised datasets comprise of metadata on the primary data used as input, metadata for the data transformation routines and metadata on the derived datasets.

Existing literature covers the documentation of the primary HDSS datasets (Kanjala et al., 2017) and the final harmonised data (Herbst et al., 2015). We still have a gap in our knowledge relating to the documentation of harmonisation routines. While there are some approaches that have been devised to document data transformations, they have mainly been outside HDSS settings. It remains to be assessed how they fair when applied to harmonised datasets.

Research contributions

This paper reports on the efforts within ALPHA to provide user-friendly access to tool-agnostic provenance metadata for retrospectively harmonised datasets. In this study, we went beyond free text description and reference to the code (Figure 1) or graphical flow diagrams (Figure 2). We developed structured metadata formatted in compliance with the internationally recommended metadata standards. In addition, we ascertained the requirements for the provision of these provenance metadata in user-friendly format. This was done through an online requirements elicitation study with experts working in population-based health research data harmonisation projects within and outside ALPHA.

Paper overview

The rest of this paper is organised as follows. The next section gives a literature survey related to the reasons for data documentation and the available models, standards and technologies. The literature is followed by a section on the study settings, a brief overview of the ALPHA

network its data management practices. Next, we consider the development of structured metadata for ALPHA data transformations. This is followed by the requirement elicitation study. The paper ends with a discussion of the findings, their implications for HDSS data management and potential further steps needed.

Related literature

There are two sides to the literature on data documentation - the demand side and the supply side. The FAIR guiding principles for scientific data management and stewardship (Wilkinson et al., 2016) provide a succinct description of the reasons for demanding data documentation and the associated benefits. On the other hand, the supply side is encapsulated within a data and statistics production architecture (Bruno, Duma, Scannapieco, Silipo, & Vaste, 2016) developed in the official statistics community. In addition, we also review work on the documentation practices of data harmonisation projects.

Data documentation demand

Players on the demand side are research funders (Pisani et al., 2016; Walport & Brest, 2011), research publishers (Federer et al., 2018) and secondary data users (Chandramohan et al., 2008). The FAIR principles state that optimal stewardship ensures that data are Findable, Accessible, Interoperable and Reusable (FAIR) (Wilkinson et al., 2016). Beyond collection, processing and storage, FAIR principles foster value addition for purposes of data discovery and reuse. Data processing routines also need to be shared to facilitate correct interpretation of derived datasets (Bergeron, Doiron, Marcon, Ferretti, & Fortier, 2018; Wilkinson et al., 2016).

Metadata are a key component in achieving the FAIR criteria for data stewardship. They are the bridge that connects data and their users, without them, data are just a collection of meaningless numbers (Ryssevik, 1999). The metadata needs to cater for both humans and software agents as equally important users (IHSN, 2012; Wilkinson et al., 2016). Descriptive documentation delivers knowledge to human users while the well-structured aspects of the documentation allows software agents to find, access, exchange and process the data and metadata in highly automated ways (IHSN, 2012). In the long run, optimal documentation enhances data quality and lowers data production costs.

Data documentation supply

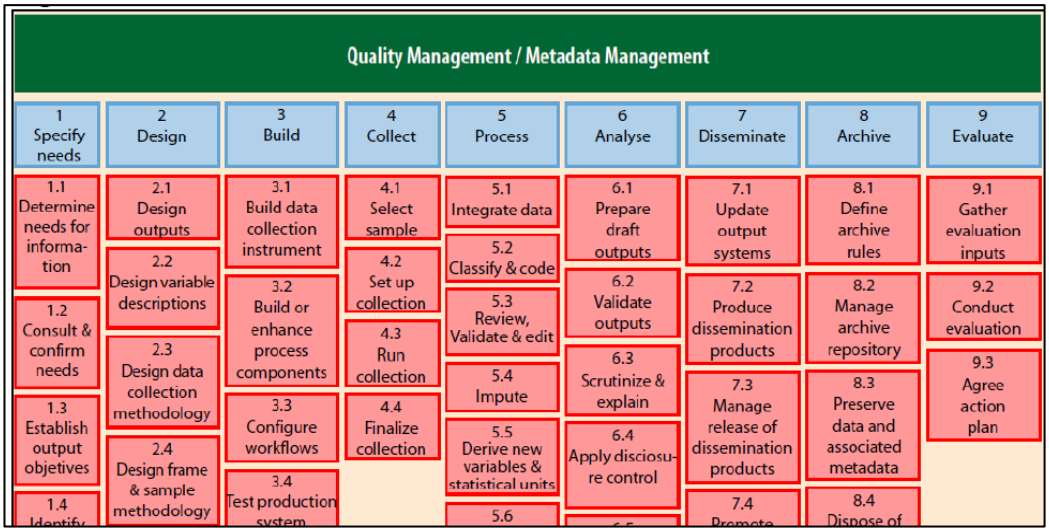
The data documentation supply side relevant for HDSS is represented by technologies, models and standards developed within official statistics and research data documentation communities. These have aimed to satisfy the FAIR principles and are connected under an enterprise architecture (EA) framework. The Enterprise architecture is driven by the understanding that business needs and strategies for a data production enterprise are the ones to guide the choices of the information models and technologies to use. According to this framework, an organisation's data production architecture is considered to comprise of a business architecture, an information architecture, an application architecture and a technology architecture.

Business architecture – Generic Business Process Models

Generic process models have been used for high level definition and description of data production processes in official statistics and longitudinal survey research (Barkow et al., 2013; UNECE, 2018). The Generic Statistical Business Process Model (GSBPM) is a reference model

for description of official statistics production (UNECE, 2018). Figure 3 shows a truncated version of the GSBPM.

Figure 3: Generic Statistical Business Process Model



(Thérèse Lalor & Steven Vale, 2013)

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. It has been widely adopted by national and international statistics offices (Ausborn, Rotondo, & Mulcahy, 2014; Brancato & Simeoni, 2012; UNECE Secretariat, 2009).
A specialisation of the GSBPM, the Generic Longitudinal Business Process Model (GLBPM), has been developed within the human science research data documentation community. It is aimed at a better description of the activities involved in the longitudinal survey data production process than that offered by the GSBPM (Barkow et al., 2013).
The GLBPM is more relevant to HDSS data production compared to the GSBPM.

Information architecture –Information Models

The official statistics developed an information model called the Generic Statistical Information Model (GSIM). It was designed to complement the GSBPM capturing, at a conceptual level, the pieces of information (information objects) used in or produced from the sub-processes of the GSBPM (UNECE 2018a). The information objects involved include data, metadata, rules and parameters among others. On the other hand, the human science research data documentation community is building an information model for the information objects flowing between activities in the GLBPM (William Block et al., 2012), the Data Documentation Initiative Information Model (DDI IM) (DDI Alliance, 2019, p. 4).

Application Architecture – Metadata standards and languages for data transformations documentation

Within this layer, we find standards which support metadata development. The relevant standards for HDSS data are the Data Documentation Initiative (DDI) (DDI Alliance, 2018) and the Statistical Data and Metadata eXchange (SDMX) standard (SDMX, 2018). DDI is focussed on documenting microdata while SDMX is more suitable for aggregated data. Both DDI and SDMX have limitations when it comes to documenting data transformations. They use textual descriptions and reference to the original code.

To address these forgoing weaknesses in DDI and SDMX, two dedicated languages for documenting data transformations have been developed. These are the Validation and Transformation Language (VTL) (SDMX Technical Working Group, 2018) and the Structured Data Transformation Language (SDTL) (C2Metadata, 2017). SDTL and VTL are concerned with the description of the granular details of the transformations. With VTL, one first creates VTL code then convert the code to a programming language of interest using parsers. To use SDTL, one starts with code prepared in SPSS, Stata, R or SAS. This code is converted to SDTL using appropriate parsers.
The work presented in this paper is focussed on the high-level description of the transformation, therefore, it does not touch on the granular details supported by SDTL and VTL.

Technology Architecture – SDMX-based and DDI-based tools and other tools

A number of metadata standards-based tools have been developed either generic enough for use across projects implementing the metadata standards or bespoke to specific projects. There are a number of tools for implementation of SDMX and DDI. VTL and SDTL are still under development and so are the tools.

Data documentation in harmonisation projects

Four data harmonisation projects are considered here as they are, to the authors' knowledge, among the best efforts in documenting public health harmonised data. Maelstrom research (Fortier et al., 2017) has provided guidelines and tools for systematic data harmonisation and their data documentation tools are DDI-based. The CLOSER project (CLOSER, 2019; O'Neill et al., 2019) has also used DDI for data documentation and has an extensive questionnaire documentation. The CLOSER harmonised data are shared through the CLOSER discovery platform (CLOSER, 2019). In addition, the Gesis institute in German has developed a data harmonisation tool called CharmStats (Winters & Netscher, 2016). It documents data transformations performed using SPSS.
In the HDSS domain, the seminal work by (Herbst et al., 2015) is widely recommended for harmonising and curating HDSS data. Herbst et al. (2015) describes an HDSS data harmonisation and curation infrastructure called the Centre in a Box (CiB). The CiB comprises of a portable mini-server hardware which contains 3 virtual servers. The first hosts a database management system used by a member institution and replicates the institution's operational database. This facilitates the transfer of data from operational databases into a data harmonisation environment. The second virtual server is a data manager's desktop which hosts the data harmonisation and documentation software. The third one manages the system's security, shared file system and a web server implementing a local instance of a data cataloguing tool called National Data archive (NADA) (International Household Survey Network, 2016).
In all these projects, data harmonisation routines are not documented in tool-agnostic and structured ways. The task of converting the transformation routines between proprietary software is therefore left to the users. Developing tool-agnostic generic documentation for transformation routines performed on HDSS data will make the harmonised data more accessible to users.

Study settings

ALPHA network overview and data management

The ALPHA network is described in the two 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/). It is a research programme focussing on broadening the evidence base of HIV epidemiology through multi-site data harmonisation, pooling and analysis (Reniers et al. 2016). It is a collaboration of ten autonomous health research institutions in Eastern and Southern Sub Saharan Africa and the London School of Hygiene and Tropical Medicine in the global north. The network members have published their individual study profiles (Asiki et al., 2013; Beguy et al., 2015; Crampin et al., 2012; Geubbels et al., 2015; Gregson et al., 2017; Kahn et al., 2012; Kishamawe et al., 2015; Odhiambo et al., 2012; Tanser et al., 2008).

ALPHA does not collect primary data, rather it transforms data collected within its members for secondary analysis. 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 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. Staff turnovers have also posed challenges relating to reproducibility of the transformations. ALPHA is currently migrating to the use of Pentaho for a more robust and standardised data processing. Pentaho belongs to a class of software called Extract-Transform and Load (ETL) software, it offers powerful data transformation options and makes reproducibility easier than Stata. Pentaho is being used within the CiB environment. 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).

Structured metadata for ALPHA data transformation routines

Methods

This study adapted the data and statistics production architecture developed in the data documentation community to prepare provenance metadata for ALPHA. The data processing routines analysed in this study are those for creating the ALPHA data specification on residency episodes. This dataset contains information on residence in the study area, including dates of birth, migration and death.

A specialised business process model was developed to describe the high-level aspects of the data transformations. The DDI information model was used to capture the information objects flowing between the sub-processes of the business process as inputs and outputs. The contents of this model were expressed as DDI XML.

The following procedure was followed:

1. The transformation routines performed in Pentaho were analysed to identify the main activities and tasks involved in producing this dataset.
2. The identified activities were mapped to the GLBPM.
3. The metadata that could be automatically harvested from Pentaho using bespoke software agents were identified and extracted.
4. Supplementary metadata were added to those automatically mined

5. The mapped steps were specialised to suit HDSS contexts, capturing events of interest (births, deaths and migration), their order of occurrence and their timing. This specialisation relied heavily on the HDSS reference data model (Benzler, Herbst, & Macleod, 1998)

Results

Table 1: Mapping ALPHA business process to GLBPM and specialisation

Sub-job	GLBPM	Algorithm overview (Specialisation)
CORE Produce Raw 61 Dataset	5.8 Anonymise data 5.1 Integrate data	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	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	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

Figure 4: Specialisation of the GLBPM expressed in DDI XML format

```
<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>
```

Figure 5: Scores for proposed features of a provenance metadata browsing platform

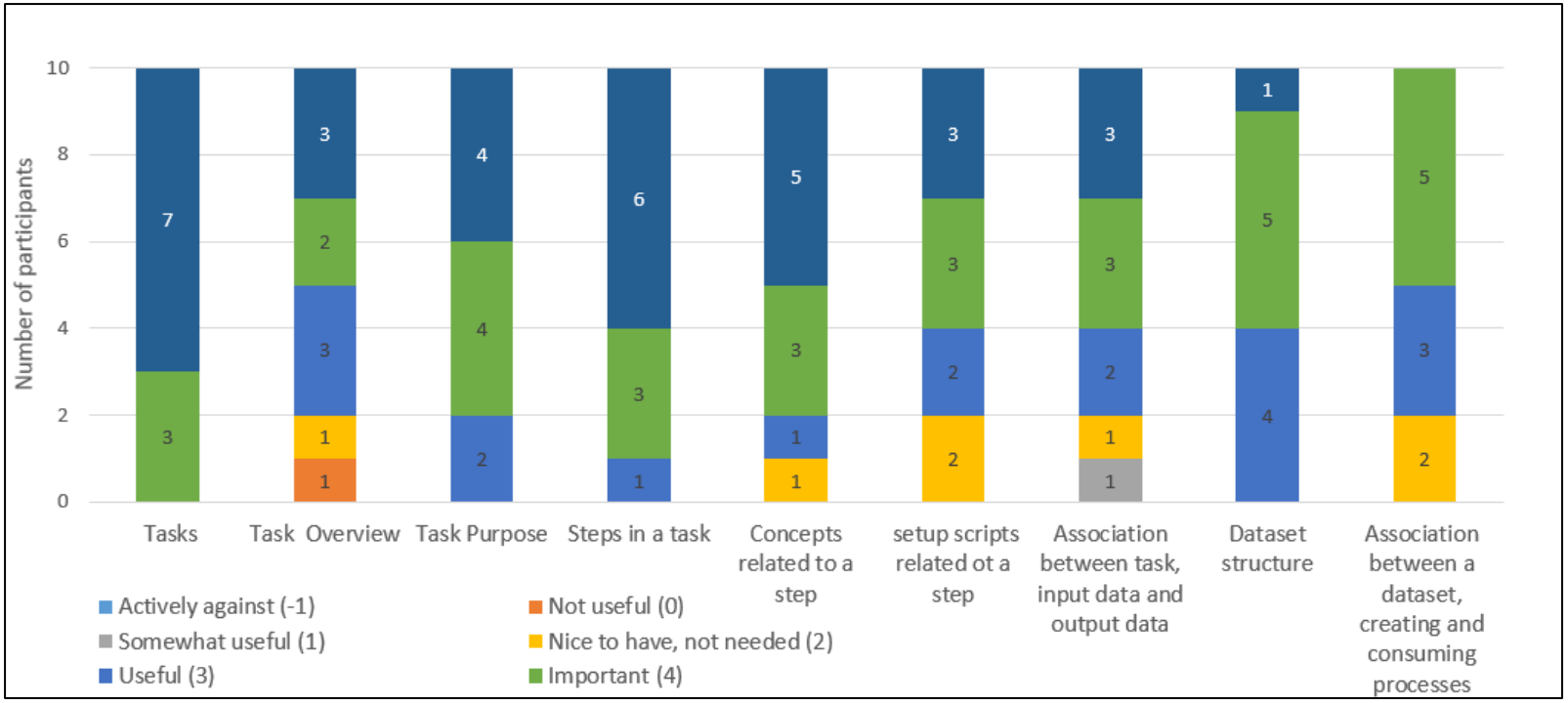


Table 1 shows the results of mapping tasks performed in Pentaho (first column) to the GLBPM (second column) and the specialisation (third column). The mapping to the GLBPM resulted in broad descriptions of tasks at hand. It was through the specialisation that the tasks were described in more concrete terms that practitioners working in HDSS can relate to. The specialisation also captured the tasks in tool-agnostic terms. A potential user would see what was done not necessarily how it was done in Pentaho.
Figure 4 shows the metadata presented in the third column and second row in Table 1 but represented in XML format. This is storage format that software agents are able to use in automating tasks of data processing and exchange. It also has the characteristic that it can be converted into forms that are human intelligible in automated ways.

We compiled all the Algorithm overviews for the entire process of creating ALPHA residency episodes harmonised data. We then integrated these with the input and output datasets. This resulted in a specialised business process model for African population-based demographic surveillance called the African Demographic and Epidemiological Surveillance Business Process Model (ADESBPM). This captures, at a business level, the tasks involved in creating a harmonised dataset in networks such as ALPHA.

Providing user-friendly access to ALPHA provenance metadata: ALPHA metadata browser requirements elicitation study

While machines need structured metadata such as those in Figure 4 to process and exchange data and metadata in automated ways, human users require human intelligible formats. Since this study is the first one to ever create structured, standards based metadata for ALPHA data transformations, the requirements for human-user friendly representation of those provenance metadata are unknown. The elicitation study was therefore done to ascertain those requirements.

Methods

Mock-ups of proposed features: A set of mock-up diagrams were created comprising of proposed features for a user friendly metadata browser. These mock-ups were based on the structured provenance metadata and they formed the core of a qualitative study to elicit the views of data managers', demographers and epidemiologists' on the features.

Recruitment, data collection and analysis: A convenience sample of 10 participants was drawn from an experienced cadre of data managers and researchers within and outside the ALPHA network. Prior to the interview, mock-up diagrams of the proposed features of the software were emailed to the participants to give them time to analyse them. Each participant was then interviewed over skype on the features in the mock-up diagrams. The participants graded each feature's importance 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. Data were analysed using the NVivo software, version 12. The analysis included identification of features specified as essential, not essential, and those not in the mock-ups but perceived as vital.

Results

Scores for proposed features

Figure 5 gives the number of interviewees who gave a particular score for each of the features that were included in the mock-up diagrams. 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. To help clarify the perspective of the participants, the rationales of the participants are presented next.

Rationale for/ against proposed features and suggested improvements

None of the proposed features were out rightly rejected. In the majority of the cases, 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, the suggestion to have the *dataset –centric* and the *task centric* views together in one diagram and the suggestion to develop concepts further into site-specific concepts.

Respondents also wanted to see an integrated platform showing the three dimensions of harmonised data documentation – the metadata for the source data from HDSS, the metadata for the transformation routines and the metadata for the resulting harmonised datasets.

Discussion

This paper sought to investigate the provision of user-friendly access to tool-agnostic provenance metadata for retrospectively harmonised datasets in HDSS data pooling networks using ALPHA as a prototype. Through the customisation of data and statistics production frameworks developed in the official statistics and data documentation communities, such metadata were developed for an ALPHA prototype dataset.

The given results include contents of a business process model specialised to HDSS, the ADESBPM and an analysis of requirements for a human user friendly platform for browsing and searching the metadata.

The presented results show that with a combination of existing models and standards, the job of documenting transformations such as those performed in ALPHA can be done. However, it is a time and labour intensive undertaking. It requires skilled personnel to analyse the transformations and relate them to the theory underpinning the logic behind the transformations. Tools are required to better streamline and automate the processes of producing these metadata.

The gathered requirements are a step towards the provision of user-friendly access to the developed metadata. There is need for developers to now take the requirements and convert them into a fit for purpose metadata browsing and searching software.

Overall, there is a gap in the market for tools to automate the described metadata development and access provision. Retrospectively harmonised longitudinal studies would be the primary beneficiaries targeted by these tools but other harmonised data production studies may benefit too.

Until documentation as described here is built into studies and is completely routine, data sharing will always be hampered.

References

Asiki, G., Murphy, G., Nakiyingi-Miiro, J., Seeley, J., Nsubuga, R. N., Karabarinde, A., … Pomilla, C. (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), 129–141.

Ausborn, S., Rotondo, J., & Mulcahy, T. (2014). Mapping the General Social Survey to the Generic Statistical Business Process Model: NORC's Experience. *IASSIST QUARTERLy*, 21.

Barkow, I., Block, W., Greenfield, J., Gregory, A., Hebing, M., Hoyle, L., & Zenk-möltgen, W. (2013). GENERIC LONGITUDINAL BUSINESS PROCESS MODEL DDI Working Paper Series – Longitudinal Best Generic Longitudinal Business Process Model. *Business*, 1–26.

Beguy, D., Elung'ata, P., Mberu, B., Oduor, C., Wamukoya, M., Nganyi, B., & Ezeh, A. (2015). HDSS Profile: The Nairobi Urban Health and Demographic Surveillance System (NUHDSS). *International Journal of Epidemiology*, dyu251.

Benzler, J., Herbst, K., & Macleod, B. (1998). A Data Model for Demographic Surveillance Systems 1. *Science*.

Bergeron, J., Doiron, D., Marcon, Y., Ferretti, V., & Fortier, I. (2018). Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit. *PloS One*, *13*(7), e0200926.

Brancato, G., & Simeoni, G. (2012). *Istat statistical process modelling and the Generic Statistical Business Process Model: a comparison*. Presented at the European Conference on Quality in Official Statistics Q.

Bruno, M., Duma, R., Scannapieco, M., Silipo, M., & Vaste, G. (2016). CORE: a concrete implementation of the CSPA architecture. *Statistical Journal of the IAOS*, *32*(4), 591–596.

C2Metadata. (2017). Structured Data Transform Language. Retrieved November 15, 2018, from Structured Data Transform Language website: http://c2metadata.gitlab.io/sdtl-docs/

Chandramohan, D., Shibuya, K., Setel, P., Cairncross, S., Lopez, A. D., Murray, C. J. L., … Binka, F. (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. Retrieved August 24, 2019, from https://wiki.ucl.ac.uk/display/CLOS/CLOSER+Discovery

Crampin, A. C., Dube, A., Mboma, S., Price, A., Chihana, M., Jahn, A., … Branson, K. (2012). Profile: the Karonga health and demographic surveillance system. *International Journal of Epidemiology*, *41*(3), 676–685.

DDI Alliance. (2018). Welcome to the Data Documentation Initiative | Data Documentation Initiative. Retrieved February 19, 2019, from https://www.ddialliance.org/

DDI Alliance. (2019). Moving Forward Project (DDI4). Retrieved May 6, 2019, from https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/491703/Moving+Forward+Project+DDI4

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLOS ONE*, *13*(5), e0194768. https://doi.org/10.1371/journal.pone.0194768

Fortier, I., Raina, P., Van den Heuvel, E. R., Griffith, L. E., Craig, C., Saliba, M., … Ferretti, V. (2017). Maelstrom Research guidelines for rigorous retrospective data harmonization. *International Journal of Epidemiology*, *46*(1), 103–105.

Geubbels, E., Amri, S., Levira, F., Schellenberg, J., Masanja, H., & Nathan, R. (2015). Health & demographic surveillance system profile: the Ifakara rural and urban health and demographic surveillance system (Ifakara HDSS). *International Journal of Epidemiology*, *44*(3), 848–861.

Gregson, S., Mugurungi, O., Eaton, J., Takaruza, A., Rhead, R., Maswera, R., … Schaefer, R. (2017). Documenting and explaining the HIV decline in east Zimbabwe: the Manicaland General Population Cohort. *BMJ Open*, *7*(10), e015898.

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.

IHSN. (2012). *IHSN technical note on metadata standards - DRAFT*. Retrieved from International Household Survey Network website: www.ihsn.org/sites/default/files/resources/DDI_SDMX_IHSN_DRAFT.pdf

International Household Survey Network. (2016). Microdata Cataloging Tool (NADA) | IHSN. Retrieved August 3, 2016, from http://www.ihsn.org/home/software/nada

Kahn, K., Collinson, M. A., Gómez-Olivé, F. X., Mokoena, O., Twine, R., Mee, P., … Khosa, A. (2012). Profile: Agincourt health and socio-demographic surveillance system. *International Journal of Epidemiology*, *41*(4), 988–1001.

Kanjala, C., Todd, J., Beckles, D., Castillo, T., Knight, G., Mtenga, B., … Zaba, B. (2017). Open-access for existing LMIC demographic surveillance data using DDI. *IASSIST Quarterly*, *40*(2), 18–18.

Kishamawe, C., Isingo, R., Mtenga, B., Zaba, B., Todd, J., Clark, B., … Urassa, M. (2015). Health & Demographic Surveillance System Profile: The Magu Health and Demographic Surveillance System (Magu HDSS). *International Journal of Epidemiology*, dyv188.

Odhiambo, F. O., Laserson, K. F., Sewe, M., Hamel, M. J., Feikin, D. R., Adazu, K., … Bayoh, N. (2012). Profile: the KEMRI/CDC health and demographic surveillance system—Western Kenya. *International Journal of Epidemiology*, *41*(4), 977–987.

O'Neill, D., Benzeval, M., Boyd, A., Calderwood, L., Cooper, C., Corti, L., … Park, A. (2019). Data Resource Profile: Cohort and Longitudinal Studies Enhancement Resources (CLOSER). *International Journal of Epidemiology*. https://doi.org/10.1093/ije/dyz004

Pisani, E., Aaby, P., Breugelmans, J. G., Carr, D., Groves, T., Helinski, M., … Marsh, V. (2016). Beyond open data: realising the health benefits of sharing data. *Bmj*, *355*, i5295.

Ryssevik, J. (1999). *PROVIDING GLOBAL ACCESS TO DISTRIBUTED DATA THROUGH METADATA*. 22–24. Geneva, Switzerland.

SDMX. (2018, October). Learning | SDMX – Statistical Data and Metadata eXchange. Retrieved August 2, 2019, from SDMX website: https://sdmx.org/?page_id=2555

SDMX Technical Working Group. (2018, April). *VTL – version 2.0 (Validation and Transformation Language) Part 1 - User Manual*. Retrieved from https://sdmx.org/wp-content/uploads/VTL-2.0-User-Manual-20180416-final.pdf

StataCorp. (2019). Stata Statistical Software (Version 16). Retrieved from www.stata.com

Tanser, F., Hosegood, V., Bärnighausen, T., Herbst, K., Nyirenda, M., Muhwava, W., … Newell, M.-L. (2008). Cohort Profile: Africa centre demographic information system (ACDIS) and population-based HIV survey. *International Journal of Epidemiology*, *37*(5), 956–962.

Thérèse Lalor, & Steven Vale. (2013). *Modernising the Production of Official Statistics*. *4*(2), 72–79.

UNECE. (2018, August). GSBPM v5.0 - Generic Statistical Business Process Model - UNECE Statswiki. Retrieved February 6, 2019, from https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0

UNECE Secretariat. (2009). *Generic Statistical Business Process Model: Version 4.0–April 2009*.

Walport, M., & Brest, P. (2011). Sharing research data to improve public health. *The Lancet*, *377*(9765), 537–539.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Bourne, P. E. (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, … Wolfgang Zenk-M¨oltgen. (2012). *Developing a Model-Driven DDI Specification* [Workshop Paper]. Retrieved from Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH website: www.ddialliance.org/system/files/DevelopingaModel-DrivenDDISpecification2013_05_15.pdf

Winters, K., & Netscher, S. (2016). Proposed standards for variable harmonization documentation and referencing: a case study using QuickCharmStats 1.1. *PloS One*, *11*(2), e0147795.