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

Study protocol

Version: 27 January, 2019

Principle investigators:

Mr Chifundo Kanjala
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine
E-mail: chifundo.kanjala@lshtm.ac.uk

Prof Jim Todd
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine
E-mail: jim.todd@lshtm.ac.uk

Dr Emma Slaymaker
Faculty of Epidemiology and Population Health
London School of Hygiene and Tropical Medicine
E-mail:emma.slaymaker@lshtm.ac.uk

Dr Jay Greenfield
Independent Consultant, Data and Metadata architect, USA

Mr Arofan Gregory
Independent Consultant, Metadata and software developer, USA

Contents

1. Introduction

The network for Analysing Longitudinal Population-based HIV/AIDS data on Africa (ALPHA), a collaboration of ten research institutions conducting population-based HIV surveillance in eastern and southern Africa, is working on a project to produce a set of harmonised data that combines both population-based and clinic data from the member studies with funding from the Wellcome Trust. Whilst community-based cohorts and demographic surveillance systems provide a rich source of data, use of the data is often limited because successful analysis requires detailed knowledge of the study's contemporary and historical procedures and of data management practices. To date the ALPHA Network has successfully extracted and harmonised 10 standard data tables from the member studies. However, these data are still complex and require considerable prior knowledge to use effectively, which in practice means the data can only be used in collaboration with one of the ALPHA staff. The main project combines three sets of activities: Using industry standard data integration methods, and a bespoke data appliance Centre in a Box - CiB (Herbst et al., 2015) to develop a robust process for deriving the ALPHA datasets. The second theme is on the integration of the existing ALPHA clinical dataset with data contributed to the IeDEA Network (which links HIV clinical cohorts). The third set of activities relates to documentation of the data. High-quality documentation of both the data and the processes used to derive these is fundamental to the success of the ALPHA main project. The CiB provides tool-specific process metadata. Generic standards-based data documentation will be developed on top of this tool-specific metadata, so that the data becomes immediately accessible to users independent of knowledge of the production tools.

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

1.1. Aims and objectives

This study aims to produce a requirements specification document for the ALPHA data documentation software from eliciting, analysing and synthesising requirements from experts in ALPHA and the CLOSER (Cohort & Longitudinal Studies Enhancement Resources) project (https://www.closer.ac.uk/).

Specific objectives include:

i. To gather a list of features considered to be important for the ALPHA data documentation software from domain experts working in ALPHA and the CLOSER project
ii. To elicit views of the domain experts working in ALPHA and the CLOSER project on the ALPHA data documentation software features provided for in mock-ups created by the ALPHA developers.
iii. To synthesise the elicited functional requirements and views on the features in the mock-ups into a requirements specification document

2. Methods

2.1. Recruitment

An overview of the study will be given at a workshop where potential participants will be present. A convenience sample will then be drawn from the professionals directly involved in the ALPHA data production process or analysis process from the member studies and secretariat.

An archivist and developer from the CLOSER project, who are known to the applicants, will draw a convenience sample of their colleagues who are data programmers or data scientists and archivists involved in receiving harmonising and preparing CLOSER data for dissemination.

2.2. Data collection

Data collection will involve, data scientists, epidemiologists, demographers and other researchers who produce and or use the ALPHA datasets. For the CLOSER project, the study will involve data scientists/ programmers, data archivist and developers who are involved in the process of receiving data from the various cohort studies, harmonising or otherwise preparing them for dissemination.

The data collection will consist of:

Background materials reading and note taking prior to interview: Prior to the interview, background materials will be sent to the potential interviewee to familiarise them with the planned project and the process of requirements elicitation. The interviewees will be encouraged to read through and make notes as they go and list any questions they may have. The annotated background material will be collected prior to the interview.

Skype interview: Each participant will be interviewed over skype first asking them open ended questions on their desired software features for documenting the ALPHA datasets. Subsequently, they will be asked to review 6 mock-ups of possible features of the metadata software. After the review, they will grade each feature as either a useful feature or not on a

scale as provided in the questionnaire. They will also provide comments on the features as they see fit.

2.3. Target population and Sample size

The 6 – 8 interviewees will be purposively selected from ALPHA domain experts and 5 – 7 participants will come from the CLOSER project. The ALPHA participants, as the producers and or internal users of the harmonised data, they will provide the viewpoint of users who are sensitive to the specifics of the data harmonisation process Analysis. CLOSER project staff will provide the viewpoint of archivists and developers who are familiar with the metadata standards and with data integration and harmonisation (they have successfully conducted an ongoing data harmonisation project involving eight UK birth cohorts). Between these two groups of users, we feel that we will be able to identify the requirements of both internal and external users.

2.4. Data Analysis

The interviews will be transcribed and analysed using NVivo software. Analyses will include identification of functionalities that interviewees specify as essential, those they do not find useful, and those not covered but considered to be important. Features existing in the mock-ups and indicated by interviewees as important will be maintained in the requirements specification document. The features in the mock-ups considered not very useful will be revised as needed and either added to the requirements specification or discarded as appropriate. The features that interviewees consider important but are not present in the mock-ups will be prioritised according to their perceived value from interviewee responses and incorporated as possible within the requirements specification document.

2.5. Dissemination of results

The requirements specification document resulting from this study will be made available to the ALPHA software developers and the entire network members, CLOSER project, and other stakeholders for use as a basis for the software evaluation survey to be done when the development is completed.

3. Human subjects' protection

3.1. Consent and risk to participants

The study protocol will be submitted for review to the appropriate ethics review board at the London School of Hygiene & Tropical Medicine.
There is no direct risk to the participants.
Participants will be informed that their participation is voluntary and that they are free to discontinue their participation at any time. The results of this will be available to the participants in form of a software requirements specification. Further we will ask participants whether they consent to having their anonymised responses shared for research purposes.

3.2. Data handling

We recognise that ensuring participant confidentiality is important for this study, therefore all data collected on study participants will be secured (locked if paper or password-protected/encrypted if electronic). An anonymised version of the raw database, stripped of identifying information like name, address, etc. will be created, and only the anonymised one will be shared with researchers on LSHTM institutional repository - Data Compass.
Only select members of the data team will have access to the raw data including the personal identifiers, for the purpose of entering the information into an electronic database, checking for errors, and producing the anonymised version. Access to the anonymised dataset will still be restricted to individuals who are actively involved in the study. The electronic database will be password protected on a secure server at the ALPHA secretariat at the London School of Hygiene & Tropical Medicine.

4. Study team

4.1. London School of Hygiene and Tropical Medicine

Chifundo Kanjala, Data documentalist, ALPHA network, Department of Population Health
Prof Jim Todd, Applied Biostatistician, ALPHA network PI, Department of Population Health
Dr Emma Slaymaker, Epidemiologist/ Demographer, ALPHA network PI, Department of Health

4.2. Collaborators

Dr Jay Greenfield, Independent Consultant Data/ Metadata Architect, New Hampshire, USA
Mr Arofan Gregory, Independent Consultant, Metadata/ software developer, New Hampshire, USA

References

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

Reniers, G., Wamukoya, M., Urassa, M., Nyaguara, A., Nakiyingi-Miiro, J., Lutalo, T., … Geubbels, E. (2016). Data Resource Profile: Network for Analysing Longitudinal Population-based HIV/AIDS data on Africa (ALPHA Network). *International Journal of Epidemiology*, *45*(1), 83–93.

Appendices

Appendix 1: Information sheet and consent form (to be translated)

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

PARTICIPANT INFORMATION SHEET

My name is Chifundo Kanjala and I am a PhD student in the Department of Population Health, London School of Hygiene and Tropical Medicine. We are carrying out a study to elicit functional requirements for a software system for browsing, searching and constructing data lineages relating to the ALPHA datasets.
You are being invited to take part in a research study. Before you decide to take part, it is important for you to understand why the research is being done and what it will involve. I will read information to you about this study. Please ask me if there is anything that is not clear or if you would like more information.

WHAT ARE WE TRYING TO LEARN WITH THIS RESEARCH STUDY?
We would like to understand what the business requirements are for a data documentation software for ALPHA datasets from the perspectives of domain experts in ALPHA and the CLOSER project.
The full utility of structured data documentation is realised when tools are available to browse, search and explore those metadata. The functionalities of such a tool for ALPHA datasets documentation is currently not fully understood. It is important to understand these requirements from the perspective the potential users and experts in the area of research data harmonisation and dissemination. By interviewing domain experts, we can gather and analyse their views. We hope that the findings from this study will help improve our understanding of the desired functional requirements for the said tools.

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

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

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

WHAT WILL MY PARTICIPATION IN THIS STUDY INVOLVE?

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

ARE THERE ANY RISKS INVOLVED WITH PARTICIPATING IN THIS STUDY?

There are no direct risks from participating.

ARE THERE ANY BENEFITS INVOLVED WITH PARTICIPATING IN THIS STUDY?

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

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

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

WHO WILL SEE THE INFORMATION THAT IS COLLECTED?

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

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

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

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

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

Appendix 2: Consent form for Participant

CONSENT FORM

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

Name of PI/Researcher responsible for project: Chifundo Kanjala

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

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