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

BACKGROUND INFORMATION

What is ALPHA?

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

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

ALPHA data and "modus operandi"

ALPHA assembles datasets on various topics related to demographic and HIV surveillance. These data are referred to as ALPHA data specifications or data specs and are described here[1]. The ALPHA data specs have a well-defined structure to which each partner of the network has to transform their data. ALPHA is organised around data analysis and HIV research capacity strengthening workshops. At the workshops, partners bring their data and are involved in data analysis training addressing research questions of interest for the particular workshop.

Data harmonisation in ALPHA

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

Whilst community-based cohorts and demographic surveillance systems provide a rich source of data, use of the data is often limited because successful analysis requires detailed

[1] http://alpha.lshtm.ac.uk/metadata/

knowledge of the study's contemporary and historical procedures and of data management practices. To date the ALPHA Network has successfully extracted and harmonised 10 standard data tables from the partner studies. However, these data are still complex and require considerable prior knowledge to use effectively, which in practice means the data can only be used in collaboration with one of the ALPHA staff.

The main project combines three sets of activities:

(1) Using industry standard data integration methods, and a bespoke data appliance Centre in a Box - CiB (Herbst et al., 2015) to develop a robust process for deriving the ALPHA datasets.

(2) Integration of the existing ALPHA clinical dataset with data contributed to the IeDEA Network (which links HIV clinical cohorts).

(3) High-quality documentation of both the data and the processes used to derive the data.

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

Mock-ups

Included in the information pack is a set of mock-up diagrams showing features that the developers have proposed as a starting point for discussion. Please note that these mock-ups are not a reflection of what the software interface will look like, they only show the features of the system. Please have a look at these before the interview as the interview questions will seek to elicit your views about the proposed features.

Terms used in mock-ups

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

Why ALPHA network interviewees?

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

Why CLOSER project interviewees?

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

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

Requirements overview

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

Types of Requirements

A requirement can be:

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

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

References

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

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