Data Documentation Template

1 Study Description

1.1 Identification

1.1.1 Title

Contains the full authoritative title of the data collection. A full title should indicate the geographic scope of the data collection as well as the time period covered.

Visualisation of sequence and demographic data to assist HIV surveillance in northern KwaZulu-Natal: extending the TasP/iSense dashboard to include markers of HIV drug resistance mutations.

1.2 Overview

1.2.1 Abstract

An unformatted summary describing the purpose, nature, and scope of the data collection, special characteristics of its contents, major subject areas covered, and what questions the PIs attempted to answer when they conducted the study. A listing of major variables in the study is important here.

This proposal aims to extend an existing collaboration between AHRI and UCL. As part of the iSense project, teams from AHRI and UCL have successfully developed a dashboard that integrates information from mobile computers used for TasP field visits with data from the clinics to display spatial coverage of homestead visits, highlighting those that require follow-up visits to ensure linkage to care. The dashboard provides a broad snapshot of the state of the study, spatially aggregating geographical zones in order to preserve the privacy of trial participants.

The aim of this proposal is to extend this framework to visualise presence and prevalence of drug resistance mutations (DRMs) within the study area. A higher prevalence of DRMs than expected may be linked to several factors, e.g. poor drug adherence, and thus of value to clinicians and healthcare workers in terms of focusing efforts and resource allocation.

1.2.2 Kind of Data

The type of data included in the file: survey data, census/enumeration data, aggregate data, clinical data, event/transaction data, program source code, machine-readable text, administrative records data, experimental data, psychological test, textual data, coded textual, coded documents, time budget diaries, observation data/ratings, process-produced data, etc. No description just a single phrase, e.g. Genetic sequences

The existing dashboard is built on pseudonymised demographic data grouped into geographic hexagons. Within the PANGEA project, we have independently assembled HIV genomes from patient clinical samples and identified DRMs for each sample. To map the DRMs to each geographical hexagon, we require a lookup table to link the TasP/PANGEA IDs to hexagons.

1.2.3 Unit of Analysis

Basic unit(s) of analysis or observation that the study describes: For PANGEA is each record a sequence from a specimen, or are there multiple records for a single specimen or study participant

Each sample ID should be linked to a single hexagon where possible.

1.3 Scope

1.3.1 Topics Classification

MeSH subject headings

HIV-1

Incidence

Phylogeny

Epidemics

Population Surveillance

Rural Population

HIV Infections

Africa

1.3.2 Keywords

Keywords summarize the content or subject matter of the survey. As topic classifications, these are used to facilitate referencing and searches in electronic survey catalogues.

Mapping, visualisation, data linkage, drug resistance, DRM, HIV-1

1.4 Coverage

1.4.1 Country

Indicates the country or countries covered in the file

South Africa

1.4.2 Geographic Coverage

Information on the geographic coverage of the data. Include the total geographic scope of the data, and any additional levels of geographic coding provided in the variables.

Demographic surveillance area of the Africa Health Research Institute's TasP study.

1.4.3 Universe

A description of the population covered by the data in the file; the group of persons or other elements that are the object of the study and to which the study results refer. Age, nationality, and residence commonly help to delineate a given universe, but any of a number of factors may be involved, such as age limits, sex, marital status, race, ethnic group, etc. The universe may consist of elements other than persons, such as specimen, sample or isolate. In general, it should be possible to tell from the description of the universe whether a given individual or element (hypothetical or real) is a member of the population under study. Also known as universe of interest, population of interest, and target population.

1.5 Producers and Sponsors

1.5.1 Investigators

The person, corporate body, or agency responsible for the data collection's substantive and intellectual content. Repeat the element for each author and use the affiliation attribute if available. Invert first and last name and use commas. Remarks: The author in this element should be the individual(s) or organization(s) directly responsible for the intellectual content of the data collection.

Name	Affiliation
Frampton, Dan	Division of Infection and Immunity, UCL, London

McKendry, Rachel	London Centre for Nanotechnology, UCL, London
Herbst, Kobus	Africa Health Research Institute
Pillay, Deenan	Africa Health Research Institute

1.5.2 Funding

The source(s) of funds for production of the data collection. If different funding agencies sponsored different stages of the production process, use the role attribute to distinguish them. Also includes a field for the grant/contract number of the project that sponsored the data collection effort.

Agency	Abbreviation	Grant number	Role
South African Medical Research Council	SAMRC	MRC-RFA-UFSP-01/2013/UKZN HIVEPI	Genotyping funding source
Engineering and Physical Sciences Research Council	EPSRC	EP/R00529X/1	Design and implementation of dashboard within the iSense project

1.5.3 Acknowledgements

Statements of responsibility not recorded in the title and statement of responsibility areas. Indicate here the persons or bodies connected with the work, or significant persons or bodies connected with previous editions and not already named in the description. For example, the name of the person who cleaned the data collection might be cited here, using the role and affiliation attributes. Does not include funders.

Name	Affiliation	Role
Jaco Dreyer	Africa Health Research Institute,	Linking sample and demographic data
Ed Manley	UCL Centre for Advanced Spatial Analysis (CASA), London	Design and implementation of iSense dashboard
Dave Concannon	UCL Centre for Advanced Spatial Analysis (CASA), London	Design and implementation of iSense dashboard

1.6 Sampling

1.6.1 Sampling Procedure

The type of sample and sample design used to select the survey respondents to represent the population. May include reference to the target sample size and the sampling fraction

HIV positive individuals within the TasP surveillance area of the Africa Health Research Insititute from 2012 to 2016.

1.7 Data Collection

1.7.1 Dates of Collection

Contains the date(s) when the data were collected/produced.

Between the start of 2012 and the end of 2016.

1.7.2 Notes on Data Collection/Production

Used to describe noteworthy aspects of the data collection/production.

Not applicable.

1.8 Data Processing

1.8.1 Other Processing

Used to indicate additional information about the methodology and processing involved in the production of the dataset.

Samples were sequenced at the Durban based laboratory of AHRI; genome assembly and downstream sequence analysis was performed at UCL.

1.9 Data Access

1.10 Contacts

1.10.1 Contact persons

Individuals listed as contact persons will be used as resource persons regarding problems or questions raised by the user community. The URI attribute should be used to indicate a URN or URL for the homepage of the contact individual. The email attribute is used to indicate an email address for the contact individual.

Name	Affiliation	Email	URI
Frampton, Dan	Division of Infection and Immunity, UCL	d.frampton@ucl.ac.uk	www.i-sense.org.uk

2 File Description

2.1 Data Files

2.1.1 Contents

Abstract or description of the file. A summary describing the purpose, nature, and scope of the data file, special characteristics of its contents, major subject areas covered, and what questions the PIs attempted to answer when they created the file. No need to repeat information already contained in the abstract in paragraph 1.2.1.

See section 1.2.1

3 Variable Description

A code book of the variables in the data file.

Name	Definition	Data Type and Codes
AC	Africa Centre (AHRI)	Community level location of sampling
HexagonID	Identifier corresponding to a aggregated geographical area within the TasP study	Integer as previously used for the iSense dashboard
SampleID	Unique sample ID with the TasP study	TASP123456, TASPX123456
Date	Date of sampling in decimal format	e.g. 2010.123