# Title:

|  |
| --- |
|  |

The title is the official name of the survey as it is stated on the questionnaire or as it appears in the design documents. The following items should be noted:

- Include the reference year(s) of the survey in the title.

- Do not include the abbreviation of the survey name in the title.

- As the survey title is a proper noun, the first letter of each word should be capitalized (except for prepositions or other conjunctions).

- Including the country name in the title is optional.

**Examples:**

- Palau Household Budget Survey 2002-2003

- Vanuatu Multiple Indicator Cluster Survey 2002

# Abbreviation:

|  |
| --- |
|  |

The abbreviation of a survey is usually the first letter of each word of the titled survey. The survey reference year(s) may be included.

**Examples:**

- DHS 2000 for "Demographic and Health Survey 2005"

- HIES 2002-2003 for "Household Income and Expenditure Survey 2003"

# Study Type:

|  |
| --- |
| * Administrative records (“Health”, “Education” or “Other”) * Agriculture Census * Agriculture Survey * Child Labour Survey * Core Welfare Indicators Questionnaire * Demographic and Health Survey * Enterprise Survey * Enterprise Census * Income / Expenditure / Household Survey * Informal Sector Survey * Integrated Survey * Multiple Indicator Cluster Survey * Labour Force Survey * Living Standards Measurement Study * Other Household Health Survey * Other Household Survey * Price Survey * Priority Survey * Population and Household Census * Sample Frame, Households * Sample Frame, Enterprises * Service Provision Assessments * Socio-Economic / Monitoring Survey |

The study type or survey type is the broad category defining the survey.

This item has a controlled vocabulary (you may customize the IHSN template to adjust this controlled vocabulary *if needed)*

*Cover Pict: Toga Raikoti, SPC, 2018*

# Series Information:

|  |
| --- |
|  |

A survey may be repeated at regular intervals (such as an annual labour force survey), or be part of an international survey program (such as the MICS, CWIQ, DHS, LSMS and others). The Series information is a description of this "collection" of surveys. A brief description of the characteristics of the survey, including when it started, how many rounds were already implemented, and who is in charge would be provided here. If the survey does not belong to a series, leave this field empty.

# ID Information:

|  |
| --- |
|  |

The ID number of a dataset is a unique number that is used to identify a particular survey. Define and use a consistent scheme to use. Such an ID could be constructed as follows, and in capital letters: Producer\_country\_survey\_year\_version where:

- producer is the abbreviation of the producing agency

- country is the 3-letter ISO country abbreviation

- year is the reference year (or the year the survey started)

- survey is the survey abbreviation (PHC: census, EUS: Employment and Unemployment Survey…)

- version is the documentation version number (first attempt is “v01”, adaptation of the v01 is going to be “v02”…etc)

- “M” if it’s an adaptation of the Master file, or “PUF” if it’s a Public-use file.

**Example:**

The Demographic and Health Survey implemented by the Uganda Bureau of Statistics in 2005 could have the following ID: FIBOS\_FJI\_2008\_HIES\_v01\_M

# Version Description:

|  |
| --- |
|  |

The version description should contain a version number followed by a version label. The version number should follow a standard convention to be adopted by the institute. We recommend using “v01” (which is the version of your documentation) followed by the type of dataset used:

**Examples:**

- v01: Basic raw data, obtained from data entry (before editing).

- v01: Edited data, second version, for internal use only.

- v01: Edited, anonymous dataset for public distribution.

A brief description of the version should follow the numerical identification.

# Date of Production:

|  |
| --- |
|  |

This is the date in ISO format (yyyy-mm-dd) of actual and final production of the data. Production dates of all versions should be carefully tracked. Provide at least the month and year. Use the calendar icon in the Metadata editor to assure that the date selected is in compliance with the ISO format.

# Kind of data:

|  |
| --- |
| * Sample survey data * Census / Enumeration data * Administrative records data * Aggregated data * Clinical data * Event / Transaction data |

This field is a broad classification of the data and it is associated with a drop-down box providing controlled vocabulary. That controlled vocabulary includes 9 items but is not limited to them.

# Unit of analysis:

|  |
| --- |
|  |

Basic unit(s) of analysis or observation that the study describes: individuals, families/households, groups, facilities, institutions/organizations, administrative units, physical locations, etc.

**Examples**:

- A living standards survey with community-level questionnaire would have the following units of analysis: individuals, households, and communities. An economic survey could have the firm and establishment as units of analysis.

# Description of scope:

|  |
| --- |
|  |

The scope is a description of the themes covered by the survey. It can be viewed as a summary of the modules that are included in the questionnaire (education, economic activities, health…). The scope does not deal with geographic coverage.

# Country:

|  |  |
| --- | --- |
| **Country name** | **Abbreviation** |
|  |  |

Enter the country name, even in cases where the survey did not cover the entire country. In the field "Abbreviation", we recommend that you enter the 3-letter ISO code of the country. If the dataset you document covers more than one country, enter all in separate rows.

# Geographic Coverage:

|  |
| --- |
|  |

This filed aims at describing at what geographic level the data are representative. Typical entries will be "National coverage", "Urban (or rural) areas only", "State of ...", "Capital city", etc.

*Note that we do not describe here where the data was collected. For example, as sample survey could be declared as "national coverage" even in cases where some districts where not included in the sample, as long as the sampling strategy was such that the representativity is national.*

# Universe:

|  |
| --- |
|  |

We are interested here in the survey universe (not the universe of particular sections of the questionnaires or variables), i.e. in the identification of the population of interest in the survey. The universe will rarely be the entire population of the country. Sample household surveys, for example, usually do not cover homeless, nomads, diplomats, community households. Population censuses do not cover diplomats. Try to provide the most detailed information possible on the population covered by the survey/census.

# Primary Investigator:

|  |
| --- |
|  |

The primary investigator will in most cases be an institution, but could also be an individual in the case of small-scale academic surveys. The two fields to be completed are the Name and the Affiliation fields.

Generally, in a survey, the Primary Investigator will be the institution implementing the survey. If various institutions have been equally involved as main investigators, then all should be mentioned. This only includes the agencies responsible for the implementation of the survey, not its funding or technical assistance.

The order in which they are listed is discretionary. It can be alphabetic or by significance of contribution. Individual persons can also be mentioned. If persons are mentioned use the appropriate format of Surname, First name.

# Funding:

|  |
| --- |
|  |

List the organizations (national or international) that have contributed, in cash or in kind, to the financing of the survey.

The government institution that has provided funding should not be forgotten.

# Other acknowledgements:

|  |
| --- |
|  |

This optional field can be used to acknowledge any other people and institutions that have in some form contributed to the survey.

# Sampling:

|  |
| --- |
|  |

This item should document the design and definition of the sample size, including: sampling frame, sampling type and final size of the sample, as well as sample loss, estimation method and accurate calculation of the results. This Item is only applicable for sample surveys.

# Sample design and size definition:

|  |
| --- |
|  |

This element provides information on the sampling frame and the methods and procedures used to select respondents. The desired sample size should also be mentioned. The titles of each of the themes should be capitalized Major deviations from the sample design: this element is used to describe the correspondence between the units that were successfully surveyed and the planned sample. Any significant deviation should be mentioned here. Estimation procedure: it explains the estimators that are used to obtain indicators of the statistical operation. It defines and justifies the selected methodology and its components. It indicates the calculation of the expansion factors and elements that determine the expansion. Each one of the titles used in the documentation of this field should be capitalized. Calculation of the accuracy of results: it describes the methodology used to estimate sampling errors and their presentation, in order to determine the level of confidence. It reviews the design of the variance estimation method. It presents the formulas for calculating the standard error and / or the coefficient of variation of the estimators. Subtitles used in the documentation of this field should be capitalized.

# Sampling Procedure

|  |
| --- |
|  |

This field only applies to sample surveys. Information on sampling procedure is crucial (although not applicable for censuses and administrative datasets). This section should include summary information that includes though is not limited to:

- Sample size

- Selection process (e.g., probability proportional to size or over sampling)

- Stratification (implicit and explicit)

- Stages of sample selection

- Design omissions in the sample

- Level of representation

- Strategy for absent respondents/not found/refusals (replacement or not)

- Sample frame used, and listing exercise conducted to update it

It is useful also to indicate here what variables in the data files identify the various levels of stratification and the primary sample unit. These are crucial to the data users who want to properly account for the sampling design in their analyses and calculations of sampling errors.

This section accepts only text format; formulae cannot be entered. In most cases, technical documents will exist that describe the sampling strategy in detail. In such cases, include here a reference (title/author/date) to this document, and make sure that the document is provided in the External Resources.

**Example:**

5,000 households were selected for the sample. Of these, 4,996 were occupied households and 4,811 were successfully interviewed for a response rate of 96.3%. Within these households, 7,815 eligible women aged 15-49 were identified for interview, of which 7,505 were successfully interviewed (response rate 96.0%), and 3,242 children aged 0-4 were identified for whom the mother or caretaker was successfully interviewed for 3,167 children (response rate 97.7%). These give overall response rates (household response rate times individual response rate) for the women's interview of 92.5% and for the children's interview of 94.1%.

# Deviation from Sample Design

|  |
| --- |
|  |

This field only applies to sample surveys.

Sometimes the reality of the field requires a deviation from the sampling design (for example due to difficulty to access to zones due to weather problems, political instability, etc). If for any reason, the sample design has deviated, this should be reported here.

# Response rate

|  |
| --- |
|  |

Response rate provides that percentage of households (or other sample unit) that participated in the survey based on the original sample size. Omissions may occur due to refusal to participate, impossibility to locate the respondent, or other. Sometimes, a household may be replaced by another by design. Check that the information provided here is consistent with the sample size indicated in the "Sampling procedure" field and the number of records found in the dataset (for example, if the sample design mention a sample of 5,000 households and the data on contain data on 4,500 households, the response rate should not be 100 percent). Provide if possible the response rates by stratum. If information is available on the causes of non-response (refusal/not found/other), provide this information as well.

# Weighting

|  |
| --- |
|  |

This field only applies to sample surveys.

Provide here the list of variables used as weighting coefficient. If more than one variable is a weighting variable, describe how these variables differ from each other and what the purpose of each one of them is.

**Example:**

Sample weights were calculated for each of the data files.

Sample weights for the household data were computed as the inverse of the probability of selection of the household, computed at the sampling domain level (urban/rural within each region). The household weights were adjusted for non-response at the domain level, and were then normalized by a constant factor so that the total weighted number of households equals the total unweighted number of households. The household weight variable is called HHWEIGHT and is used with the HH data and the HL data.

Sample weights for the women's data used the un-normalized household weights, adjusted for non-response for the women's questionnaire, and were then normalized by a constant factor so that the total weighted number of women's cases equals the total unweighted number of women's cases.

Sample weights for the children's data followed the same approach as the women's and used the un-normalized household weights, adjusted for non-response for the children's questionnaire, and were then normalized by a constant factor so that the total weighted number of children's cases equals the total unweighted number of children's cases.

# Dates of collection:

|  |
| --- |
|  |

Enter the dates (at least month and year) of the start and end of the data collection.

DATE MUST BE ENTERED IN THE ISO FORMAT YYYY-MM-DD

# Time period:

|  |
| --- |
|  |

This field will usually be left empty. Time period differs from the dates of collection as they represent the period for which the data collected are applicable or relevant.

NOTE: DATE MUST BE ENTERED IN THE FORMAT YYYY-MM-DD

# Mode of collection:

|  |
| --- |
| * CAPI (Computer Assisted Personal Interview) * CATI (Computer Assisted Telephone Interview) * Face to face * Mail questionnaire * Focus Group * Internet * Other |

The mode of data collection is the manner in which the interview was conducted or information was gathered. This field is a controlled vocabulary field.

# Notes on Data Collection:

|  |
| --- |
|  |

This element is provided in order to document any specific observations, occurrences or events during data collection. Consider stating such items like:

- Was a training of enumerators held? (elaborate)

- Any events that could have a bearing on the data quality?

- How long did an interview take on average?

- Was there a process of negotiation between households, the community and the implementing agency?

- Are anecdotal events recorded?

- Have the field teams contributed by supplying information on issues and occurrences during data collection?

- In what language was the interview conducted?

- Was a pilot survey conducted?

- Were there any corrective actions taken by management when problems occurred in the field?

# Questionnaires:

|  |
| --- |
|  |

This element is provided to describe the questionnaire(s) used for the data collection. The following should be mentioned:

- List of questionnaires and short description of each (all questionnaires must be provided as External Resources)

- In what language were the questionnaires published?

- Information on the questionnaire design process (based on a previous questionnaire, based on a standard model questionnaire, review by stakeholders). If a document was compiled that contains the comments provided by the stakeholders on the draft questionnaire, or a report prepared on the questionnaire testing, a reference to these documents should be provided here and the documents should be provided as External Resources.

# Data Collectors:

|  |
| --- |
|  |

This element is provided in order to record information regarding the persons and/or agencies that took charge of the data collection. This element includes 3 fields: Name, Abbreviation and the Affiliation. In most cases, we will record here the name of the agency, not the name of interviewers. Only in the case of very small-scale surveys, with a very limited number of interviewers, the name of person will be included as well. The field Affiliation is optional and not relevant in all cases.

# Supervision:

|  |
| --- |
|  |

This element will provide information on the oversight of the data collection. The following should be considered:

- Were the enumerators organized in teams that included a controller and a supervisor? With how many controllers/supervisors per interviewer?

- What were the main roles of the controllers/supervisors?

- Were there visits to the field by upper management? How often?

# Data Editing:

|  |
| --- |
|  |

The data editing should contain information on how the data was treated or controlled for in terms of consistency and coherence. This item does not concern the data entry phase but only the editing of data whether manual or automatic.

- Was a hot deck or a cold deck technique used to edit the data?

- Were corrections made automatically (by program), or by visual control of the questionnaire?

- What software was used?

If materials are available (specifications for data editing, report on data editing, programs used for data editing), they should be listed here and provided as external resources.

# Other processing:

|  |
| --- |
|  |

Use this field to provide as much information as possible on the data entry design. This includes such details as:

- Mode of data entry (manual or by scanning, in the field/in regions/at headquarters)

- Computer architecture (laptop computers in the field, desktop computers, scanners, PDA, other; indicate the number of computers used)

- Software used

- Use (and rate) of double data entry

Information on tabulation and analysis can also be provided here.

All available materials (data entry/tabulation/analysis programs; reports on data entry) should be listed here and provided as external resources.

# Estimates of Sampling Error:

|  |
| --- |
|  |

For sampling surveys, it is good practice to calculate and publish sampling error. This field is used to provide information on these calculations. This includes:

- A list of ratios/indicators for which sampling errors were computed.

- Details regarding the software used for computing the sampling error, and reference to the programs used (to be provided as external resources) as the program used to perform the calculations.

- Reference to the reports or other document where the results can be found (to be provided as external resources).

# Other forms of data validation:

|  |
| --- |
|  |

This section can be used to report any other action taken to assess the reliability of the data, or any observations regarding data quality. This item can include:

- For a population census, information on the post enumeration survey (a report should be provided in external resources and mentioned here).

- For any survey/census, a comparison with data from another source.

# Access authority:

|  |
| --- |
|  |

This section is composed of various sections: Name-Affiliation-email-URI. This information provides the contact person or entity to gain authority to access the data. It is advisable to use a generic email contact such as data@popstatsoffice.org <mailto:data@popstatsoffice.org> whenever possible to avoid tying access to a particular individual whose functions may change over time.

# Confidentiality:

|  |
| --- |
|  |

If the dataset is not anonymized, we may indicate here what Affidavit of Confidentiality must be signed before the data can be accessed. Another option is to include this information in the next element (Access conditions). If there is no confidentiality issue, this field can be left blank.

An example of statement could be the following:

Confidentiality of respondents is guaranteed by Articles N to NN of the National Statistics Act of [date].

Before being granted access to the dataset, all users have to formally agree:

1. To make no copies of any files or portions of files to which s/he is granted access except those authorized by the data depositor.

2. Not to use any technique in an attempt to learn the identity of any person, establishment, or sampling unit not identified on public use data files.

3. To hold in strictest confidence the identification of any establishment or individual that may be inadvertently revealed in any documents or discussion, or analysis. Such inadvertent identification revealed in her/his analysis will be immediately brought to the attention of the data depositor.

This statement does not replace a more comprehensive data agreement (see Access condition).

# Citation Requirement:

|  |
| --- |
|  |

Citation requirement is the way that the dataset should be referenced when cited in any publication. Every dataset should have a citation requirement. This will guarantee that the data producer gets proper credit, and that analytical results can be linked to the proper version of the dataset. The Access Policy should explicitly mention the obligation to comply with the citation requirement. The citation should include at least the primary investigator, the name and abbreviation of the dataset, the reference year, and the version number. Include also a website where the data or information on the data is made available by the official data depositor.

**Example:**

"National Statistics Office of Popstan, Multiple Indicators Cluster Survey 2000 (MICS 2000), Version 1.1 of the public use dataset (April 2001), provided by the National Data Archive. www.nda\_popstan.org"

# Disclaimer:

|  |
| --- |
|  |

A disclaimer limits the liability that the Statistics Office has regarding the use of the data. A standard legal statement should be used for all datasets from a same agency. The IHSN recommends the following formulation:

The user of the data acknowledges that the original collector of the data, the authorized distributor of the data, and the relevant funding agency bear no responsibility for use of the data or for interpretations or inferences based upon such uses.

# Copyright:

|  |
| --- |
|  |

Include here a copyright statement on the dataset, such as:

(c) 2007, Popstan Central Statistics Agency

# Contact persons:

|  |
| --- |
|  |

Users of the data may need further clarification and information. This section may include the name-affiliation-email-URI of one or multiple contact persons. Avoid putting the name of individuals. The information provided here should be valid for the long term. It is therefore preferable to identify contact persons by a title. The same applies for the email field. Ideally, a "generic" email address should be provided. It is easy to configure a mail server in such a way that all messages sent to the generic email address would be automatically forwarded to some staff members.