

---

# **Guide for Data Archivists Documentation**

**Olivier Dupriez, Diana Marcela Sanchez Castro, Matthew Welch**

**Mar 13, 2020**



---

## Table of contents

---

<b>1</b>	<b>Content</b>	<b>3</b>
1.1	Introduction . . . . .	3
1.2	Before you start: organizing your files . . . . .	4
1.3	Gathering and preparing the data set . . . . .	5
1.4	Gathering and preparing the documentation . . . . .	20
1.5	Importing data and establishing relationships . . . . .	21
1.6	Importing external resources . . . . .	24
1.7	Completing metadata . . . . .	25
1.8	Creating variable groups . . . . .	56
1.9	Running validations and diagnostics . . . . .	56
1.10	Generating the survey documentation in PDF . . . . .	57
1.11	Independent quality review . . . . .	57
1.12	Section A . Data Validations in Stata: Practical Examples . . . . .	58





## **International Household Survey Network**

### **IHSN**

#### **Version 2019 -04**

**Authors of this guide:** Olivier Dupriez, Diana Marcela Sanchez Castro, Matthew Welch (The World Bank)

The production of this guide was made possible through a grant from the TFSCB - DFID funding to the World Bank P167116/TF0A7461.

**Acknowledgments:** Francois Fonteneau (PARIS21), Geoffrey Greenwell (PARIS21), Chris Rockmore (World Bank) and Jan Smit (ESCAP) provided valuable input to an earlier version of the document. Trevor Croft (UNICEF) provided many of the examples of good practices for completing survey metadata.



## 1.1 Introduction

This *Quick Reference Guide for Data Archivists* provides data archivists with guidelines to document a micro-dataset in compliance with the Data Documentation Initiative (DDI) and the Dublin Core (DCMI) metadata standards<sup>1</sup>, using the World Bank Metadata Editor.

The World Bank Metadata Editor is an application designed to help document data collection operations undertaken for different kinds of research projects. The application is developed as an open-source tool by the World Bank. A number of Metadata standards recognized as global models for defining and describing different types of data have been integrated into the Metadata Editor, these are: The Data Documentation Initiative (DDI Codebook), The Dublin Core Metadata Initiative (DCMI) and the ISO 19139 for geospatial data.

The Metadata Editor is modelled on the Nesstar Publisher. As such it should provide a familiar environment for the Nesstar users. As an added benefit, the Metadata Editor is flexible and can support the documentation of multiple-data types. Out of the box it supports the documentation of survey data, time series data, geospatial data, statistical tables, images, analytical scripts and standalone publications or documents.

This Guide summarizes the process in 10 chronological steps:

1. Gathering and preparing the data set
2. Gathering and preparing the documentation
3. Importing data and establishing relationships
4. Importing external resources
5. Adding metadata
6. Creating variable groups (optional)
7. Running diagnostics
8. Generating the standard survey documentation using the PDF generator

---

<sup>1</sup> DDI (Data Documentation Initiative) and DCMI (Dublin Core Metadata Initiative) are international XML metadata specifications. For more information on these standards and on the IHSN Toolkit, please visit [www.surveynetwork.org](http://www.surveynetwork.org).

9. Quality assessment

10. Producing the output for publication

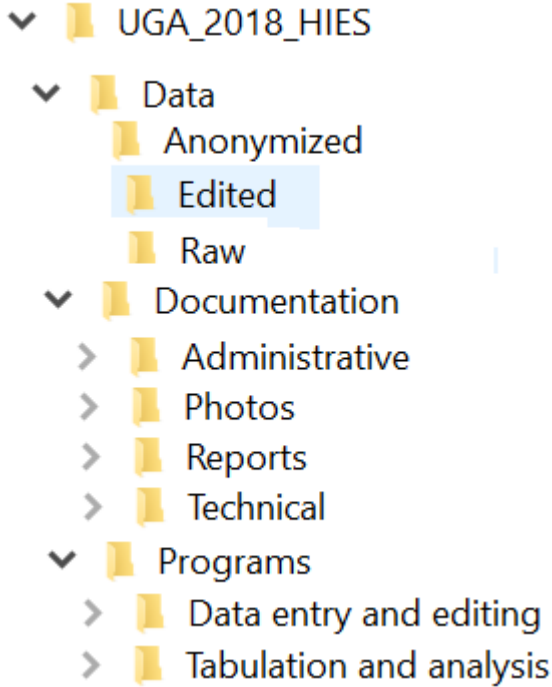
Also provided (in appendix) is the *IHSN DDI Reviewers' Feedback Form* which provides a standard tool for the assessment of survey metadata by an external reviewer.

This Guide is not a Metadata Editor reference or training manual. It is assumed that users are already familiar with the Editor. A *Metadata Editor User's Guide* is available at <https://metadata-editor.readthedocs.io/en/latest/>.

## 1.2 Before you start: organizing your files

Documentation of a dataset will be most efficient if you organize your data and other files properly. We recommend that, before anything else, you create the necessary directories as follows:



	<ul style="list-style-type: none"> <li>• Create a directory for the survey. We suggest you name it using the country, survey's year and the abbreviated name, e.g. "UGA_2018_HIES" for "Household Income and Expenditure Survey" of Uganda collected in 2018.</li> </ul>
	<ul style="list-style-type: none"> <li>• Create various sub-directories for the data files (and for the various versions of the dataset if relevant)</li> </ul>
	<ul style="list-style-type: none"> <li>• Create sub-directories for the documentation and for the program files if relevant (see example).</li> </ul>

### 1.3 Gathering and preparing the data set

Gathering and preparing data is a process that requires great care. Prior to documenting a dataset, it is important to ensure that you are working with the most appropriate version of all the concerned data files. If the dataset is meant for public release, one should work with the final, edited, anonymous version of the dataset. If the dataset is being documented for archiving and internal use only, one may include the raw data as well as the final, fully edited files. The Metadata Editor provides you with the possibility of documenting the specificity of each version of the dataset.

Much of the quality of the output generated by the Editor will depend upon the prior preparatory work. Although the Editor can make some changes to the data, it is highly recommended that the necessary checks and changes be made in advance using a statistical package and a script. This ensures accurate and replicable results.

This section describes the various checks and balances involved in the data preparation process, such as: making a diagnostic on the structure of your data, cleaning it and identifying the various variables at the outset. Listed below are some common data problems that users encounter:

- Absence of variables that uniquely identify each record of the dataset
- Duplicate observations
- Errors from merging multiple datasets
- Encountering incomplete data when comparing the content of the data files with the original survey questionnaire
- Unlabelled data
- Variables with missing values
- Unnecessary or temporary variables in the data files
- Data with sensitive information or direct identifiers

Some practical examples using a statistical package are provided in *Section A “Data Validations in Stata: Practical Examples”*.

---

**Note:** If you are working in a data archive, be careful not to overwrite your original variables. Since managing databases involves several data-checking procedures, archive a new version in addition to the original. Work on this new version, leaving the original data files untouched.

---

The following procedures are recommended for preparing your dataset(s):

### 1.3.1 Data files should be organized in a hierarchical format

Look at your data and visualize it to understand its structure. It is preferable to organize your files in a hierarchical format instead of a flat format. In a hierarchical format, columns contain specific information about all possible units of analysis and rows form the individual observations (households, establishments, products, communities/countries, or any combination of those). Hierarchical files are easier to analyse, as they contain fewer columns that store the same information and are more compact. A flat format contains multiple columns with information on only one specific unit of analysis, so the information becomes redundant. For example, the information provided in one column is about the household head, and the row provides information on the child in the household.

**Table 1. Flat Format**

household ID	ID member 1	ID member 2	ID member 3	age member 1	age member 2	age member 3	relationship with hh member 1	relationship with hh member 2	relationship with hh member 3
1	1	2	3	30	28	10	Head of household	Spouse of head	Unmarried child
2	1	2	3	28	62	68	Head of household	Father	Mother
3	1	2	3	40	25	23	Head of household	Spouse of head	Married child
4	1	2	3	39	80	82	Head of household	Grandfather	Grandmother
5	1	2	3	55	31	5	Head of household	Daughter-in-law	Grandchild

**Table 2. Hierarchical Format**

household ID	ID members	age	relationship with head of household
1	1	30	Head of household
1	2	28	Spouse of head
1	3	10	Unmarried child
2	1	28	Head of household
2	2	62	Father
2	3	68	Mother
3	1	40	Head of household
3	2	25	Spouse of head
3	3	23	Married child
4	1	39	Head of household
4	2	80	Grandfather
4	3	82	Grandmother
5	1	55	Head of household
5	2	31	Daughter-in-law
5	3	5	Grandchild

*Tables 1 and 2* illustrate the two data structures. They contain the same information about six people on age and their relationship with the head of the household. The flat dataset (Table 1) stores the information on each family member in a new column. Note that for every additional member or characteristic, the dataset gets flatter and wider. The hierarchical dataset (see Table 2) has one observation and one row per person. Each variable contains a value that measures the same attribute across people and each record contains all values measured on the same person across variables. For every additional member characteristic, the dataset maintains the same number of columns, gains additional rows and gets longer and less flat compared to *Table 1*. The first two columns of this dataset have a hierarchical structure, where the ID member column is nested inside the ID household column.

Hierarchical files are easier to manage. Suppose in this example that there were many characteristics measured for everyone, the hierarchical structure would be a more convenient format because for each new characteristic, the dataset creates only one additional column, whereas, in the flat structure, it would create as many columns as there are people in the data with such characteristics.

### 1.3.2 Datasets with multiple units of analysis should be stored in different data files

It is recommended that you store your data in different files when you have multiple observational units. For example, *Table 3* shows a dataset that has both household-level data (columns on the type of dwelling and walls material) and individual-level data (columns on the marital status, work status, and worker category). Note that storing both levels of information in one dataset will result in a repetition of household characteristics for each household member. In *Table 3*, the information about the columns ‘type of dwelling’ and ‘wall material’ is repeated for everyone. Sometimes, this duplication is inefficient, and it is easier to have the dataset broken down by observational unit, into multiple files. In this example, it would be simpler to create two files: one for the household characteristics and another for the individual characteristics. The two files can be connected through a unique identifier, which in this case will be the household ID and member ID. We discuss the need for this unique identifier further on in this text as well.

**Table 3. Single data set with more than one observational unit**

Household ID	Number of persons	Type of dwelling	Walls material	ID member	Marital Status	Employed	Worker Category
1	3	Private House	Wood	1	Married	Yes	Gov't Employee
1	3	Private House	Wood	2	Common-Law	Yes	Private employee
1	3	Private House	Wood	3	Not stated	No	.
2	4	Townhouse	Concrete	1	Widowed	Yes	employee
2	4	Townhouse	Concrete	2	Married	Yes	Unpaid family worker
2	4	Townhouse	Concrete	3	Single	Yes	Self-employed
2	4	Townhouse	Concrete	4	Not stated	Yes	Private employee
3	2	Apartment	Blocks	1	Married	Yes	employee
3	2	Apartment	Blocks	2	Married	Yes	employee

### 1.3.3 Columns in a dataset should represent variables, not values

It is recommended that columns represent variables (e.g., sex, age, marital status) and rows represent observations (e.g., individuals, households, firms, products and so forth). In some datasets, columns instead of describing variables or attributes, describe values, which means that one variable is broken into segments and each one is stored in different columns. While this dataset structure can be useful for some analysis, the standard data structure where columns are variables and not values is the norm.

For example, *Table 4* (options 1 and 2) gives information at the individual-level on marital status, relationship with the head of the household and age. The difference between both tables is how the variable 'age' is reported. Option 1 had broken the variable 'age' into segments. This practice makes your data: i) messier, it has values of the variable as headings, and ii) inefficient, it increases the size of the dataset. Option 2 is recommended since there is only one heading and store of the information occupies less space, allowing the user to identify the structure of the data in a clear manner.

**Table 4. Data Structures: Hypothetical datasets**

**Option 1:**

Household ID	ID members	Marital Status	Relationship with head of household	age_10	age_20	age_30	age_40	age_50	age_60	age_70	age_n
1	1	Not stated	Head of household	0	0	1	0	0	0	0	0
1	2	Widowed	Spouse of head	0	0	0	1	0	0	0	0
1	3	Single	Unmarried child	1	0	0	0	0	0	0	0
2	1	Married	Head of household	0	0	0	0	1	0	0	0
2	2	Not stated	Father	0	0	0	0	0	1	0	0
2	3	Married	Mother	0	0	0	0	0	1	0	0
3	1	Common-Law	Head of household	0	0	0	1	0	0	0	0
3	2	Not stated	Spouse of head	0	0	0	0	1	0	0	0
3	3	Married	Married child	0	1	0	0	0	0	0	0
4	1	Widowed	Head of household	0	0	1	0	0	0	0	0
4	2	Married	Grandfather	0	0	0	0	0	0	1	0
4	3	Married	Grandmother	0	0	0	0	0	1	0	0

**Option 2:**

Household ID	ID members	Marital Status	Relationship with head of household	age
1	1	Not stated	Head of household	30
1	2	Widowed	Spouse of head	40
1	3	Single	Unmarried child	10
2	1	Married	Head of household	50
2	2	Not stated	Father	60
2	3	Married	Mother	60
3	1	Common-Law	Head of household	40
3	2	Not stated	Spouse of head	50
3	3	Married	Married child	20
4	1	Widowed	Head of household	30
4	2	Married	Grandfather	70
4	3	Married	Grandmother	60

### 1.3.4 Each observation in every file must have a unique identifier

Before you check for uniqueness of the identifiers in your files, you need to figure out the unit of analysis. Even if you are not the data producer, it is often easy to identify it. You can always review the documentation to see if the information has been provided. Below, some examples of units of analysis:

**Table 5. Unit of Analysis by Study type**

Study Type	Unit of analysis
<b>Income/Expenditure/Household Surveys</b>	- Households - Individuals - Consumption items
<b>Enterprise Surveys/Census</b>	- Firms - Establishments/Plants
<b>Agricultural Surveys/Census</b>	- Households - Crop area
<b>Research Data</b>	- Schools - Financial transactions - Exported products - Municipalities/precincts

Once you recognize the unit of analysis, the next step is to identify the column that uniquely identifies each record. If a dataset contains multiple related files, each record in every file must have a unique identifier. The data producer

can also choose multiple variables to define a unique identifier. In that case, more than one column in a dataset is used to guarantee uniqueness. These identifiers are also called **key variables** or **ID variables**<sup>2</sup>. The variable(s) should not contain missing values or have any duplicates. They are used by statistical packages such as SPSS, R or Stata when data files need to be merged for analysis

The absence of a unique identifier is a data quality issue, so one needs to ensure that the unique IDs remain fixed/present during the data cleaning process. If this correction is not possible, the archivist should note the anomalies in the documentation process.

### Best Practices

- It is recommended that ID variables be defined as a numeric since sorting and filtering records is much more efficient when variables are numeric.
- ID variables should not contain spaces, special characters or accents, since they may suffer modifications when the dataset is converted in different formats.
- For the convenience of users of the data, avoid identifiers consisting of too many variables. For example, in a household survey, the household identifier should ideally be a single variable (which you may create by concatenating a group of variables<sup>3</sup>), and the individual identifier should be the combination of only two variables (the household ID, and the sequential number of each member).
- It is recommended that you generate an ID based on a sequential number, however, keep in mind that it should not be too long because statistical packages and spreadsheet programs store a number of digits of precision, so opening a data set that contains ID variables with many characters, might result in truncated fields. For instance, the limit of the number of characters in Microsoft Excel is 15, so it changes any digits past the fifteenth place to zeroes.
- If you prepare your data files for public dissemination, it may be preferable to generate a unique household identification that would **not** be a compilation of geographic codes (because geographic codes are highly identifying). This recommendation is to ensure anonymity and will be explained in further detail later on in this text. The following example shows how to construct a unique identifier without using detailed information provided by the geographic codes.

### Example

- Suppose the unique identification of a household is a combination of variables PROV (Province), DIST (District), EA (Enumeration Area), HHNUM (Household Number). Options 2 and 3 are recommended. Note that if option 3 is chosen, it is crucial to preserve (but not distribute) a file that would provide the mapping between the original codes and the new HHID.

Option 1: Use a combination of four variables				Option 2: Generate a concatenated ID	Option 3: Generate a sequential number
PROV	DIST	EA	HHNUM	HHID	HHID
12	01	014	004	1201014004	1
12	01	015	001	1201015001	2
13	07	008	112	1307008112	3
Etc	Etc	Etc	Etc	Etc	Etc

Once you recognize the unit of analysis and the variable that uniquely identifies it, the following checks are suggested:

- Even if the data set has a variable with a label “unique identifier”, it is important to confirm that this variable truly does uniquely identify each record. To confirm or even to find out what the unique identifier is, you can

<sup>2</sup> See section 3 – *Importing data and establishing relationships* for more information on key variables.

<sup>3</sup> In Stata, this can be done through the use of the *group* function from the *egen* command. For example, to create a variable hhid based on a combination of variables *province*, *district*, *ea* and *hnum*, use the command “`egen hhid=group(province district ea hh_num)`”.

make use of the *-duplicate-* function in SPSS or the *-isid-* command in Stata (for R, do as shown in *Table 6*). For more details, refer to *Example 1* and *Example 2* of *Section A*.

**Table 6. Check for unique identifiers: STATA/R/SPSS Commands**

STATA	R	SPSS
<pre>use "household.dta"  isid key1 key2</pre>	<pre>my_data &lt;-   load("household.rda")  id &lt;-c( "key1" , " key2") library(eeptools) isid(my_data, id,   verbose=FALSE   verbose = FALSE) *</pre>	<p>Load dataset and choose from the menu:</p> <ul style="list-style-type: none"> <li>• Data&gt;Identify Duplicate Cases</li> <li>• Select Key Variables</li> </ul>

- Finally, check that the ID variable for the unit of observation doesn't have missing or assigned zero/null values. Ensure that the datasets are sorted and arranged by their unique identifiers.

*Table 7* below gives a hypothetical example. In this dataset, the highlighted columns (hh1, hh2, hh3) are the key variables, which means that they are supposed to make up the unique identifier. However, looking at those variables, we can identify some problems: the key variables do not uniquely identify each observation as they have the same values in rows 4 and 5, they also have some missing values (represented by asterisks), assigned zero values and some null values (those that say NA, don't know). All these issues suggest that those variables are not the key variables, and one needs to go back and double-check the data documentation. Alternatively, the archivist could check with the data producer and ask them how to fix these variables, in case those are indeed the key variables.

**Table 7. Check for unique identifiers: Hypothetical data set**

	Household ID			Number of persons	Type of dwelling	Walls material	Type of fuel
	hh1	hh2	hh3				
1	10	001	01	3	Private House	Wood	Coal
2	10	*	02	4	Townhouse	Concrete/Concrete Blocks	Wood
3	10	*	03	2	Apartment	Brick/Blocks	Gas/LPG/Cooking gas
4	20	001	01	8	Undivided Private House	Wood/Timber	Kerosene
5	20	001	01	3	Part of a Private House	Concrete/Concrete Blocks	Electricity
6	30	004		5	Flat, Apartment	Wood & Concrete	Other
7	*	004		2	Townhouse	Stone	Gas/LPG/Cooking gas
8	*	005	03	4	Double House	Brick/Blocks	Kerosene
9	30	005	NA	7	Combined Business & Dwelling	Plywood	Kerosene
10	40	006	don't know	8	Barns	Makeshift	Electricity
11	50	0	02	2	Other	Other	Coal
12	50	006	03	6	Condominium	Timber	Wood
13	50	006	04	3	Duplex	Don't Know	Other

*Example 3* provides further details and describes the steps involved in performing a validation when the identifier is made of multiple variables (see *Section A*).

### 1.3.5 Identifying duplicate observations

One way to rule out problems with the unique identifier is to check if there are duplicate observations (records with identical values for all variables, not just the unique identifiers). Duplicate observations can generate erroneous analysis and cause data management problems. Some possible reasons for duplicate data are, for example, the same record being entered twice during data collection. They could also arise from an incorrect reading of the questionnaires during the scanning process if paper-based methods are being used.

Identifying duplicate observations is a crucial step. Correcting this issue may involve eliminating the duplicates from the dataset or giving them some other appropriate treatment.

Statistical packages have several commands that help identify duplicates. *Table 8* shows examples of these commands in STATA, R and SPSS. The STATA command *-duplicates report-* generates a table that summarizes the number of copies for each record (across all variables). The command *-duplicates tag-* allows us to distinguish between duplicates and unique observations. For more details, refer to *Example 4* of *Section A*.

**Table 8. Check for duplicates observations: STATA/R/SPSS Commands**

STATA	R	SPSS
<pre>use "household.dta"  duplicates report  duplicates tag, generate (newvar)</pre>	<pre>my_data &lt;-   load("household.rda")  household[duplicated(household), ↵]</pre>	<p>Load dataset and choose from the menu:</p> <ul style="list-style-type: none"> <li>• Data&gt;Identify Duplicate Cases execute.</li> </ul>

### 1.3.6 Ensure that each individual dataset can be combined into a single database

For organizational purposes, surveys are often stored in different datasets. Therefore, checking the relationship between the data files is an essential step to keep in mind throughout the data validation process. The role of the data producer is to store the information as efficiently as possible, which implies storing data in different files. The role of the data user is to analyse the data as holistically as possible, which could sometimes mean that they might have to join all the different data files into a single file to facilitate analysis. It is essential to ensure that each of the separate files can be combined (merged or appended depending on the case) into a single file, should the data user want to undertake this step.

Use statistical software to validate that all files can be combined into one. For a household survey, for example, verify that all records in the individual-level files have a corresponding household in the household-level master file. Also, verify that all households have at least one corresponding record in the household-roster file that lists all individuals. Below, some considerations to keep in mind before merging data files:

- The variable name of the identifier should be the same across all datasets.
- The ID variables need to be the same type (either both numeric or both string) across all databases.
- Except for ID variables, it is highly recommended that the databases don't share the same variable names or labels.

#### Example

- A household survey is disseminated in two datasets; one contains information about household characteristics and the other contains information on the children (administered only to mothers or caretakers). To build a dataset containing all the information about the household characteristics, including where the children live, one needs to combine these files. Users are thus assured that all observations in the child-level file have corresponding household information.

#### Joining data files: Hypothetical data set



Household-level Data File						Child-level Data File							
	hh1	hh2	Number of persons	Type of dwelling	Walls material		hh1	hh2	Indiv	age	School Attendance	School level	Health Insurance
1	10	001	3	Private House	Wood	1	10	001	01	3	Yes	Pre-kindergarten	Yes
2	10	002	4	Townhouse	Concrete	2	10	001	02	5	Yes	Kindergarten	Yes
3	10	003	2	Apartment	Brick/Blocks	3	10	001	03	4	Yes	Kindergarten	Yes
4	20	001	8	Undivided Private House	Wood/Timber	4	10	002	01	8	Yes	2nd grade	Yes
5	20	002	3	Part of a Private House	Concrete	5	10	002	02	1	No	*	Yes
6	30	001	5	Flat, Apartment	Wood & Concrete	6	10	002	03	6	Yes	1st grade	Yes
7	30	002	2	Townhouse	Stone	7	10	003	01	2	No	*	Yes
						8	20	001	01	2	No	*	Yes
						9	20	001	02	4	Yes	Kindergarten	Yes
						10	20	002	01	3	Yes	Pre-kindergarten	Yes
						11	20	002	02	8	Yes	2nd grade	Yes
						12	20	002	03	2	No	*	Yes
						13	30	001	01	6	Yes	1st grade	Yes
						14	30	001	02	7	Yes	1st grade	Yes
						15	30	002	01	2	No	*	Yes
						16	30	002	02	5	Yes	Kindergarten	Yes
						17	30	002	03	1	No	*	Yes

Single and Combined Data File										
	hh1	hh2	Indiv	age	School Attendance	School level	Health Insurance	Number of persons	Type of dwelling	Walls material
1	10	001	01	3	Yes	Pre-kindergarten	Yes	3	Private House	Wood
2	10	001	02	5	Yes	Kindergarten	Yes	3	Private House	Wood
3	10	001	03	4	Yes	Kindergarten	Yes	3	Private House	Wood
4	10	002	01	8	Yes	2nd grade	Yes	4	Townhouse	Concrete
5	10	002	02	1	No	*	Yes	4	Townhouse	Concrete
6	10	002	03	6	Yes	1st grade	Yes	4	Townhouse	Concrete
7	10	003	01	2	No	*	Yes	2	Apartment	Brick/Blocks
8	20	001	01	2	No	*	Yes	8	Undivided Private House	Wood/Timber
9	20	001	02	4	Yes	Kindergarten	Yes	8	Undivided Private House	Wood/Timber
10	20	002	01	3	Yes	Pre-kindergarten	Yes	3	Part of a Private House	Concrete
11	20	002	02	8	Yes	2nd grade	Yes	3	Part of a Private House	Concrete
12	20	002	03	2	No	*	Yes	3	Part of a Private House	Concrete
13	30	001	01	6	Yes	1st grade	Yes	5	Flat, Apartment	Wood & Concrete
14	30	001	02	7	Yes	1st grade	Yes	5	Flat, Apartment	Wood & Concrete
15	30	002	01	2	No	*	Yes	2	Townhouse	Stone
16	30	002	02	5	Yes	Kindergarten	Yes	2	Townhouse	Stone
17	30	002	03	1	No	*	Yes	2	Townhouse	Stone

Statistical packages have some commands that allows us to combine datasets using one or multiple unique identifiers. Table 9 shows examples of these commands/functions in STATA, R and SPSS. For more details, refer to *Example 5* of Section A.

**Table 9. Joining data files: STATA/R/SPSS Commands**

STATA	R	SPSS
<pre>use "household.dta"  merge 1:m hh1 hh2 hh3 using "individuals.dta"</pre>	<pre>household &lt;- load(   ↪ "household.rda")  individuals &lt;- load(   ↪ "individuals.rda")  md &lt;- merge(household, ↪   ↪ individuals,   by = c("hh1",   ↪ "hh2", "hh3"),   all = TRUE)</pre>	<p>Load dataset and choose from the menu:</p> <ul style="list-style-type: none"> <li>• Data&gt;Merge Files&gt; Add Variables</li> <li>• Select the data file to merge</li> <li>• Select Key Variables</li> </ul>

Panel datasets should be stored in different files as well. Having one file per data collection period is a good practice. To combine the different periods of a panel dataset, the data user could merge them (Adding variables to the existing observations for the same period) or append them (Adding observations for a different period to the existing variables). To make sure that panels can be properly appended, the following checks are suggested:

- Check for the column(s) that identifies the period of the data (Year, Wave, Serie, etc.).
- The variable names and variable types should be the same across all datasets.
- Ensure that the variables use the same label and the same coding across all datasets.

In SPSS, use the function “Append new records” and in STATA the command *-append-* to combine datasets vertically.

### 1.3.7 Check for variables with missing values

Getting data ready for documentation also involves checking for variables that do not provide complete information because they are full of missing values. This step is important because missing values can have unexpected effects on the data analysis process. Typically, missing values are defined as a character (.a, .b, single period or asterisks), special numeric (-1, -2) or blanks. Variables entirely comprised of missing values should ideally not be included in the dataset. However, before excluding them, it is useful to check whether the missing values are expected according to the questionnaire, and the skip patterns.

For example, a hypothetical household survey at the individual-level (Table 10) provides information about the respondent’s employment status. The survey identifies if the respondent is employed in Column D, and then provides information about the worker category in Column E, but only for those who reported being employed in Column D. This means that those who answered ‘unemployed’ in column D should have a valid missing value in column E. In other words, this is a pattern in the missing values that should be observed and duly noted.

On the other hand, Columns F and G are used to determine if the people who are not employed are looking for a job and are actively seeking it. These questions are not asked to the employed people (those who answered “yes” in Column D), which mean that again, the missing values in those columns correspond with what is expected. However, Column H contains information for all employed individuals, so missing values in this column suggest that there is a problem in the data and should be addressed. Therefore, one should not blindly delete missing values at the outset without checking for these patterns.

**Table 10. Checking for Missing Values: Hypothetical data set**

	A	B	C	D	E	F	G	H
	ID member	Age	Marital Status	Employed	Worker Category	Looking for Job	Steps to look for work	Health Insurance
1	1	22	Married	Yes	Gov't Employee	*	*	*
2	2	30	Common-Law	Yes	Private employee	*	*	*
3	3	19	Not stated	No	*	No	*	*
4	1	45	Widowed	No	*	No	*	*
5	2	32	Married	No	*	No	*	*
6	3	21	Single	Yes	Self-employed	*	*	*
7	4	29	Not stated	Yes	Private employee	*	*	*
8	1	39	Married	Yes	employee	*	*	*
9	2	50	Married	Yes	employee	*	*	*

In SPSS, use the function “Missing Value Analysis” and in R, do as shown in Table 11. You can also use the STATA command `-misstable summarize-` that produces a report that counts all the missing values. You can also use the `-rowmiss()-` command with `-egen-` to generate the number of missing values among the specified variables. For more details, refer to Example 6 of Section A.

**Table 11. Counting Missing Values: STATA/R/SPSS Commands**

STATA	R	SPSS
<pre>use "individual.dta"  misstable summarize</pre>	<pre>my_data &lt;-   load("individual.rda")  colSums(is.na(individual))  colMeans(is.   ↪na(individual))</pre>	<p>Load dataset and choose from the menu:</p> <ul style="list-style-type: none"> <li>• Data&gt;Analyze&gt; Missing Value Analysis</li> <li>• Select “Use All Variables”</li> </ul>

#### *Best Practices*

Since there are different reasons for missing values, data producer should code them with negative integers or letters to distinguish the missing values and valid data. For instance, ( 1) might be the code for “Don’t Know”, (-2) the code for “Refused to Answer” and (-9) code for “Not Applicable”.

### 1.3.8 Check Improper value ranges

It is helpful to generate descriptive statistics for all variables (frequencies for discrete variables; min/max/mean for continuous variables) and verify that these statistics look reasonable. Just as there are variables that must take on only specific values, such as “F” and “M” for gender, there are also some variables that can take on several values (such as age or height). However, those values must fit a particular range. For example, we don’t expect negative values, or typically see values over 115 years for age.

Values for categorical variables should be guided by the questionnaire (or separate documentation for constructed variables). If we have an education variable that has 9 response options in the questionnaire, the corresponding ‘education’ variable in the dataset should have 9 categories. We should not observe more than 9 unique values for this variable. Similarly, for any questions in the survey for which the options are only “yes”, “no” and “other”, we should not observe more than these 3 unique values. When out of range values exist, this might signal data cleaning issues.

Table 12 shows examples of some commands/functions in STATA, R and SPSS.

**Table 12. Generate descriptive statistics: STATA/R/SPSS Commands**

STATA	R	SPSS
<pre>use "individual.dta"  summarize</pre>	<pre>individual &lt;-   load("individual.rda")  summary(individual)</pre>	<p>Load dataset and choose from the menu:</p> <ul style="list-style-type: none"> <li>• Data&gt;Analyze&gt; Descriptive Statistics&gt; Frequencies</li> <li>• Select “Statistics”</li> </ul>

## 1.3.9 Verify that the number of records in each file corresponds to what is expected

The technical documentation helps to form some expectations about the size of the dataset. Make sure that in all the files, the number of records is the same as (or is similar to) what is explicitly stated in the sample design of your survey.

Suppose that you have a household survey and according to the documentation the sample size is 50,321 households. Consequently, the file that contains the household-level information should have a similar number of observations. When this is not the case, you should be able to account for this difference in data documentation.

On the other hand, even if the number of individual records is not available in the documentation, you can still perform a rough check on the files. For example, if you have the household level file and the person level file, the latter should be between 2 or 6 times larger than the former, depending on the average household size in the country for which the information has been collected. Another example is to compare the household level file of an expenditure survey with the consumption level file (at the product-level).

The latter should have  $n$  times the number of observations than the former, where  $n$  is the average number of products that each household records in the survey.

## 1.3.10 Datasets must contain all variables from the questionnaire and be in a logic sequence

Verify the completeness of your data files by comparing the content of these files with the survey questionnaire. All variables in the questionnaire should appear in the dataset, except those excluded on purpose by the producer of the data because of reasons of confidentiality (see numeral 1.15). Cross-checking with the questionnaire(s) is needed to ensure that all sections are included in the dataset.

Additionally, it is a good practice to make sure that the database is sorted in the same order as the questionnaire. This practice will help users navigate seamlessly across the dataset using the questionnaire as a route map.

The Stata command *-describe-* displays the names, variable labels and other characteristics, which helps us verify that no variables have been omitted in the database. It simultaneously confirms that all variables are correctly ordered. Refer to *Example 7* of *Section A* for further details.

## 1.3.11 Include the relevant weighting coefficients and variables identifying the stratification levels

All data files of a sample survey should have clearly labelled variable(s) with information on the survey weights. Sample surveys need to be representative of a broader population for which the data is collected, and the user needs the survey weights for almost every analysis performed. In the case of household surveys, the survey weights are equal among members of the same household but differ across households. Weights are positive and strictly higher than zero. They should not have a larger value than the population for which the survey is representative.

A more detailed description of how the survey weights would look like should be provided in the documentation of the survey. Based on it, you can perform some basic range checks. Notice that Census datasets do not need weights since

a census collects data on all the individuals in the population. There are however some exceptions, for example in the case of IPUMS, the data collected are not full censuses but census samples, so weights are required in this context.

Additionally, for sample surveys, verify that the variables identifying the various levels of stratification and the primary sampling unit are included and easily identifiable in at least one of the data files. These variables are needed for the calculation of sampling errors.

### 1.3.12 Variables and codes for categorical variables must be labelled

#### Variable labels

Labels should be short and precise. They should provide a clear indication of what information is contained in the variables. Variable labels are brief descriptions or attributes of each variable. Without variable labels, users are not able to link the variables in the database to the questions of the questionnaire. So, one should ensure that all variables are labelled.

Additionally, even if variables are fully labelled, the following practices must be considered:

- Variable labels can be up to 80 characters long in Stata and 255 in SPSS, however, it is recommended that labels be informative, short and accurate.
- It is a common practice to have a literal question from the survey as a variable label. However, the literal questions are usually longer than the maximum number of characters, so this is not an advisable practice.
- The same label should not be used for two different variables.

#### Value labels

Label values are used for categorical variables. To ensure the correct encoding of data, it is important to check that the stored values in those variables correspond to what is expected according to the questionnaire. In the case of continuous variables, we also suggest the checking of ranges. For instance, if the question is about the number of working hours, the variable should not have negative values.

You can compare variable labels in the dataset to those in the questionnaire using the `-codebook-` Stata command or `-labelbook-`. Refer to *Example 8* of *Section A* for further details.

### 1.3.13 Temporary, calculated or derived variables should not be disseminated

Remove all unnecessary or temporary variables from the data files. These variables are not collected in the field and present no interest for users.

The data producer could generate variables that are only needed during the quality control process but are not relevant to the final data user. For example, the variable “\_merge” in Stata is generated automatically after performing the check described in the Numeral 1.6, when the data producer wants to see if the datasets match properly. Variables that group categories of a question, dummy variables that identify a question’s category are all variables produced during the coding process that are not relevant once the analysis is completed.

There are cases in which calculated variables may be useful to the users, so they must be documented in the meta-data. For example, most Labor Force Surveys (LFS) contain derived dummy variables to identify the sections of the population that are employed or unemployed. These variables are generated using multiple questions from the dataset and are essential elements of any LFS. Most data users prefer to make use of them instead of computing them on their own, to reduce the risk of error. This is a strong argument to make a case for keeping these variables in the dataset, despite them being a by-product of other original variables.

To be useful, those variables that remain in the dataset must be well documented, else they, they may be useless to or misunderstood by users.

### 1.3.14 Check that the data types are correct

Do not include string variables if they can be converted into numeric variables. Look at your data and check the variables' types, particularly for those that you expect to be numeric (age, years, number of persons/employees/hours, income, purchases/expenditures, weights, and so forth). If there are numeric variables stored as string variables, your data needs cleaning.

For example, *Table 13* contains a data set at the individual-level with some variables that should be numeric. The columns B (Age) and E (Working Weeks) are stored as numeric variables, which is fine. However, the variables 'Number of working of hours per week' (Column G), 'Number of persons working at the business' (Column H) and 'Monthly Income' (Column I) are loaded as strings because there are non-numeric values (don't know, skip, refused to answer) and some missing values present. Those variables need to be cleaned and converted from string variables to numeric variables.

**Table 13. Checking Data Types: Hypothetical data set**

	A	B	C	D	E	F	G	H	I
	ID member	Age	Marital Status	Employed	Worker Category	Working Weeks	Working hours per week	Number of persons working at the business	Monthly Income
1	1	22	Married	Yes	Gov't Employee	48	40	4	5000
2	2	30	Common-Law	Yes	Private employee	50	40	19	Refused to answer
3	3	19	Not stated	Yes	Private employee	32	40	49	4000
4	1	45	Widowed	Yes	employee	42	40	99	7000
5	2	32	Married	Yes	Gov't Employee	30	Skip	Don't know	Refused to answer
6	3	21	Single	Yes	Self-employed	50	40	20	6000
7	4	29	Not stated	Yes	Private employee	50	40	12	5800
8	1	39	Married	Yes	employee	50	40	NA	Skip
9	2	50	Married	Yes	employee	20	-	30	2500

Statistical packages have some commands that allows us to make such conversions. *Table 14* shows examples of these commands/functions in STATA, R and SPSS.

**Table 14. Convert string variables to numeric: STATA/R/SPSS Commands**

STATA	R	SPSS
<pre>use "individual.dta"  destring (varname), {generate replace}</pre>	<pre>individual &lt;- load("individual.rda")  individual\$varname &lt;- as.numeric(individual ↪\$varname)</pre>	<p>Load dataset and choose from the menu:</p> <ul style="list-style-type: none"> <li>• Data&gt;Transform&gt; Recode into Same Different Variables</li> <li>• Select the variable</li> <li>• Select "Old and New Values" and Recode it</li> <li>• Select "Convert numeric strings to numbers ('5'-&gt;5)</li> </ul>

### 1.3.15 Datasets must not have directed identifiers

One must verify that in all data files, sensitive information or direct identifiers that could reveal the identity of the respondent directly (names, addresses, GPS coordinates, phone numbers, etc.) have been removed. Check to ensure this information is not in the dataset(s). If it is, those variables need to be removed from shared datasets.

Keep in mind that if you are preparing a dataset for public release, you need a cleaned, anonymous dataset. Removing all direct identifiers is the first key step to ensuring the anonymity of the participants. However, before you start any privacy procedures, you should always check your data.

For more information on how to apply statistical disclosure control (SDC) methods to data before release, see the document “Introduction to Statistical Disclosure Control (SDC)” available at <http://ihnsn.org/sites/default/files/resources/ihnsn-working-paper-007-Oct27.pdf>

### 1.3.16 Compress the variables to reduce the file size

Compress the variables consist of reducing the size of the data file without loss of precision or modifying the information that it provides. Listed below are some reasons why compressing a data set may be a useful practice for at least three reasons: First, it makes faster the process of creating backups, uploading and downloading data files from your data repository or any Survey Catalog. Second, it reduces the time that data users will need to spend working with the data. Additionally, it will make the data more accessible to the different type of users; sometimes the data size will impose restrictions on those users who lack high computational power. Third, it will help to free up disk space in the server where you store your data

#### Example

- *Table 15* shows two versions of one dataset that provides individual-level information about the year of the first union, age, school attendance, and health insurance. There is no difference in the appearance of both datasets. However, version 1 was saving uncompressed and version 2 compressed. In the uncompressed version, the variables “ID” and “Year” are stored as double, which means that they can store number with high decimal precision, but they are designed to only record information of integer numbers between -32,767 and 32,740. So, the compressed version changed the storage type of these variables to int and saves 6 bytes per observation. Similarly, other variables like “age” and “school attendance” are stored as a byte in the compressed version, which saves 7 bytes per observation when are compared to the uncompressed version. Let’s suppose that one has a data set with 500 variables like these, the total savings would be 3,500 bytes per observation; if this data set has 50,000 observations, it means that the savings in memory space would be around 175 megabytes.

**Table 15. Compressing the Variables: Hypothetical data set**

Data Set Version 1: Uncompressed File					Variable Properties		
A	B	C	D	E	Variable Name	Storage Type	Display Format
ID	Year	Age	School_Attendance	Health_Insurance			
1	1	1980	15 Lower Secondary	A	ID	double	%10.0g
2	2	2017	55 Tertiary Education		Year	double	%10.0g
3	10	2017	67 Upper Secondary		Age	double	%10.0g
4	24	2017	34 Lower Secondary		School_Attended	double	%10.0g
5	55	Inconsistent	Upper Secondary		Health_Insurance	str3	%3s
6	100	2018	47 Upper Secondary	A			
7	234	2018	Tertiary Education				
8	876	2018	77 Tertiary Education				
9	106	2019	Lower Secondary				
10	1170	DK	33 Tertiary Education	A			

Data Set Version 2: Compressed File

	A	B	C	D	E
	ID	Year	Age	School_Attendance	Health_Insurance
1	1	1980	15	Lower Secondary	A
2	2	2017	55	Tertiary Education	
3	10	2017	67	Upper Secondary	
4	24	2017	34	Lower Secondary	
5	55	Inconsistent		Upper Secondary	
6	100	2018	47	Upper Secondary	A
7	234	2018		Tertiary Education	
8	876	2018	77	Tertiary Education	
9	106	2019		Lower Secondary	
10	1170	DK	33	Tertiary Education	A

Variable Properties

Variable Name	Storage Type	Display Format
ID	int	%10.0g
Year	int	%10.0g
Age	byte	%10.0g
School_Attended	byte	%10.0g
Health_Insurance	str1	%9s

Use the *compress* command in Stata, or the *compress* option when you save a SPSS data file.

**Suggestion** If you are in the process of establishing a data archive and plan to document a collection of surveys, undertake a full inventory of all existing data and metadata before you start the documentation. Use the *IHSN Inventory Guidelines and Forms* to before you start the documentation. Use the *IHSN Inventory Guidelines and Forms* to facilitate this inventory (available at [www.surveynetwork.org](http://www.surveynetwork.org)).

## Suggestion:

If you are in the process of establishing a data archive and plan to document a collection of surveys, undertake a full inventory of all existing data and metadata before you start the documentation. Use the *IHSN Inventory Guidelines and Forms* to before you start the documentation. Use the *IHSN Inventory Guidelines and Forms* to facilitate this inventory (available at [www.surveynetwork.org](http://www.surveynetwork.org)).

## 1.4 Gathering and preparing the documentation

All information related to the survey may be useful and should be archived (even if not all will be disseminated to the public). This includes not only technical documents such as the questionnaires or list of codes (obviously needed by data users), but also administrative reports (potentially useful for implementation of future surveys), and other documents such as a compilation of the comments provided by stakeholders at the time the questionnaire was designed, etc. Resources to be included if available include:

- The survey questionnaire(s); make sure that the cover page and all sections are included. If the questionnaire exists in multiple languages, provide all versions.
- All technical, analytical and administrative documents
  - Sampling information
  - Interviewers and supervisor's manuals
  - List of codes
  - Instructions for data editing
  - Survey report (tabulation and analysis)
  - Analytical papers and policy briefs that made use of the data
  - Survey budget and other key planning documents
  - PowerPoint presentations and other related material
- Computer programs (used for data entry, editing, tabulation and analysis)
- Photos



- Tables
- Maps
- Survey promotional/informational materials (flyers, videos, posters, songs, etc.)

Documents available in electronic format (MS-Word, Excel, and others) must be preserved in their original format and in PDF format.

All documents available only on hard copy must be scanned. Use low resolution graphics, and black & white option (unless it is crucial to preserve colours) to avoid large file sizes. A scanning resolution of 300 dpi is recommended. Save the scanned documents in PDF format. OCR is useful, although not required.

Scan all resources with an updated virus detection application.

## 1.5 Importing data and establishing relationships

After all data and documentation files are gathered and checked, import the data files in the Editor. In the Metadata Editor, order the files in a logical fashion (e.g., sequentially through sections).

---

**Note:** If you are documenting a population census and have very large data files, it is recommended to split the files by geographic area. Typically, you will have a file at individual level, one at the household level, and possibly one at the community level, for each State or Province. In such case, import all files for one State or Province only. You will import the other data files after you complete the documentation of the files. This will considerably reduce the time needed to save your files. The Metadata Editor will allow you to replicate the metadata from the documented files to all other data files that you will import later.

---

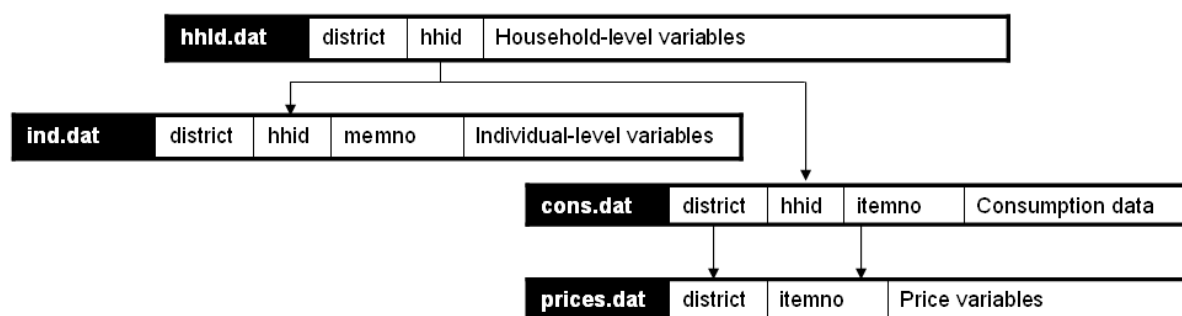
After all files are imported and ordered in a proper sequence, define the key variables for each data file. The base key variable(s) in a data file is (are) the variable(s) that provide the unique identifier of each record in that specific data file.

Then establish the relations and validate them using the *Tool > Validate Dataset Relations* in the Editor. This automatic validation is a way to check the structural integrity of the identifier variables and assure there are no duplicates in the data.

## Establishing relationships – An example

In this example, we assume that the dataset is obtained from a household budget survey and comprises:

- A household-level file “hhld.dat” with the household characteristics (one record per household). Each household is identified by a variable named *hhid*.
- A household-level file “hhld.dat” with the household (one record per person). Each household member is identified by the combination of variables *hhid* and *memno*.
- A consumption data file “cons.dat”, with one record per item (goods and services) per household. Each record is uniquely identified by the combination of variables *hhid* and *itemno*. The file also contains a variable *district* identifying the district where the household resides.
- A data file “prices.dat” with average price per commodity, per district (one record per item per district). Each record is uniquely identified by variables *district* and *itemno*.



In the Metadata Editor, these relationships will be established as follows in the “Key variables and relationships” section of each data file:

hhld.dta	ind.dat	cons.dat	prices.dat
<div><div>RELATIONS</div><div><div>Relations</div><div>ind</div><div>cons</div></div></div> <div><div>BASE KEY VARIABLES</div><div><div><div>Number</div><div>V2</div></div><div><div>Name</div><div>hhid</div></div></div></div>	<div><div>RELATIONS</div><div><div>Relations</div><div>hhid</div></div></div> <div><div>BASE KEY VARIABLES</div><div><div><div>Number</div><div>V2</div><div>V3</div></div><div><div>Name</div><div>hhid</div><div>memno</div></div></div></div>	<div><div>RELATIONS</div><div><div>Relations</div><div>hhld</div></div></div> <div><div>BASE KEY VARIABLES</div><div><div><div>Number</div><div>V2</div><div>V3</div></div><div><div>Name</div><div>hhid</div><div>memno</div></div></div></div>	<div><div>RELATIONS</div><div><div>Relations</div><div>cons</div></div></div> <div><div>BASE KEY VARIABLES</div><div><div><div>Number</div><div>V1</div><div>V2</div></div><div><div>Name</div><div>district</div><div>itemno</div></div></div></div>

If you have imported your data from any format other than fixed ASCII, re-sequence the data using the *Variables > Resequence* option in the Editor. This re-sequencing tool will automatically fill the “StartCol” and “EndCol” columns in the variable description section. This must be done for each data file.

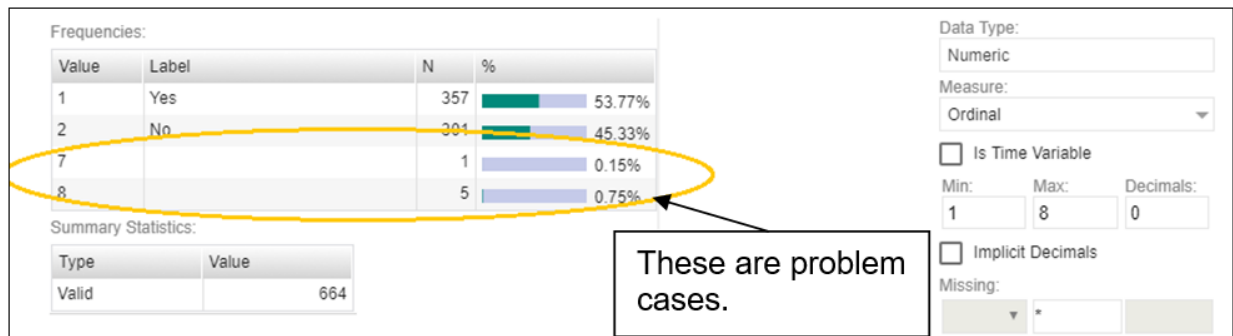
**VARIABLES**

Number	Name	Label	Width	Start Col	End Col	Data Type	Measure
V1	hid	Household ID	2	*	*	Numeric	Nominal
V2	hi6	Area	1	*	*	Numeric	Nominal
V3	hi10	Result of household interview	1	*	*	Numeric	Nominal
V4	hi1	Line number	2	*	*	Numeric	Nominal

**VARIABLES**

Number	Name	Label	Width	Start Col	End Col	Data Type	Measure
V1	hid	Household ID	2	1	2	Numeric	Nominal
V2	hi6	Area	1	3	3	Numeric	Nominal
V3	hi10	Result of household interview	1	4	4	Numeric	Nominal
V4	hi1	Line number	2	5	6	Numeric	Nominal

Before going further, quickly browse all variables in all data files to visually check the frequencies. This will allow you to easily spot some outliers or invalid codes, which will require recoding (which can be done in the Editor or in the source data files which will then have to be re-imported).



Save the project. The Editor saves the full project, the associated data and documentation in one zip file. We recommend you save the project using the survey abbreviation, year and version number as project name (e.g., UGA\_2018\_HIES\_v01\_M). Note that it is good practice to avoid using spaces in a file name (use underscore characters instead).

## 1.6 Importing external resources

Before importing your external resources, create folders in the Editor as necessary (these are directories in the External Resources section in the Editor, not new directories on your hard drive). If you have very few external resources, all resources can be listed in the root directory. If you have many, organize them by type of resources (in the example below, we have created separate directories for the Questionnaires, Technical Documents, Computer Programs, Reports, Tables, Photos and Maps).

	<p>Create an entry for each resource by entering a label in the Resource Information field. This label should be short but explicit. Then identify the resource file in the “Resource” field. The field “Resource” is used to indicate the filename or URL location (website) of the external resource. The resource consists of the filename, and a relative path. The reason for entering a relative path is that it will allow you to move the whole study directory and its subdirectories to another location or another drive, without having to re-enter the location of the files.</p>
--	--

Example:

Let’s assume your study is a Household Budget Survey conducted in 2018. If you followed the recommendations made in the introductory chapter “Before you start – Organizing your files”, you will have created a directory like C:UGA\_2018\_HIES. Suppose also that a document titled Report2018.pdf is saved in a directory C:UGA\_2018\_HIESDoc. When you fill the resource field in the External Resources page, do NOT enter “C:UGA\_2018\_HIESDocReport2018.pdf.pdf. Enter the file name as follows: DocReportsReport2018.pdf

Some resources might be composed of more than one file (for example, the CSPro data entry application includes multiple files that should not be separated). In such cases, zip them into one single file, and import it as a single

resource.

For documents available in multiple formats (for example, a questionnaire available in Excel and in PDF), you may create two separate resources, or zip the files into one single file. In such case, list the different formats available in the “Content/ Description” field.

*Best Practices – Naming Convention for External Resources*

- Use file names short, but self-explanatory about the content of the document.
- Preferably, use lower cases.
- Avoid spaces to delimit words.
- Be consistent with your method of naming across all files. For instance, if you use underscores to delimit words, keep it that way in all files.
- Use only alphanumeric characters, underscores or dashes. Avoid using special characters (!@#\$\$%^&\*()~) or any accented characters.
- If you intend to have an archive useable and downloadable across multiple countries, use English names for your files.

## 1.7 Completing metadata

The Metadata Editor makes use of the Data Documentation Initiative (DDI Version 2.5), the Dublin Core (DCMI version X) metadata standards and ISO 19139 for geospatial information.

The table below provides an overview of the different metadata standards as related to the project. Each metadata standard is integrated into the template that will define the project.

Table: Standards included in the templates		Metadata Standard		
		DDI	Dublin Core	ISO19139
Projects	Survey	✓	✓	
	Time Series	✓	✓	
	Geospatial			✓
	Document		✓	
	Table		✓	

A thorough completion of the DDI and DCMI elements will significantly raise the value of the archiving work by providing users with the necessary information to put the study into its proper context and to understand its purpose.

The DDI requires completion of the following sections: Document Description, Study Description, Datasets, Variables Groups, and External Resources. Recommendations for each field included in the IHSN template are provided below.

The IHSN recommends using the standardized IHSN DDI/DCMI templates (Study Template and External Resources Template). This Quick Reference Guide is based on these two templates. Visit the IHSN website to download the latest version of these templates, available in multiple languages.

### Overall recommendations:

- As an archivist, you may need to seek assistance from key experts involved in some of the technical aspects of the survey.

- As a general rule, avoid using ALL CAPS when you fill DDI fields. Also, check the spelling of all entries. The Editor does not provide (yet) an automatic spell checker.
- Some of the examples below present an optimal documentation of some fields. In many cases, for past surveys, you will not find such detailed information. Try to provide as much detail as possible. For future surveys, the information should be compiled and provided during the whole life cycle of the survey. This will ensure that the best possible documentation is available at completion of that survey.

## 1.7.1 Good practices for completing the Document Description

Documenting a study using the DDI and DCMI metadata standards consists of generating a metadata file which will be saved in XML format in what is called an *XML Document*. The *Document Description* described below is a description of that XML file. The IHSN Template selected 4 elements to describe the DDI document.

Metadata Producer	Name of the person(s) or organization(s) who documented the dataset. Use the “role” attribute to distinguish different stages of involvement in the production process. Example:	
	<i>Name</i>	<i>Role</i>
	<i>National Statistics Office (NSO)</i>	<i>Documentation of the study</i>
	<i>International Household Survey Network (IHSN)</i>	<i>Review of the metadata</i>
Date of Production	This is the date (in ISO format YYYY-MM-DD) the DDI document was produced (not distributed or archived). This date will be automatically imputed when you save the file.	
DDI Document Version	Documenting a dataset is not a trivial exercise. Producing “perfect” metadata is probably impossible. It may therefore happen that, having identified errors in a DDI document or having received suggestions for improvement, you decide to modify the Document even after a first version has been disseminated. This element is used to identify and describe the current version of the document. It is good practice to provide a version number (and date), and information on what distinguishes this version from the previous one(s) if relevant. <b>Example:</b> <i>Version 02 (July 2017). This version is identical to version 01, except for the section on Data Appraisal which was updated.</i>	
DDI Document ID Number	The ID number of a DDI document is a unique number that is used to identify this DDI file. Define and use a consistent scheme to use. Such an ID could be constructed as follows: DDI_COUNTRY_PRODUCER_SURVEY_YEAR where <ul style="list-style-type: none"> <li>• <i>country</i> is the 3-letter ISO country abbreviation</li> <li>• <i>producer</i> is the abbreviation of the producing agency</li> <li>• <i>survey</i> is the survey abbreviation</li> <li>• <i>year</i> is the reference year (or the year the survey started)</li> <li>• DDI document version number</li> </ul> <b>Example:</b> <i>The DDI file related to the Demographic and Health Survey documented by staff from the Uganda Bureau of Statistics in 2005 would have the following ID: DDI_UGA_UBOS_DHS_2005_v01. If the same survey is documented by a staff from the IHSN, this would be DDI_UGA_IHSN_DHS_205_v01.</i>	
Collection	This field allows viewed and searched the study by collection	
Programs	Link surveys to projects / trust funds	

## 1.7.2 Good practices for completing the Study Description

In the DDI standard, the Study Description is the section that contains all elements needed to describe the study itself (investigators, dates and methods, scope and coverage, etc.)

<b>Identification</b>	
Survey Title	<p>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:</p> <ul style="list-style-type: none"> <li>• Include the reference year(s) of the survey in the title.</li> <li>• Do not include the abbreviation of the survey name in the title.</li> <li>• As the survey title is a proper noun, the first letter of each word should be capitalized (except for prepositions or other conjunctions).</li> <li>• Including the country name in the title is optional.</li> </ul> <p><b>Example:</b></p> <ul style="list-style-type: none"> <li>• <i>National Household Budget Survey 2012-2013</i></li> <li>• <i>Popstan Multiple Indicator Cluster Survey 2012</i></li> </ul>
Survey Subtitle	<p>Subtitle is optional and rarely used. A subtitle can be used to add information usually associated with a sequential qualifier for a survey.</p> <p><b>Example:</b> Title: <i>Welfare Monitoring Survey 2007</i>  Subtitle: <i>Fifth round</i></p>
Abbreviation or Acronym	<p>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.</p> <p><b>Example:</b></p> <ul style="list-style-type: none"> <li>• <i>DHS 2015 for "Demographic and Health Survey 2005"</i></li> <li>• <i>HIES 2012-2013 for "Household Income and Expenditure Survey 2003"</i></li> </ul>
Study type	<p>The study type or <i>survey type</i> 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).</p>

Continued on next page

Table 1 – continued from previous page

Series information	<p>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.</p> <p><b>Example:</b> <i>The Multiple Indicator Cluster Survey, Round 3 (MICS3) is the third round of MICS surveys, previously conducted around 1995 (MICS1) and 2000 (MICS2). MICS surveys are designed by UNICEF, and implemented by national agencies in participating countries. MICS was designed to monitor various indicators identified at the World Summit for Children and the Millennium Development Goals. Many questions and indicators in MICS3 are consistent and compatible with the prior round of MICS (MICS2) but less so with MICS1, although there have been a number of changes in definition of indicators between rounds. Round 1 covered X countries, round 2 covered Y countries, and Round Z covered N countries.</i></p>
Translated title	<p>In countries with more than one official language, a translation of the title may be provided. Likewise, the translated title may simply be a translation into English from a country’s own language. Special characters should be properly displayed (such as accents and other stress marks or different alphabets).</p>
Unique user defined ID Number	<p>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: country-producer-survey-year-version where</p> <ul style="list-style-type: none"> <li>• <i>country</i> is the 3-letter ISO country abbreviation</li> <li>• <i>producer</i> is the abbreviation of the producing agency</li> <li>• <i>survey</i> is the survey abbreviation</li> <li>• <i>year</i> is the reference year (or the year the survey started)</li> <li>• <i>version</i> is the number dataset version number (see Version Description below)</li> </ul> <p><b>Example:</b> <i>The Demographic and Health Survey implemented by the Uganda Bureau of Statistics in 2005 could have the following ID: UGA-UBOS-DHS-2005-v01.</i></p>
Depositor	<p>The name of the person (or institution) who provided this data collection to the archive storing it.</p>

Continued on next page



Table 1 – continued from previous page

Date of Deposit	The date that the data collection was deposited with the archive that originally received it.
<b>Version</b>	
Version Description	<p>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 that larger series be defined by a number to the left of a decimal and iterations of the same series by a sequential number that identifies the release. Larger series will typically include (0) the raw, unedited dataset; (1) the edited dataset, non anonymized, for internal use at the data producing agency; and (2) the edited dataset, prepared for dissemination to secondary users (possibly anonymized).</p> <p><b>Example:</b></p> <ul style="list-style-type: none"> <li>• <i>v00: Basic raw data, obtained from data entry (before editing).</i></li> <li>• <i>v01: Edited data, second version, for internal use only.</i></li> <li>• <i>v02: Edited, anonymous dataset for public distribution.</i></li> </ul> <p>A brief description of the version should follow the numerical identification.</p>
Production date	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.
Version Notes	Version notes should provide a brief report on the changes made through the versioning process. The note should indicate how this version differs from other versions of the same dataset.
<b>Study Authorization</b>	
Authorizing Agency	Name of the agent or agency that authorized the study. The “affiliation” attribute indicates the institutional affiliation of the authorizing agent or agency. The “abbr” attribute holds the abbreviation of the authorizing agent’s or agency’s name
Authorization Statement	<p>The text of the authorization. Use XHTML to capture significant structure in the document</p> <p><b>Example:</b></p> <p>Required documentation covering the study purpose, disclosure information, questionnaire content, and consent statements was delivered to the OUHS on 2010-10-01 and reviewed by the compliance officer. Statement of authorization for the described study was received on 2010-11-04</p>
Legal basis	Decree or law authorizing or requiring the study (e.g. census act)
<b>Overview</b>	

Continued on next page

Table 1 – continued from previous page

Study Budget	Describe the budget of the project in as much detail as needed. Internal structure is allowed using XHTML elements. Different organizations express their budgets in different formats and this open format allows flexibility. Attributes: Budget line ID, Budget line label, Amount, Currency and Source of funding.
Abstract	The abstract should provide a clear summary of the purposes, objectives and content of the survey. It should be written by a researcher or survey statistician aware of the survey.
Objectives of the study	Describe the Main (explicit) and secondary (explicit) objectives of the survey.
Kind of 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	A survey could have various units of analysis. These are fairly standard and are usually: <ul style="list-style-type: none"> <li>• Household (household survey, census)</li> <li>• Person (household survey, census)</li> <li>• Enterprise (enterprise survey)</li> <li>• Commodity (household survey, price survey)</li> <li>• Plots of land (agricultural survey)</li> </ul>
<b>Scope</b>	

Continued on next page

Table 1 – continued from previous page

Description of scope	<p>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. The scope does not deal with geographic coverage.</p> <p><b>Example:</b> The scope of the Multiple Indicator Cluster Survey includes:</p> <ul style="list-style-type: none"> <li>• <i>HOUSEHOLD: Household characteristics, household listing, orphaned and vulnerable children, education, child labour, water and sanitation, household use of insecticide treated mosquito nets, and salt iodization, with optional modules for child discipline, child disability, maternal mortality and security of tenure and durability of housing.</i></li> <li>• <i>WOMEN: Women's characteristics, child mortality, tetanus toxoid, maternal and newborn health, marriage, polygyny, female genital cutting, contraception, and HIV/AIDS knowledge, with optional modules for unmet need, domestic violence, and sexual behavior.</i></li> <li>• <i>CHILDREN: Children's characteristics, birth registration and early learning, vitamin A, breastfeeding, care of illness, malaria, immunization, and anthropometry, with an optional module for child development.</i></li> </ul>
Topic classifications	<p>A topic classification facilitates referencing and searches in electronic survey catalogs. Topics should be selected from a standard thesaurus, preferably an international, multilingual thesaurus. The IHSN recommends the use of the thesaurus used by the Council of European Social Science Data Archives (CESSDA). The CESSDA thesaurus has been introduced as a controlled vocabulary in the IHSN Study Template version 1.3 (available <a href="http://www.surveynetwork.org/toolkit">www.surveynetwork.org/toolkit</a>).</p>
Keywords	<p>Keywords summarize the content or subject matter of the survey. As topic classifications, these are used to facilitate referencing and searches in electronic survey catalogs. Keywords should be selected from a standard thesaurus, preferably an international, multilingual thesaurus. Entering a list of keywords is tedious. This option is provided for advanced users only.</p>
<b>Quality Statement</b>	
Standards Compliance	<p>This section lists all specific standards complied with during the execution of this study. Note the standard name and producer and how the study complied with the standard</p>
Other Quality Statement	<p>Enter any additional quality statements</p>
<b>Post Evaluation Procedures</b>	

Continued on next page

Table 1 – continued from previous page

Evaluator Type	<p>The evaluator element identifies persons or organizations involved in the evaluation. The Affiliation attribute contains the affiliation of the individual or organization. The Abbr. attribute holds an abbreviation for the individual or organization. The Role attribute indicates the role played by the individual or organization in the evaluation process.</p> <p>Example:</p> <ul style="list-style-type: none"> <li>• Affiliation: United Nations</li> <li>• Abbr.: UNSD</li> <li>• Role: Consultant</li> </ul>
Evaluation Process	Describes the evaluation process followed. Ex-Post Evaluations are frequently done within large statistical or research agencies, in particular when the survey is intended to be repeated or on-going
Evaluation Outcomes	Describe the outcomes of the evaluation
<b>Coverage</b>	
Country	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	<p>This field 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.</p> <p>Note that we do not describe here where the data was collected. For example, a sample survey could be declared as “national coverage” even in cases where some districts were not included in the sample, as long as the sampling strategy was such that the representativity is national.</p>
Geographic Unit	Lowest level of geographic aggregation covered by the data.
Universe	<p>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.</p> <p><b>Example:</b> <i>The survey covered all de jure household members (usual residents), all women aged 15-49 years resident in the household, and all children aged 0-4 years (under age 5) resident in the household.</i></p>

Continued on next page

Table 1 – continued from previous page

Geographic bounding box	The geographic bounding box is the minimum box, defined by west and east longitudes and north and south latitudes, that includes the largest geographic extent of the dataset's geographic coverage. This element is used in the first pass of a coordinate-based search. If the Geographic bounding Polygon element is included, then this field element <b>MUST</b> be included
Geographic Bounding Polygon	This field allows the creation of multiple polygons to describe in a more detailed manner the geographic area covered by the dataset. It should only be used to define the outer boundaries of a covered area. Example: In the United States, such polygons can be created to define boundaries for Hawaii, Alaska, and the continental United States, but not interior boundaries for the contiguous states. This field is used to refine a coordinate-based search, not to actually map an area
<b>Producers and Sponsors.</b>	
Authoring Entity/Primary investigators	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.
Producers	This field is provided to list other interested parties and persons that have played a significant but not the leading technical role in implementing and producing the data. The specific fields to be completed are: Name of the organization, Abbreviation, Affiliation and Role. If any of the fields are not applicable these can be left blank. The abbreviations should be the official abbreviation of the organization. The role should be a short and succinct phrase or description on the specific assistance provided by the organization in order to produce the data. The roles should be standard vocabulary such as: <ul style="list-style-type: none"> <li>• [Technical assistance in] questionnaire design</li> <li>• [Technical assistance in] sampling methodology / selection</li> <li>• [Technical assistance in] data collection</li> <li>• [Technical assistance in] data processing</li> <li>• [Technical assistance in] data analysis</li> </ul> Do not include here the financial sponsors.

Continued on next page

Table 1 – continued from previous page

Funding Agency/Sponsor	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 Identifications/ acknowledgements	This optional field can be used to acknowledge any other people and institutions that have in some form contributed to the survey.
<b>Sampling</b>	
Sampling procedure	<p>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:</p> <ul style="list-style-type: none"> <li>• Sample size</li> <li>• Selection process (e.g., probability proportional to size or over sampling)</li> <li>• Stratification (implicit and explicit)</li> <li>• Stages of sample selection</li> <li>• Design omissions in the sample</li> <li>• Level of representation</li> <li>• Strategy for absent respondents/not found/refusals (replacement or not)</li> <li>• Sample frame used, and listing exercise conducted to update it</li> </ul> <p>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.</p> <p>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.</p> <p><b>Example:</b> <i>5000 households were selected for the sample. Of these, 4996 were occupied households and 4811 were successfully interviewed for a response rate of 96.3%. Within these households, 7815 eligible women aged 15-49 were identified for interview, of which 7505 were successfully interviewed (response rate 96.0%), and 3242 children aged 0-4 were identified for whom the mother or caretaker was successfully interviewed for 3167 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%.</i></p>

Continued on next page

Table 1 – continued from previous page

Sample Frame Name	Sample frame describes the sampling frame used for identifying the population from which the sample was taken. Label and text describing the sample frame
Update of listing	Describes operations conducted to update the sample frame
Valid Period	Defines a time period for the validity of the sampling frame. Enter dates in YYYY-MM-DD format.
Custodian	Custodian identifies the agency or individual who is responsible for creating or maintaining the sample frame. Attribute affiliation provides the affiliation of the custodian with an agency or organization. Attribute abbr. provides an abbreviation for the custodian.
Use Statement	Sample frame use statement
Frame Unit	Provides information about the sampling frame unit. The attribute “is Primary” is boolean, indicating whether the unit is primary or not.
Reference Period	Indicates the period of time in which the sampling frame was actually used for the study in question. Use ISO 8601 date/time formats to enter the relevant date(s).
Sample Size	This element provides the targeted sample size in integer format. Attributes: Planned / Actual and Unit and Number.
Sample Size Formula	This element includes the formula that was used to determine the sample size
Stratification	Describe the Stratification (implicit and explicit) and the Variables identifying strata and PSU
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 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. This field can also in some cases be used to describe non-responses in population censuses.

Continued on next page

Table 1 – continued from previous page

Weighting	<p>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.</p> <p><b>Example:</b> <i>Sample weights were calculated for each of the data files.</i></p> <p><i>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.</i></p> <p><i>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.</i></p> <p><i>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.</i></p>
<b>Data Collection</b>	
Dates of data collection	<p>Enter the dates (at least month and year) of the start and end of the data collection. They should be in the standard ISO format of YYYY-MM-DD.</p> <p>In some cases, data collection for a same survey can be conducted in waves. In such case, you should enter the start and end date of each wave separately, and identify each wave in the “cycle” field.</p>
Collector Training	<p>Describes the training provided to data collectors including interviewer training, process testing, compliance with standards etc. This is repeatable for language and to capture different aspects of the training process. The type attribute allows specification of the type of training being described</p>
Frequency of Data Collection	<p>For data collected at more than one point in time, the frequency with which the data were collected. The “freq” attribute is included to permit the development of a controlled vocabulary for this element.</p>

Continued on next page



Table 1 – continued from previous page

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.
Data Sources	Used to list the book(s), article(s), serial(s), and/or machine-readable data file(s)—if any—that served as the source(s) of the data collection
Alternatives to data collection	Sources of data available / potentially considered
Mode of data collection	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. Use the drop-down button in the Toolkit to select one option. In most cases, the response will be “face to face interview”. But for some specific kinds of datasets, such as for example data on rain falls, the response will be different.
Data Capture	Where was data capture done (e.g., In the field or at the office) and when was data capture done. Also, describe the technology using for data capture (e.g. scanning, PDAs OR Web)

Continued on next page

Table 1 – continued from previous page

Notes on data collection	<p>This element is provided in order to document any specific observations, occurrences or events during data collection. Consider stating such items like:</p> <ul style="list-style-type: none"> <li>• Was a training of enumerators held? (elaborate)</li> <li>• Any events that could have a bearing on the data quality?</li> <li>• How long did an interview take on average?</li> <li>• Was there a process of negotiation between households, the community and the implementing agency?</li> <li>• Are anecdotal events recorded?</li> <li>• Have the field teams contributed by supplying information on issues and occurrences during data collection?</li> <li>• In what language was the interview conducted?</li> <li>• Was a pilot survey conducted?</li> <li>• Were there any corrective actions taken by management when problems occurred in the field?</li> </ul> <p><b>Example:</b> <i>The pre-test for the survey took place from August 15, 2006 - August 25, 2006 and included 14 interviewers who would later become supervisors for the main survey.</i></p> <p><i>Each interviewing team comprised of 3-4 female interviewers (no male interviewers were used due to the sensitivity of the subject matter), together with a field editor and a supervisor and a driver. A total of 52 interviewers, 14 supervisors and 14 field editors were used. Data collection took place over a period of about 6 weeks from September 2, 2006 until October 17, 2006. Interviewing took place everyday throughout the fieldwork period, although interviewing teams were permitted to take one day off per week.</i></p> <p><i>Interviews averaged 35 minutes for the household questionnaire (excluding salt testing), 23 minutes for the women's questionnaire, and 27 for the under five children's questionnaire (excluding the anthropometry). Interviews were conducted primarily in English and Mumbo-jumbo, but occasionally used local translation in double-Dutch, when the respondent did not speak English or Mumbo-jumbo.</i></p> <p><i>Six staff members of GenCenStat provided overall fieldwork coordination and supervision. The overall field coordinator was Mrs. Doe.</i></p>
--------------------------	--

Continued on next page

Table 1 – continued from previous page

Questionnaires	<p>This element is provided to describe the questionnaire(s) used for the data collection. The following should be mentioned:</p> <ul style="list-style-type: none"> <li>• List of questionnaires and short description of each (all questionnaires must be provided as External Resources)</li> <li>• In what language were the questionnaires published?</li> <li>• 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.</li> </ul> <p><b>Example:</b> <i>The questionnaires for the Generic MICS were structured questionnaires based on the MICS3 Model Questionnaire with some modifications and additions. A household questionnaire was administered in each household, which collected various information on household members including sex, age, relationship, and orphanhood status. The household questionnaire includes household characteristics, support to orphaned and vulnerable children, education, child labour, water and sanitation, household use of insecticide treated mosquito nets, and salt iodization, with optional modules for child discipline, child disability, maternal mortality and security of tenure and durability of housing. In addition to a household questionnaire, questionnaires were administered in each household for women age 15-49 and children under age five. For children, the questionnaire was administered to the mother or caretaker of the child. The women's questionnaire include women's characteristics, child mortality, tetanus toxoid, maternal and newborn health, marriage, polygyny, female genital cutting, contraception, and HIV/AIDS knowledge, with optional modules for unmet need, domestic violence, and sexual behavior. The children's questionnaire includes children's characteristics, birth registration and early learning, vitamin A, breastfeeding, care of illness, malaria, immunization, and anthropometry, with an optional module for child development. The questionnaires were developed in English from the MICS3 Model Questionnaires, and were translated into Mumbo-jumbo. After an initial review the questionnaires were translated back into English by an independent translator with no prior knowledge of the survey. The back translation from the Mumbo-jumbo version was independently reviewed and compared to the English original. Differences in translation were reviewed and resolved in collaboration with the original</i></p>
1.7. Completing metadata	<p><i>translation from the Mumbo-jumbo version was independently reviewed and compared to the English original. Differences in translation were reviewed and resolved in collaboration with the original</i></p>

Table 1 – continued from previous page

Instrument Development	Describe any development work on the data collection instrument. Type attribute allows for the optional use of a defined development type with or without use of a controlled vocabulary.
Review process for survey instrument	Description of the review process / list of agencies/people consulted
Pilot/testing of survey instrument and data collection	Description of pilot survey
Survey management team	Attributes: Name, Title, Agency, Role
Data collectors	<p>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, the Affiliation and the Role. 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. The role attribute specifies the role of person in the data collection process.</p> <p><b>Example:</b> <i>Abbreviation: CSO</i>  <i>Affiliation: Ministry of Planning</i>  <i>Role: Planner</i></p>
Compliance with international data collection standards	Describe if the survey comply with international survey recommendations

Continued on next page

Table 1 – continued from previous page

Supervision	<p>This element will provide information on the oversight of the data collection. The following should be considered:</p> <ul style="list-style-type: none"> <li>• Were the enumerators organized in teams that included a controller and a supervisor? With how many controllers/supervisors per interviewer?</li> <li>• What were the main roles of the controllers/supervisors?</li> <li>• Were there visits to the field by upper management? How often?</li> </ul> <p><b>Example:</b> <i>Interviewing was conducted by teams of interviewers. Each interviewing team comprised of 3-4 female interviewers, a field editor and a supervisor, and a driver. Each team used a 4 wheel drive vehicle to travel from cluster to cluster (and where necessary within cluster).</i></p> <p><i>The role of the supervisor was to coordinator field data collection activities, including management of the field teams, supplies and equipment, finances, maps and listings, coordinate with local authorities concerning the survey plan and make arrangements for accommodation and travel. Additionally, the field supervisor assigned the work to the interviewers, spot checked work, maintained field control documents, and sent completed questionnaires and progress reports to the central office.</i></p> <p><i>The field editor was responsible for reviewing each questionnaire at the end of the day, checking for missed questions, skip errors, fields incorrectly completed, and checking for inconsistencies in the data. The field editor also observed interviews and conducted review sessions with interviewers.</i></p> <p><i>Responsibilities of the supervisors and field editors are described in the Instructions for Supervisors and Field Editors, together with the different field controls that were in place to control the quality of the fieldwork.</i></p> <p><i>Field visits were also made by a team of central staff on a periodic basis during fieldwork. The senior staff of GenCenStat also made 3 visits to field teams to provide support and to review progress.</i></p>
Data Processing	

Continued on next page

Table 1 – continued from previous page

Data entry and editing	<p>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.</p> <ul style="list-style-type: none"> <li>• Was a hot deck or a cold deck technique used to edit the data?</li> <li>• Were corrections made automatically (by program), or by visual control of the questionnaire?</li> <li>• What software was used?</li> </ul> <p>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.</p> <p><b>Example:</b> <i>Data editing took place at a number of stages throughout the processing, including:</i></p> <ol style="list-style-type: none"> <li><i>Office editing and coding</i></li> <li><i>During data entry</i></li> <li><i>Structure checking and completeness</i></li> <li><i>Secondary editing</i></li> <li><i>Structural checking of SPSS data files</i></li> </ol> <p><i>Detailed documentation of the editing of data can be found in the “Data processing guidelines” document provided as an external resource.</i></p>
Software	<p>List of software used for the key activities (especially data entry, editing, tabulation, analysis).</p> <p>Attributes: Purpose and Software</p>

Continued on next page

Table 1 – continued from previous page

Other processing	<p>Use this field to provide as much information as possible on the data entry design. This includes such details as:</p> <ul style="list-style-type: none"> <li>• Mode of data entry (manual or by scanning, in the field/in regions/at headquarters)</li> <li>• Computer architecture (laptop computers in the field, desktop computers, scanners, PDA, other; indicate the number of computers used)</li> <li>• Software used</li> <li>• Use (and rate) of double data entry</li> <li>• Average productivity of data entry operators; number of data entry operators involved and their work schedule</li> </ul> <p>Information on tabulation and analysis can also be provided here.</p> <p>All available materials (data entry/tabulation/analysis programs; reports on data entry) should be listed here and provided as external resources.</p> <p><b>Example:</b> <i>Data were processed in clusters, with each cluster being processed as a complete unit through each stage of data processing. Each cluster goes through the following steps:</i></p> <ol style="list-style-type: none"> <li>1. <i>Questionnaire reception</i></li> <li>2. <i>Office editing and coding</i></li> <li>3. <i>Data entry</i></li> <li>4. <i>Structure and completeness checking</i></li> <li>5. <i>Verification entry</i></li> <li>6. <i>Comparison of verification data</i></li> <li>7. <i>Back up of raw data</i></li> <li>8. <i>Secondary editing</i></li> <li>9. <i>Edited data back up</i></li> </ol> <p><i>After all clusters are processed, all data is concatenated together and then the following steps are completed for all data files:</i></p> <ol style="list-style-type: none"> <li>10. <i>Export to SPSS in 4 files (hh - household, hl - household members, wm - women, ch - children under 5)</i></li> <li>11. <i>Recoding of variables needed for analysis</i></li> <li>12. <i>Adding of sample weights</i></li> <li>13. <i>Calculation of wealth quintiles and merging into data</i></li> <li>14. <i>Structural checking of SPSS files</i></li> <li>15. <i>Data quality tabulations</i></li> <li>16. <i>Production of analysis tabulations</i></li> </ol> <p><i>Details of each of these steps can be found in the data processing documentation, data editing guidelines, data processing programs in CPro and SPSS, and tabulation guidelines.</i></p> <p><i>Data entry was conducted by 12 data entry operators in tow shifts, supervised by 2 data entry supervisors, using a total of 7 computers (6 data entry computers plus one supervisors' computer). All data entry was conducted at the GenCenStat head office using manual data entry. For data entry, CPro version 2.6.007 was used with a highly structured data entry program, using system controlled approach that controlled entry of each variable.</i></p>
1.7. Completing metadata	<p>All range checks and skips were controlled by the program and operators could not override these. A limited set of consistency checks were also included in the data entry program. In addition, the calculation of anthropo-</p>

Table 1 – continued from previous page

Coding Instructions	
Coding Instructions Text	<p>Describe specific coding instructions used in data processing, cleaning, assessment, or tabulation. Use this field to describe instructions in a human readable form.</p> <p><b>Example:</b> <i>Due to an error in the data collection system the value of “27” was entered for the variable NBWFBPC which should be coded as an invalid value of “99”</i></p>
Command	<p>Provide command code for the coding instruction. The formalLanguage attribute identifies the language of the command code.</p> <p><b>Example:</b> <i>SPSS”&gt;RECODE V1 TO V100 (10 THROUGH HIGH = 0)</i></p>
Data Appraisal	

Continued on next page



Table 1 – continued from previous page

Estimate of sampling error	<p>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:</p> <ul style="list-style-type: none"> <li>• A list of ratios/indicators for which sampling errors were computed.</li> <li>• 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.</li> <li>• Reference to the reports or other document where the results can be found (to be provided as external resources).</li> </ul> <p><b>Example:</b> <i>Estimates from a sample survey are affected by two types of errors: 1) non-sampling errors and 2) sampling errors. Non-sampling errors are the results of mistakes made in the implementation of data collection and data processing. Numerous efforts were made during implementation of the 2005-2006 MICS to minimize this type of error, however, non-sampling errors are impossible to avoid and difficult to evaluate statistically. If the sample of respondents had been a simple random sample, it would have been possible to use straightforward formulae for calculating sampling errors. However, the 2005-2006 MICS sample is the result of a multi-stage stratified design, and consequently needs to use more complex formulae. The SPSS complex samples module has been used to calculate sampling errors for the 2005-2006 MICS. This module uses the Taylor linearization method of variance estimation for survey estimates that are means or proportions. This method is documented in the SPSS file CSDescriptives.pdf found under the Help, Algorithms options in SPSS.</i></p> <p><i>Sampling errors have been calculated for a select set of statistics (all of which are proportions due to the limitations of the Taylor linearization method) for the national sample, urban and rural areas, and for each of the five regions. For each statistic, the estimate, its standard error, the coefficient of variation (or relative error – the ratio between the standard error and the estimate), the design effect, and the square root design effect (DEFT – the ratio between the standard error using the given sample design and the standard error that would result if a simple random sample had been used), as well as the 95 percent confidence intervals (+/-2 standard errors).</i></p> <p><i>Details of the sampling errors are presented in the sampling errors appendix to the report and in the sampling errors table presented in the external resources.</i></p>
----------------------------	--

Continued on next page

Table 1 – continued from previous page

Other forms data appraisal	<p>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:</p> <ul style="list-style-type: none"> <li>• For a population census, information on the post enumeration survey (a report should be provided in external resources and mentioned here).</li> <li>• For any survey/census, a comparison with data from another source.</li> <li>• Etc.</li> </ul> <p><b>Example:</b> <i>A series of data quality tables and graphs are available to review the quality of the data and include the following:</i></p> <ul style="list-style-type: none"> <li>• <i>Age distribution of the household population</i></li> <li>• <i>Age distribution of eligible women and interviewed women</i></li> <li>• <i>Age distribution of eligible children and children for whom the mother or caretaker was interviewed</i></li> <li>• <i>Age distribution of children under age 5 by 3 month groups</i></li> <li>• <i>Age and period ratios at boundaries of eligibility</i></li> <li>• <i>Percent of observations with missing information on selected variables</i></li> <li>• <i>Presence of mother in the household and person interviewed for the under 5 questionnaire</i></li> <li>• <i>School attendance by single year age</i></li> <li>• <i>Sex ratio at birth among children ever born, surviving and dead by age of respondent</i></li> <li>• <i>Distribution of women by time since last birth</i></li> <li>• <i>Scatter plot of weight by height, weight by age and height by age</i></li> <li>• <i>Graph of male and female population by single years of age</i></li> <li>• <i>Population pyramid</i></li> </ul> <p><i>The results of each of these data quality tables are shown in the appendix of the final report and are also given in the external resources section.</i></p> <p><i>The general rule for presentation of missing data in the final report tabulations is that a column is presented for missing data if the percentage of cases with missing data is 1% or more. Cases with missing data on the background characteristics (e.g. education) are included in the tables, but the missing data rows are suppressed and noted at the bottom of the tables in the report (not in the SPSS output, however).</i></p>
Data Access	

Continued on next page

Table 1 – continued from previous page

Access authority	<p>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 <a href="mailto:data@popstatsoffice.org">data@popstatsoffice.org</a> whenever possible to avoid tying access to a particular individual whose functions may change over time.</p>
Confidentiality Declaration	<p>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.</p> <p>An example of statement could be the following:  <i>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:</i></p> <ol style="list-style-type: none"> <li><i>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.</i></li> <li><i>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.</i></li> <li><i>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.</i></li> </ol> <p><i>This statement does not replace a more comprehensive data agreement(see Access condition).</i></p>

Continued on next page

Table 1 – continued from previous page

Access conditions	<p>Each dataset should have an “Access policy” attached to it. The IHSN recommends three levels of accessibility:</p> <ul style="list-style-type: none"> <li>• Public use files, accessible to all</li> <li>• Licensed datasets, accessible under conditions</li> <li>• Datasets only accessible in a data enclave, for the most sensitive and confidential data.</li> </ul> <p>The IHSN has formulated standard, generic policies and access forms for each one of these three levels (which each country can customize to its specific needs). One of the three policies may be copy/pasted in this field once it has been edited as needed and approved by the appropriate authority. Before you fill this field, a decision has to be made by the management of the data depositor agency. Avoid writing a specific statement for each dataset.</p> <p>If the access policy is subject to regular changes, you should enter here a URL where the user will find detailed information on access policy which applies to this specific dataset. If the datasets are sold, pricing information should also be provided on a website instead of being entered here.</p> <p>If the access policy is not subject to regular changes, you may enter more detailed information here. For a public use file for example, you could enter information like:  <i>The dataset has been anonymized and is available as a Public Use Dataset. It is accessible to all for statistical and research purposes only, under the following terms and conditions:</i></p> <ol style="list-style-type: none"> <li>1. <i>The data and other materials will not be redistributed or sold to other individuals, institutions, or organizations without the written agreement of the [National Data Archive].</i></li> <li>2. <i>The data will be used for statistical and scientific research purposes only. They will be used solely for reporting of aggregated information, and not for investigation of specific individuals or organizations.</i></li> <li>3. <i>No attempt will be made to re-identify respondents, and no use will be made of the identity of any person or establishment discovered inadvertently. Any such discovery would immediately be reported to the [National Data Archive].</i></li> <li>4. <i>No attempt will be made to produce links among datasets provided by the [National Data Archive], or among data from the [National Data Archive] and other datasets that could identify individuals or organizations.</i></li> <li>5. <i>Any books, articles, conference papers, theses, dissertations, reports, or other publications that employ data obtained from the [National Data Archive] will cite the source of data in accordance with the Citation Requirement provided with each dataset.</i></li> <li>6. <i>An electronic copy of all reports and publications based on the requested data will be sent to the [National Data Archive].</i></li> </ol>
48	<p><b>Chapter 1. Content</b></p> <ol style="list-style-type: none"> <li>7. <i>The original collector of the data, the [National Data Archive], and the relevant funding agencies bear no responsibility for use of the data or for in-</i></li> </ol>

Table 1 – continued from previous page

Citation requirement	<p>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 (in the example above, see item 5). 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.</p> <p><b>Example:</b> “<i>National Statistics Office of Popstan, Multiple Indicators Cluster Survey 2000 (MICS 2000), Version 01 of the public use dataset (April 2001), provided by the National Data Archive. www.nda_popstan.org</i>”</p>
Location of Data Collection	Location where the data collection is currently stored. Use the URI attribute to provide a URN or URL for the storage site or the actual address from which the data may be downloaded.
URL for Location of Data Collection	Location where the data collection is currently stored. Provide a URN or URL for the storage site or the actual address from which the data may be downloaded.
Archive Where Study Originally Stored	Statement of collection availability. An archive may need to indicate that a collection is unavailable because it is embargoed for a period of time, because it has been superseded, because a new edition is imminent, etc. It is anticipated that a controlled vocabulary will be developed for this element.
Availability Status	Statement of collection availability. An archive may need to indicate that a collection is unavailable because it is embargoed for a period of time, because it has been superseded, because a new edition is imminent, etc. It is anticipated that a controlled vocabulary will be developed for this element.
Deposit Requirement	Information regarding user responsibility for informing archives of their use of data through providing citations to the published work or providing copies of the manuscripts.
<b>Disclaimer and Copyright</b>	
Disclaimer	<p>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:</p> <p><i>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.</i></p>

Continued on next page

Table 1 – continued from previous page

Copyright	Include here a copyright statement on the dataset, such as: © 2017, <i>Popstan Central Statistics Agency</i>
<b>Contacts</b>	
Contact persons	<p>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.</p> <p><b>Example:</b> <i>Name: Head, Data Processing Division</i>  <i>Affiliation: National Statistics Office</i>  <i>Email: dataproc@cso.org</i>  <i>URI: www.cso.org/databank</i></p>

### 1.7.3 Good practices for completing the File Description

The File Description is the DDI section that aims to provide a detailed description of each data file. The IHSN has selected six of the available DDI elements.

File Citation	Provides for a full bibliographic citation option for each data file described in File Description.
Contents of Files	<p>A data filename usually provides little information on its content. Provide here a description of this content. This description should clearly distinguish collected variables and derived variables. It is also useful to indicate the availability in the data file of some particular variables such as the weighting coefficients. If the file contains derived variables, it is good practice to refer to the computer program that generated it.</p> <p><b>Examples:</b></p> <ul style="list-style-type: none"> <li>• <i>The file contains data related to section 3A of the household survey questionnaire (Education of household members aged 6 to 24 years). It also contains the weighting coefficient, and various recoded variables on levels of education.</i></li> <li>• <i>The file contains derived data on household consumption, annualized and aggregated by category of products and services. The file also contains a regional price deflator variable and the household weighting coefficient. The file was generated using a Stata program named "cons_aggregate.do" available in the external resources.</i></li> </ul>
File Producer	Put the name of the agency that produced the data file. Most data files will have been produced by the survey primary investigator. In some cases however, auxiliary or derived files from other producers may be released with a data set. This may for example include CPI data generated by a different agency, or files containing derived variables generated by a researcher.
Version	A data file may undergo various changes and modifications. These file specific versions can be tracked in this element. This field will in most cases be left empty. It is more important to fill the field identifying the version of the dataset (see above).
Processing Checks	<p>Use this element if needed to provide information about the types of checks and operations that have been performed on the data file to make sure that the data are as correct as possible, e.g. consistency checking, wildcode checking, etc. Note that the information included here should be specific to the data file. Information about data processing checks that have been carried out on the data collection (study) as a whole should be provided in the "Data editing" element at the study level.</p> <p>You may also provide here a reference to an external resource that contains the specifications for the data processing checks (that same information may be provided also in the "Data Editing" filed in the Study Description section).</p>
Files Notes	This field, aiming to provide information to the user on items not covered elsewhere, will in most cases be left empty.

### **1.7.4 Good practices for completing the Variables Description**

The Variable Description is the section of the DDI document that provides detailed information on each variable.



VARIABLES	
Names	<p>These are the names given to the variables. Ideally, the variable names should be a maximum of 8 characters, and use a logical naming convention (e.g., section (S) and question (Q) numbers to name the question). If the variable names do not follow these principles, DO NOT CHANGE THE VARIABLE NAMES IN THE EDITOR, but make recommendations to the data processor for consideration for future surveys.</p>
Labels	<p>All variables should have a label that</p> <ul style="list-style-type: none"> <li>• Provides the item or question number in the original data collection instrument (unless item number serves as the variable name)</li> <li>• Provides a clear indication of what the variable contains</li> <li>• Provides an indication of whether the variable is constructed from other items</li> </ul> <p>Recommendations:</p> <ul style="list-style-type: none"> <li>• Do not use ALL CAPS in labels.</li> <li>• Make sure that different variables have different labels (avoid duplicate labels).</li> <li>• For expenditure or income: indicating the currency and period of reference is crucial (e.g. “Annual per capita real expenditure in local currency”)</li> </ul>
Width, StartCol, Endcol	<p>When you import your data files from Stata or SPSS, the information on StartCol and EndCol will be empty. It is crucial to add this information, in order to allow users to export the data to ASCII fixed format. To do so, use the “Variables &gt; Resequence” function in the Editor, for each data file.</p>
Data type	<p>Four types of variables are recognized by the Editor:</p> <ul style="list-style-type: none"> <li>• <i>Numeric</i>: Numeric variables are used to store any number, integer or floating point (decimals).</li> <li>• <i>Fixed string</i>: A fixed string variable has a predefined length (default length is 8 but it can range from 1 to 255 characters in length) which enables the publisher to handle this data type more efficiently.</li> <li>• <i>Dynamic string</i>: Dynamic string variables can be used to store open-ended questions.</li> <li>• <i>Date</i>: date variables stored in ISO format (YYYY-MM-DD?)—should specify</li> </ul> <p>The data type is usually properly identified when the data is imported. It is important to avoid the use of string variables when this is not absolutely needed. Such issues must be taken care of before the data is imported in the Editor. See the section on “1. Gathering and preparing the dataset” above.</p>
Measure	<p>The Metadata Editor will allow you to define the measure of a variable as:</p> <ul style="list-style-type: none"> <li>• <i>Nominal</i>: variable with numeric assignments for responses; the number assigned to each response does not have a meaning by itself.</li> </ul> <p><b>Example:</b> Variable <i>sex</i>: 1 = Male, 2 = Female (the number does not have a meaning by itself; w53 could as well have assigned Male = 2 and Female = 1). When variables are nominal, we can produce frequency tables by code, but calculating</p>
1.7. Completing metadata	

### **1.7.5 Good practices for completing the External Resources description**

The External Resources are all materials related to the study others than the data files. They include documents (such as the questionnaires, interviewer's manuals, reports, etc), programs (data entry, editing, tabulation, and analysis), maps, photos, and others. To document external resources, the Metadata Editor uses the Dublin Core metadata standard (which complements the DDI standard).

Label	This is the label that will be used to display a hyper link to the attached document. It can be the title, name, or an abbreviated version of the title.
Resource	The resource is used to point to the file that will be attached and distributed. The folder where the document is found is a relative path and should be the folder that will be pasted into the document path. Once you have pointed to the specified resource make sure you check file access by clicking the folder icon to the right of the entry field.
Type	<p>This is crucial information. A controlled vocabulary is provided. The selection of the type is important as it determines the way it will be presented or displayed to the user in the final output. The following are the choices:</p> <ul style="list-style-type: none"> <li>• Document Administrative: This includes materials such as the survey budget; grant agreement with sponsors; list of staff and interviewers, etc.</li> <li>• Document Analytical: Documents that present analytical output (academic papers, etc. This does not include the descriptive survey report (see below).</li> <li>• Document Questionnaire: the actual questionnaire(s) used in the field.</li> <li>• Document Reference: Any reference documents that are not directly related to the specific dataset, but that provide background information regarding methodology, etc. For international standard surveys, this may for example include the generic guidelines provided by the survey sponsor.</li> <li>• Document Report: Survey reports, studies and other reports that use the data as the basis for their findings.</li> <li>• Document Technical: Methodological documents related to survey design, interviewer's and supervisor's manuals, editing specifications, data entry operator's manual, tabulation and analysis plan, etc.</li> <li>• Document Other: Miscellaneous items</li> <li>• Audio: audio type files.</li> <li>• Map: Any cartographic information.</li> <li>• Photo: Photos can provide good documentary evidence of a survey.</li> <li>• Program: programs generated during data entry and analysis (data entry, editing, tabulation and analysis). These can be zipped together (include a brief summary report to describe the contents)</li> <li>• Table: Tabulations such as confidence intervals that may not be included in a general report.</li> <li>• Video: video type files provided as additional visual information</li> <li>• Website: Link to related website(s), such as a link to a Redatam server, or to the website of the survey sponsor in the case of international survey programs like the DHS, LSMS, or MICS).</li> <li>• Database: any databases related to the survey (e.g., a Devinfo database providing the aggregated results of the survey).</li> </ul>
<b>1.7. Completing metadata</b>	
Title	Full title of the document as it is provided on the cover

## 1.8 Creating variable groups

Variable groups are optional, but will help organize the data for the user into specific subject of use categories. This will be particularly useful to the user in the case of data files that contain many variables and are not organized by topic (some flat files contain hundreds or even thousands of variables).

The Metadata Editor allows you to group variables found in various separate data files. For example, education data may be found in various locations and the disparate variables grouped together. Also, a same variable can belong to more than one group.

Variable groups are “virtual”. The variables themselves are not moved or grouped. They remain untouched in the data files.

The variable groups will appear under a menu item “Data dictionary”. The only reason for grouping variables is to allow users to easily locate variables related to their topic of interest. If your dataset contains very few variables, there is no justification for grouping them.

If you decide to create variable groups (and sub-groups if needed), make sure that ALL variables in the dataset belong to at least one group.

Variable groups also have their own DDI elements which include Type, Label, Text, Definition, Universe, and Notes. These elements are optional and will in most cases be left empty.

Type	This is a controlled vocabulary field. It best identifies the manner the variables are grouped together. This field is optional.
Label	The label used to identify the group should be clear and relate to the type chosen. If these are grouped by subject, then the subject should be clearly stated etc.
Text	Include additional text to clarify the reason or purpose for grouping the variables. This field is optional.
Definition	This optional field is used to define the variable group.
Universe	This optional field defines the universe relevant to the selected grouped variables. The variables for example can be grouped as “Fertility Data” and the universe restricted to women between the ages of 15-49.
Notes	Additional space for further optional explanatory notes.

## 1.9 Running validations and diagnostics

The Metadata Editor includes a useful series of diagnostic and validation modules (see the drop down menu *Tools*): these range from very simple validations (such as the *Tools-Validate Metadata*) to complex visual displays that iterate through each variable and provides feedback to the archivist at the variable level.

- *Validate Metadata*: verifies that all mandatory fields are filled in.
- *Validate External Resources*: verifies all mandatory fields in the External Resources are filled in.
- *Health Check*: displays a popup window that provides some information and diagnostics regarding the R package. The Metadata Editor uses R to import and export the data. The health check option tells the which version of R is being used on the machine. In addition, the Health Check will provide information on the Environment Path and the result of the execution of the R script.
- *Validate Dataset Relations*: this option is used to validate hierarchically related datasets. The ‘Base key variables’ should be the variables that uniquely identify a case within that file.
- *Translation Manager*: provides the user with the ability to translate the interface of the Metadata Editor into any language. Selecting this option will display the Translation Manager interface. When using the option for the first time, the English template will be displayed with all the labels that require translation.

In addition to these validations, it is recommended that you generate the DDI document (in the Editor, use the Export DDI" function) and verify the size of the resulting [.xml] file. A fully documented survey with a large number of variables should not produce a file larger than 10Mb. Very large DDI files often indicate errors in the selection of summary statistics (for example, frequencies are produced for a variable like the household ID in a sample household file).

## 1.10 Generating the survey documentation in PDF

Once you are confident that all necessary checks have been completed, you may generate the survey documentation in PDF. The Metadata Editor includes a useful tool for producing a PDF document summarizing all metadata entered in the Editor (see Documentation > PDF Documentation). Generating this report is one of the final stages of properly preparing a survey for publication and dissemination. If previous versions exist and changes have been made to the data files or the metadata make sure you re-run the PDF generator.

This report should be generated, saved and attached as an *External Resource*.

! The PDF report will include a list of all external resources related to the study. This list should include this PDF report itself. **Before** you generate it, make sure you create one entry in the External Resources for documenting this report. Immediately after you generate the PDF report, import it in the Editor.

One thing to keep in mind is that in a survey with a large number of variables may produce a document that is very long. If the report is in excess of 300 or 350 pages, you may want to split this report (e.g., produce one report with the study metadata, and one with the files and variables metadata), or change the content options (e.g., not including a frequency table for all variables).

If your agency has a website, you may upload this PDF directly to the web server. The IHSN recommends the use of a proper DDI-compliant cataloguing system, such as the one provided by its National Data Archive (NADA) application. NADA is an open source package, available free of charge at [www.surveynetwork.org](http://www.surveynetwork.org).

## 1.11 Independent quality review

An independent review of the data and metadata is highly recommended prior to publishing the final output. The Appendix provides a blank review form (the *DDI Reviewer's Feedback Form*) to be used by an external reviewer.

The external review can be based on:

1. The DDI file (xml file, containing no microdata and no external resources)
2. The Metadata Editor project (containing microdata and DDI/DCMI metadata)
3. The PDF Document
4. The Microdata

The preferred option is to review the metadata and microdata, as it allows a full check of the final output. If data are highly confidential and cannot be shared with the reviewer, options 1-3 are the most appropriate.

In order to prepare for the independent quality review, proceed to step 10 if you will use options three or four. Follow the guidance there, and then finalize the archiving before producing the final output. Else, send the DDI-XML to the reviewer.

## 1.12 Section A . Data Validations in Stata: Practical Examples

### Example 1 . Check for unique identifiers (single variable)

This example uses a Country Opinion Survey and uses the Stata command `-isid-` to check whether the variable “id” uniquely identifies the observations. Each row of the data represents a different Country’s Stakeholders and the variable that identifies each one is named “id”.

\*Example using a Country Opinion Survey data

	id	a2_1	a2_2	a2_3	a2_4	a2_5	a2_6	a2_7	a2_8
1	101	0	0	0	0	1	0	0	0
2	102	1	0	0	1	0	0	0	0
3	103	.	.	.	.	.	.	.	.
4	104	0	0	0	0	0	0	0	0
5	105	0	0	0	0	0	0	0	0
6	106	0	0	0	1	0	0	0	0
7	107	0	0	1	0	1	0	0	0
8	108	0	0	1	0	1	0	0	0
9	109	0	0	1	0	1	0	0	0
10	110	0	0	0	0	0	0	0	0
11	111	0	0	0	1	1	0	0	0
12	112	0	0	0	1	0	0	0	0
13	113	0	0	0	1	0	0	0	0
14	114	0	0	0	1	0	0	0	0
15	115	0	0	0	1	0	0	0	0

#### STATA COMMAND

```
isid id
```

#### RESULTS

```
<blank>
```

If, after running the `-isid-` command you have not got an error message, it indicates that the “id” is unique and identifies each unit of analysis.

### Example 2 . Check for unique identifiers (single variable)

For this example, we use the same data than *Example 1*, but in this case, there are some hypothetical observations with the same values for the variable “id”. The highlight observations are the duplicates IDs.

\*Example using a Country Opinion Survey data

	id	a1	a2_1	a2_2	a2_3	a2_4	a2_5	a2_6	a2_7	a2_8
1	101	3	0	0	0	0	1	0	0	0
2	102	1	1	0	0	1	0	0	0	0
3	103	.	.	.	.	.	.	.	.	.
4	104	3	0	0	0	0	0	0	0	0
5	105	3	0	0	0	0	0	0	0	0
6	106	3	0	0	0	1	0	0	0	0
7	104	3	0	0	0	0	0	0	0	0
8	108	1	0	0	1	0	1	0	0	0
9	109	3	0	0	1	0	1	0	0	0
10	101	3	0	0	0	0	1	0	0	0
11	111	1	0	0	0	1	1	0	0	0
12	112	1	0	0	0	1	0	0	0	0
13	113	.	0	0	0	1	0	0	0	0
14	114	1	0	0	0	1	0	0	0	0
15	111	1	0	0	0	1	1	0	0	0

#### STATA COMMAND

```
isid id
```

#### RESULTS

variable id does not uniquely identify the observations

Since there are not unique IDs in the data, it is also useful to see the list of all duplicates. To do that, we can use the Stata command *-duplicates list-*.

#### STATA COMMAND

```
duplicates list id
```

#### RESULTS

Duplicates in terms of id

```
+-----+
|group:  obs:   id |
+-----+
|1         1    101|
|1        10    101|
|2         4    104|
|2         7    104|
|3        11    111|
|3        15    111|
+-----+
```

#### Example 3 . Check for unique identifiers (ID made of multiple variables)

The Multiple Indicator Cluster Survey at Women-level data is used in this example. According to the study's metadata, the unique identification of each woman is the combination of variables HH1 (Cluster Number), HH2 (Household Number) and Ln (Line Number of women), so instead of checking the unique identifier in just one variable, we are checking this condition in this group of variables.

\* Example using a Multiple Indicator Cluster Survey  
(Women-level data set)

	HH1	HH2	ln	WM1	WM2	WM5	WM6D	WM6M	WM6Y
1	1	1	2	1	1	12	5	december	2016
2	1	1	7	1	1	12	5	december	2016
3	1	3	2	1	3	12	6	december	2016
4	1	4	2	1	4	12	6	december	2016
5	1	5	5	1	5	14	7	december	2016
6	1	5	6	1	5	14	7	december	2016
7	1	5	7	1	5	12	7	december	2016
8	1	7	2	1	7	15	6	december	2016
9	1	7	3	1	7	15	6	december	2016
10	1	8	2	1	8	14	7	december	2016
11	1	9	2	1	9	14	6	december	2016
12	1	10	1	1	10	14	6	december	2016
13	1	11	2	1	11	14	6	december	2016
14	1	11	4	1	11	14	6	december	2016
15	1	12	1	1	12	14	6	december	2016

## STATA COMMAND

```
Isid HH1 HH2 ln
```

## RESULTS

```
<blank>
```

After running this validation, it is possible to see that the combination of “HH1”, “HH2” and “ln” generates a unique ID.

## Example 4 . Check for duplicate observations

This example uses the dataset mentioned in example 2, in which there are 3 duplicated identifiers.

	id	a1	a2_1	a2_2	a2_3	a2_4	a2_5	a2_6	a2_7	a2_8
1	101	3	0	0	0	0	1	0	0	0
2	102	1	1	0	0	1	0	0	0	0
3	103	.	.	.	.	.	.	.	.	.
4	104	3	0	0	0	0	0	0	0	0
5	105	3	0	0	0	0	0	0	0	0
6	106	3	0	0	0	1	0	0	0	0
7	104	3	0	0	0	0	0	0	0	0
8	108	1	0	0	1	0	1	0	0	0
9	109	3	0	0	1	0	1	0	0	0
10	101	3	0	0	0	0	1	0	0	0
11	111	1	0	0	0	1	1	0	0	0
12	112	1	0	0	0	1	0	0	0	0
13	113	.	0	0	0	1	0	0	0	0
14	114	1	0	0	0	1	0	0	0	0
15	111	1	0	0	0	1	1	0	0	0

## STATA COMMAND

```
duplicates report
```

## RESULTS

```
+-----+
|copies  observations surplus|
+-----+
|1          446          0|
|2           6           3|
+-----+
```



As shown, 446 records are unique in the database, but there are six observations for which there are two copies of each one. Now, it is necessary to identify duplicates. The code below allows one to generate a variable that tags the duplicates with a value 1 or more, depending on the number of times a record is duplicated.

**CODE**

```
duplicates tag, generate(duplicates)
tabulate id if duplicates>0
```

**RESULTS**

```
+-----+
|id      Freq.    Percent    Cum. |
+-----+
|101      2        33.33     33.33 |
|104      2        33.33     66.67 |
|111      2        33.33    100.00 |
+-----+
|Total    6         100.00      |
+-----+
```

The table above shows that records with the IDs 101, 104 and 111 each have one duplicate.

**Example 5 . Check the merge between datafiles**

The code below helps us combine the data collected at the individual-level with the data collected at the household-level. In this case, we have two hierarchical datasets, in the household data each row represents one household, and each household has members or individuals. So, we need to combine many observations from one data set (individual-level) with one observation from the other (household-level). The ID of the household data set is the unique identifier that we are using for the merge (“HH1”and “HH2”).

\* Example using a Multiple Indicator Cluster Survey

Individual-level data set (hl)

	HL1	HL2	HL1	HL3	HL4	HL5M	HL5Y
1	1	1	1	Head	Male	dk	1997
2	1	1	2	Spouse...	Female	May	1968
3	1	1	3	Son /...	Male	dk	1970
4	1	1	4	Son /...	Male	dk	1990
5	1	1	5	Son /...	Male	dk	1990
6	1	1	6	Son /...	Male	March	1997
7	1	1	7	Son /...	Female	February	2001
8	1	1	8	Son /...	Male	Sept...	2007
9	1	2	1	Head	Male	dk	dk
10	1	3	1	Head	Male	dk	dk
11	1	3	2	Spouse...	Female	March	1990
12	1	3	3	Son /...	Male	March	2014
13	1	3	4	Son /...	Female	October	2015
14	1	4	1	Head	Male	July	1978
15	1	4	2	Spouse...	Female	February	1990

Household-level data set (hh)

	HH1	HH2	HH3	HH4	HH5D	HH5M	HH5Y
1	1	1	12	11	8	December	2016
2	1	2	12	11	7	December	2016
3	1	3	12	11	6	December	2016
4	1	4	12	11	6	December	2016
5	1	5	13	11	6	December	2016
6	1	6	13	11	5	December	2016
7	1	7	13	11	5	December	2016
8	1	8	13	11	5	December	2016
9	1	9	14	11	5	December	2016
10	1	10	14	11	5	December	2016
11	1	11	14	11	6	December	2016
12	1	12	14	11	6	December	2016
13	1	13	15	11	6	December	2016
14	1	14	15	11	5	December	2016
15	1	15	15	11	5	December	2016

**CODE**

```
use "hl.dta"
merge m:1 HH1 HH2 using "hh.dta"
```

**RESULTS**

```
+-----+
|Result      # of obs. |
+-----+
|not matched      0 |
|matched      172,369 | (_merge==3)
+-----+
```

The report shows that all observations in the Individual file have a corresponding household in the household data set and that all households have at least one member. However, let's consider a hypothetical example that contains some records that do not match, below the results of the merge:

## RESULTS (hypothetical)

Result	# of obs.	
not matched	17	
from master	15	( _merge==1)
from using	2	( _merge==2)
matched	172,354	( _merge==3)

The merge command resulted in: 15 nonmatched observations originated from the master data and 2 from the using data. The inconsistencies between databases could be the result of a data entry error or processing errors and these also need to be referred to the data producer before documentation begins.

## Example 6 . Check variables full of missing values

In the example above, there are 3 variables (WM1\_1, WM3\_1 and WM5\_1) full of missing values. As shown in the table, the *-misstable summarize-* command allows one to identify all those cases at once. It is like tabulating every single variable to identify missing values but more efficiently.

\* Example using a hypothetical Multiple Indicator Cluster Survey

```
CODE
use "wm.dta"
misstable summarize
```

RESULTS OF THE CODE						Table of frequencies			
Variable	Obs=.	Obs>.	Obs<.	Unique values	Min	Max			
WM1_1	36,176	0	0	0	.	.			
WM3_1	36,176	0	0	0	.	.			
WM5_1	36,176	0	0	0	.	.			
WM10H	1,800	34,376	18	6	23				
WM10M	1,800	34,376	60	0	59				
WM11H	1,800	34,376	18	6	23				
WM11M	1,800	34,376	60	0	59				
WB1M	1,800	34,376	14	1	99				
WB1Y	1,800	34,376	39	1966	9999				
WB2	1,800	34,376	35	15	49				
WB3	1,800	34,376	3	1	9				
WB4	9,054	27,122	5	0	4				
WB5	13,854	22,322	22	0	36				
WB7	18,908	17,268	6	1	9				
MT2	15,353	20,823	5	1	9				
MT3	1,800	34,376	5	1	9				
MT4	1,800	34,376	5	1	9				
MT6	23,650	12,526	3	1	9				
MT7	133,794	2,382	2	1	2				

tab WM1 1,m			
Variable	#1	Freq.	Percent
hypothetical	.	36,176	100.00
Total		36,176	100.00

tab WM3 1,m			
Variable	#3	Freq.	Percent
hypothetical	.	36,176	100.00
Total		36,176	100.00

tab MT6 ,m			
Ever used a	computer	Freq.	Percent
Yes		2,382	6.58
No		10,103	27.93
Missing		41	0.11
Total		23,650	65.37

## Example 7 . Check the completeness of the data files

This example shows how you can check for discrepancies (if any) between the variables from the MICS women's questionnaire and the data set.

\* Example using a Multiple Indicator Cluster Survey  
(Women-level data set)

STATA COMMAND  
use "wm.dta"  
describe

RESULTS					MICS women's questionnaire	
variable name	storage type	display format	value	variable label		
WB1	int	%8.0g		Cluster number		
WB2	byte	%8.0g		Household number		
WB3	byte	%8.0g		Line Number of Women		
WB4	int	%8.0g		Cluster number		
WB5	byte	%8.0g		Household number		
WB6	int	%8.0g		Interviewer number		
WB7	byte	%8.0g		Day of interview		
WB8	byte	%8.0g		Month of interview		
WB9	int	%8.0g		Year of interview		
WB10	byte	%8.0g		Result of woman's interview		
WB11	byte	%8.0g		Start of interview - Hour		
WB12	byte	%8.0g		Start of interview - Minutes		
WB13	byte	%8.0g		End of interview - Hour		
WB14	byte	%8.0g		End of interview - Minutes		
WB15	byte	%8.0g		Month of birth of woman		
WB16	int	%8.0g		Year of birth of woman		
WB17	byte	%8.0g		Age of woman		
WB18	byte	%8.0g		Ever attended school		
WB19	byte	%8.0g		Highest level of school attended		
WB20	byte	%8.0g		Highest grade completed at that level		
WB21	byte	%8.0g		Can read part of the sentence		
WB22	byte	%8.0g		Frequency of reading newspaper or magazine		
WB23	byte	%8.0g		Frequency of listening to the radio		
WB24	byte	%8.0g		Frequency of watching TV		
WB25	byte	%8.0g		Ever used a computer		
WB26	byte	%8.0g		Computer usage in the last 12 months		

As shown, this dataset contains all variables from the section “Woman’s background,” and they are organized according to the questionnaire. The comparison between variables in the questionnaire to those in the data set should be made for every section in the questionnaire.

### Example 8 . Check all variables are labelled

\* Example using a Multiple Indicator Cluster Survey  
(Women-level data set)

STATA COMMAND  
use "wm.dta"  
labelbook

RESULTS					MICS women's questionnaire	
value label	WB14					
values	range: (11,99)	display	length: (5,3)			
%	12	unique at full length:	yes			
order	yes	unique at length 12:	no			
missing	*4 0	leading/trailing blanks:	no			
		numeric >= maximum:	no			
definition						
(1) Independent's home						
(2) Other home						
(3) Government hospital						
(4) Government clinic / health centre						
(5) Government health post						
(6) Other public						
(7) Private hospital						
(8) Private clinic						
(9) Other private medical						
(10) Other						
(11) Missing						
variables:	WB14					
value label	WB16					
values	range: (1,9)	display	length: (2,1)			
%	3	unique at full length:	yes			
order	yes	unique at length 12:	yes			
missing	*4 0	leading/trailing blanks:	no			
		numeric >= maximum:	no			
definition						
(1) Yes						
(2) No						
(3) Missing						

The following Form is available at [www.surveynetwork.org](http://www.surveynetwork.org)



## International Household Survey Network DDI Reviewers' Feedback Form

Country:	Language:
Dataset name:	
Dataset ID:	
Submitted by:	Date submitted:
Format provided: <input type="checkbox"/> DDI in XML <input type="checkbox"/> Nesstar file <input type="checkbox"/> Toolkit CD-ROM    DDI file size: _____ Mb	
Data provided? <input type="checkbox"/> Yes <input type="checkbox"/> No	External resources provided? <input type="checkbox"/> Yes <input type="checkbox"/> No
Reviewed by:	Review date (yyyy/mm/dd): ____ / ____ / ____
IHSN Study template used? <input type="checkbox"/> Yes <input type="checkbox"/> No	IHSN External Resource template used? <input type="checkbox"/> Yes <input type="checkbox"/> No
Has a new DDI been produced by the reviewer? <input type="checkbox"/> Yes (name: _____) <input type="checkbox"/> No	

### DOCUMENT DESCRIPTION

Document Description				
DDI Element	Expected	Status	Reviewer's comments	Action
Study Title	Proper noun format, years separated by hyphen	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Metadata producer	Name of the person and affiliation.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Date of production	Date in ISO format.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
DDI Document Version	Version number based on a standard naming convention.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
DDI Document ID Number	Number coherent with the Study Description ID Number.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

## STUDY DESCRIPTION

STUDY DESCRIPTION - Identification				
DDI Element	Expected	Status	Reviewer's comments	Action
Title	Full name of the survey, including the reference year. Proper noun format, years (if more than one) separated by hyphen. Example: <i>Household Budget Survey 2006-2007</i>	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Sub-title	In most cases, there will be no subtitle. If there is one, it should provide additional information related to the title.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Abbreviation	All capitalized; includes the reference year. Example: <i>DHS 2004</i>	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Study type	Preferably taken from a controlled vocabulary.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Series information	Clear description of the series (objectives, ownership, scope and coverage, period) and indication on how many rounds/surveys belong to the series.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Translated title	Optional (will be empty in most cases). Make sure special characters are readable.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
ID Number	Clear and consistent (based on a standard convention). Should include country abbreviation and year. Example: <i>UGA-UBOS-DHS-2004</i>	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

STUDY DESCRIPTION – Version				
DDI Element	Expected	Status	Reviewer's comments	Action
Description	Version number based on naming convention; should include a "label". Examples: <i>V0 – Raw data, unedited</i> <i>V1.1 – Edited non-anonymized data</i> <i>V2.2 – Public use dataset, 2<sup>nd</sup> release (Nov. 2007)</i>	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Production date	Date in ISO format (at least month and year)	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Notes	More information on what distinguishes this version from any other.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

STUDY DESCRIPTION - Overview				
DDI Element	Expected	Status	Reviewer's comments	Action
Abstract	Clear and concise abstract providing summary information of survey objectives, scope and coverage; and where applicable key findings of the survey.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Kind of data	Taken from the IHSN controlled vocabulary.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Unit of analysis	Usually stated as person, household, enterprise etc. (could be several units)	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

STUDY DESCRIPTION - Scope				
DDI Element	Expected	Status	Reviewer's comments	Action
Description of scope	Typically, list of questionnaire modules.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Keywords	In the future: list of keywords based on an international multilingual thesaurus	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Topics classification	In the future: list of topics based on an international multilingual thesaurus	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

STUDY DESCRIPTION - Coverage				
DDI Element	Expected	Status	Reviewer's comments	Action
Country	Country name in full	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix
Geographic coverage	Clear and concise statement of geographic coverage. Examples: - National, except province of ... - Rural only	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Universe	Population covered by the survey. This should almost never be "All population". A census for example does not cover diplomats. A household survey typically does not cover community households, homeless, and nomads.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

STUDY DESCRIPTION - Producers and Sponsors				
DDI Element	Expected	Status	Reviewer's comments	Action
Primary investigator	Full name of the agency that coordinated the data collection activities.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Other producers	Agencies that were not in charge, but participated in the implementation of the study as co-producer.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Funding	List of donors (national and international; having provided cash or in-kind contributions); national government should not be forgotten.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Other acknowledgment	Optional: acknowledgments of technical experts or others (persons or agencies) who contributed to the success of the operation.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix

STUDY DESCRIPTION – Sampling				
DDI Element	Expected	Status	Reviewer's comments	Action
Sampling procedure	Sample size, stratification, information on sample frame, replacement policy. Identification of the variables that represent for stratum, psu in the data files. For a census, this will be N.A. Reference to more detailed information in external resource. Verify that the sample size corresponds to what is found in the data files.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Deviation from sample design	Information on discrepancies between planned and actual sample. This may be inaccessibility of regions at the time of survey (insecurity or climatic issues), budget problems, etc. For a census, this will be N.A.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

Response rates	Response rates by stratum and stated as a percentage of the design. For a census, this will be N.A.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Weighting	Information on the weight variables available in the data files. If self-weighted, this must be explicitly stated here.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

#### STUDY DESCRIPTION - Data Collection

DDI Element	Expected	Status	Reviewer's comments	Action
Dates of data collection	Dates in ISO format: YYYY-MM-DD. At least month and year.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Time periods	In most cases, this will be empty.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Mode of data collection	Taken from IHSN controlled vocabulary.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Notes on data collection	Information on number and profile of interviewers and supervisors; on their training; observations on particular issues during data collection.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Questionnaires	List of questionnaire(s) and their content.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Data collectors	Optional. We do not expect a list of interviewers here.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix

Supervision	Clear description of field supervisory structure including: team size, control mechanisms etc.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
-------------	--	---	---	---

#### STUDY DESCRIPTION - Data Processing

DDI Element	Expected	Status	Reviewer's comments	Action
Data editing	Statement on method and software used. Ideally, provide a reference to external resources (documents/programs).	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Other processing	Statement on method and software used for data entry, tabulation and analysis (e.g., data entry in the field or at HQ; manual or by scanning; percentage of double entry). Ideally, provide a reference to external resources.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

#### STUDY DESCRIPTION - Data Appraisal

DDI Element	Expected	Status	Reviewer's comments	Action
Estimates of sampling errors	Summary statement on the reliability of the data is clearly stated with reference to the tests that have been run to check the variance. A link to an external resource that documents the procedure and software used is recommended when this has been done.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Other forms of data appraisal	Statement is clear and comprehensive.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

## STUDY DESCRIPTION - Data Access

DDI Element	Expected	Status	Reviewer's comments	Action
Access authority	Full name of the agency (or person) which (who) has authority to grant access to the data.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Confidentiality	Standard statement that serves for all surveys. Can use a customized version of the IHSN recommended statement.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Access conditions	One of the three standard statements to be adopted by the country (public use file / licensed file / confidential file). Can use a customized version of the IHSN recommended statement.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Citation requirements	Citation based on a standard format. Must include the following components: name of the dataset, version if available, producer, country, reference year(s).	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

## STUDY DESCRIPTION - Disclaimer and Copyright

DDI Element	Expected	Status	Reviewer's comments	Action
Disclaimer	Standard statement that serves for all surveys. Can use a customized version of the IHSN disclaimer statement.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Copyright	Standard format (Year, copyright statement).	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

## STUDY DESCRIPTION - Contacts

DDI Element	Expected	Status	Reviewer's comments	Action
Contact persons	Name and/or title of the person(s) who can provide more information on the survey. Preferably, do not use names (use title and agency).	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided <input type="checkbox"/> N.A.	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

## DATA FILES

### DATA FILES - File checks

Item	Expected	Status	Reviewer's comments	Action
Order of appearance	Data files are listed in a logical order.	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> None <input type="checkbox"/> Fix
Relationships validated	All files have numeric key variables that provide a unique ID, and the variables are validated in Toolkit (or in statistical package)	<input type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> N.A.		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix
Completeness	Data are available for all sections of all questionnaires; derived files are available as well.	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> None <input type="checkbox"/> Fix <input type="checkbox"/> Check
Re-sequence data	Position and length of all variables is available (required for ASCII export).	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> None <input type="checkbox"/> Add

### DATA FILES - File Description

DDI Element	Expected	Status	Reviewer's comments	Action
Name	File name should not be changed. No action to be taken. Reviewer can however formulate recommendation if file naming convention is not appropriate.	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> None



Content	Short but clear description of the file content. Ideally, should establish the link with questionnaire sections. Example: <i>Section 3A of the Household questionnaire: Education.</i>	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Producer	In most cases, the producer of the file is the producer of the survey.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Version	In most cases, there will be no versioning of individual file (as we have a version of the dataset).	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Processing Checks	Optional. In most cases, information on data editing will be contained at the Study Level.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Missing data	Optional. The variable description provides fields to describe missing values used for each variable.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Notes	Optional.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

## VARIABLES

### VARIABLES - Variable checks

DDI Element	Expected	Status	Reviewer's comments	Action
Variable Names	Variable names should not be changed. No action is expected. Reviewers can however comment if variable naming does not follow good practice.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None
Variable Labels	All variables should have a unique, clear label. Use the "Validate Variable" tool in the Toolkit to check.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Categories	All nominal variables have value labels. Use the "Validate Variable" tool in the Toolkit to check. If data are available, view the "Data Entry" page in the Toolkit. Entries in blue fonts are problems.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Statistics Options	Options should be properly set; frequencies should not be produced for variables such as household ID or similar (large-size DDI files often indicates such error).	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Weights	The application of weights is optional, but good to have when applicable. The appropriate weight must be applied to each variable.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None <input type="checkbox"/> N.A.		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Data type	Most variables should be numeric. String variables should be recoded, encoded, or de-string when possible.			<input type="checkbox"/> None <input type="checkbox"/> Fix <input type="checkbox"/> Check
Measure	Importing data from some formats in the Toolkit will not automatically impute the most appropriate measure.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Is Time variable	Rarely used.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Min	If data is available, check that no data are out of range by viewing the "Data entry" page in the Toolkit. Out of range values will appear in red.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Max	If data is available, check that no data are out of range by viewing the "Data entry" page in the Toolkit. Out of range values will appear in red.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

Decimals		<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Implicit decimals	Needed when the files were imported from ASCII.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Missing data	By default, missing data are indicated by ". If values have been used to indicate missing (e.g., 9999) this must be specified here. Missing values should NOT be declared in the Categories.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

VARIABLES - Description				
DDI Element	Expected	Status	Reviewer's comments	Action
Definition	Will be empty in most cases. For household surveys, make sure that we have a definition of "Household" attached to the hhjd variable.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Universe	Ideally, universe should be indicated for each variable. It can be in words (e.g. "Household members aged 15 and over", or in logical terms based on variables (e.g.: A05> 15 and A07=1)	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Source of information	Empty in most cases.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Concepts	Empty in most cases, except ones defined in the enumerator's or controller's manuals.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

VARIABLES - Question				
DDI Element	Expected	Status	Reviewer's comments	Action
Pre-question text	Pre-question text are instructions to interviewers provided <u>in the questionnaire</u> , prior to asking the question.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Literal question	All questions in the questionnaires attached to the corresponding variables.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Post-question text	Pre-question text are instructions to interviewers provided <u>in the questionnaire</u> , after the question is asked. Can include instructions on skips.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Interviewer Instructions	All instructions available in the interviewer's manual should be provided here.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

VARIABLES - Imputation and derivation				
DDI Element	Expected	Status	Reviewer's comments	Action
Imputation	Empty in most cases.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Recoding and derivation	Information should be provided for all calculated variables.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

VARIABLES - Others				
DDI Element	Expected	Status	Reviewer's comments	Action
Security	Empty in most cases. If available, should indicate a level of confidentiality.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Notes	Optional. Will be empty in most cases.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

VARIABLES - Variable groups				
Item	Expected	Status	Reviewer's comments	Action
Variable groups	Optional. If groups are provided, they should cover all variables in the file.	<input type="checkbox"/> Provided <input type="checkbox"/> Not provided		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

## EXTERNAL RESOURCES

EXTERNAL RESOURCES - External resources checks				
Item	Expected	Status	Reviewer's comments	Action
Questionnaire	All questionnaires must be provided in PDF format (and in original format as well, if possible)	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Resource (links)	All links to external resources should be valid (no broken links). Links should be relative addresses (no absolute paths).	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

PDF documentation	Technical document generated using the IHSN Toolkit PDF Generator is provided and documented in the external resources	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Programs	Data entry, editing, tabulation and analysis programs should be preserved and provided.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Report	All survey reports and analytical output must be provided in PDF (and in original format if available).	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Label	All external resources should have a short but explicit label.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

EXTERNAL RESOURCES - Identification				
DDI Element	Expected	Status	Reviewer's comments	Action
Type	All documents must have a "Type" indicated, taken from the IHSN controlled vocabulary. Make sure we have at least one "Questionnaire" and one "Report".	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Title	For documents, title as it appears on the cover page. For programs/photos/maps etc, a short title describing the content should be provided.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Subtitle	Optional. For documents only. Should correspond to what is displayed on the cover page.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Author(s)	All resources should have an author (person or agency).	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

Date	At least month and year.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Country	Country to which the resource is related.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Language	For documents only.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix
Format	Based on the IHSN controlled vocabulary.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
ID Number	Optional	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

## EXTERNAL RESOURCES - Contributors and rights

DDI Element	Expected	Status	Reviewer's comments	Action
Contributor(s)	Optional	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Publisher(s)	Optional	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Rights	Optional but recommended.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

## EXTERNAL RESOURCES - Content

DDI Element	Expected	Status	Reviewer's comments	Action
Description	<u>Short description</u> of the resource. Very important for computer programs (must describe the purpose, software needed to run it).	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Abstract	For documents only; optional.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Table of Contents	Optional. No need to include page numbers.	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Subjects	Optional (should be based on an international multilingual thesaurus)	<input type="checkbox"/> All <input type="checkbox"/> Some <input type="checkbox"/> None	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check

## CD-ROM / Website (only in cases where Toolkit-generated CD-ROM or Website is provided)

### CD-ROM / Website Check

Item	Expected	Status	Reviewer's comments	Action
CD-ROM name	Should be the abbreviation of the survey including year.			<input type="checkbox"/> None <input type="checkbox"/> Fix
Branding	An agency-specific branding is used, with at least the name of the agency and country if relevant	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> None <input type="checkbox"/> Fix
Home page	Informative message on home page	<input type="checkbox"/> Yes <input type="checkbox"/> No	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix

Autorun	For CD-ROM only: check that the autorun is available and works.	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix <input type="checkbox"/> Check
Empty pages	CD-ROM/website should not include any empty page. Check in particular photos and maps.	<input type="checkbox"/> Some <input type="checkbox"/> None		<input type="checkbox"/> None <input type="checkbox"/> Fix
Static pages	Text in static pages should be informative.	<input type="checkbox"/> Yes <input type="checkbox"/> No	<input type="checkbox"/> Typos/spelling errors detected	<input type="checkbox"/> None <input type="checkbox"/> Add <input type="checkbox"/> Fix
Links	All links in the CD-ROM/website checked.	<input type="checkbox"/> Yes <input type="checkbox"/> No		<input type="checkbox"/> None <input type="checkbox"/> Fix <input type="checkbox"/> Check

**Other comments**