International Household Survey Network

IHSN

Quick Reference Guide for Data Archivists

DRAFT - Version 2007.03
June 2007

Olivier Dupriez (World Bank)
and
Geoffrey Greenwell (PARIS21)

Content

Acknowledgments

This document (or an updated version of it) is available at www.surveynetwork.org

Comments and suggestions are always appreciated and can be sent to
odupriez@worldbank.org or toolkit@surveynetwork.org.

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, using the Microdata Management Toolkit provided by the International Household Survey Network (IHSN)[1]. It 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
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 an IHSN Toolkit reference or training manual. It is assumed that users are already familiar with the Toolkit. A *Toolkit User's Guide* is available at www.surveynetwork.org/toolkit.

Before you start: organizing your files

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

- Create a directory for the survey. We suggest you name it using the survey's abbreviated name and year, e.g. "HIES2007" for "Household Income and Expenditure Survey 2007"

- Create various sub-directories for the data files (and for the various versions of the dataset if relevant)

- Create sub-directories for the documentation and for the program files if relevant (see example).

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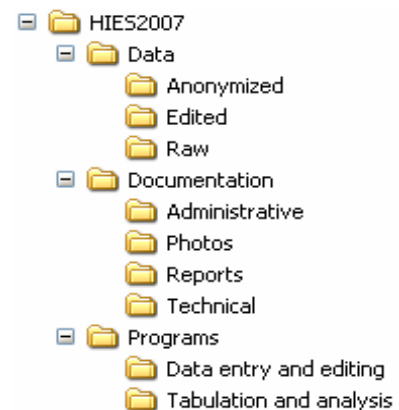
[1] DDI (Data Documentation Initiative) and DCMI (Dublin Core Metadata Initiative) are international XML metadata specifications. For more information on these standards and on the IHSN Toolkit, please visit www.surveynetwork.org.

1. Gathering and preparing the data set

Gathering and preparing data requires great care. Various versions of the data may have been produced. Before starting to document a dataset, it is important to ensure that you have the most appropriate version of all data files. If you are preparing a dataset for public release, you need the most final, edited, and anonymized version of the data. If you are documenting a dataset for archiving and internal use only, you may want to include the raw data (before editing), as well as the final, fully edited non-anonymized files. The Toolkit provides you with a possibility to document the specificity of each version of a dataset.

Much of the quality of the output generated by the Toolkit will depend upon the preparatory work. Although some changes in the data can be made in the Toolkit, it is highly recommended that the necessary checks and changes be made in advance using a statistical package. The following are recommended:

- Preferably, organize your files in a hierarchical format (not in a flat format). Hierarchical files are easier to analyze (as they contain less variables to store the same information), and are more efficient in terms of data storage (smaller files).

  Flat file:

  All data are in one "flat" file. Individual-level variables are repeated as many times as needed to accommodate the largest household. If the household contains only one member, variables created for members 2 to *n* are left empty. With such a file, one single variable (e.g. sex) will be contained in n variables (sex for member 1, sex for member 2, etc)
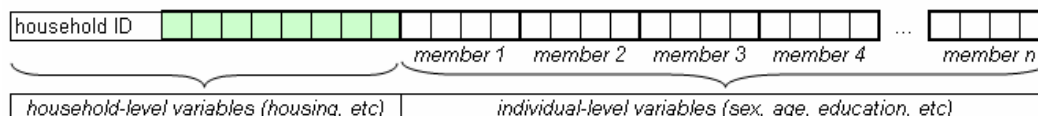
  Hierarchical file:
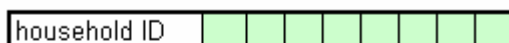
- If a dataset contains multiple related files, each record in each file must have a unique identifier. This identifier may be made of one or multiple variables. These identifiers are also called "key variables", as they are used by statistical

packages such as SPSS or Stata when data files need to be merged.[2] Identifiers must be composed of <u>numeric</u> variables only (some software applications, including the IHSN Toolkit, require that identifiers be numeric. The reason is that sorting and filtering records is much more efficient when variables are numeric).

For the convenience of users of the data, avoid identifiers made of too many variables. For example, in a household survey, the household identifier will preferably be a single variable (which you may create by concatenating a group of variables[3]), and the individual identifier should be the combination of only two variables (the household ID, and the sequential number of each member).

If you prepare your data files for public dissemination, it may be preferable to generate a unique household identification that would <u>not</u> be a compilation of geographic codes (because geographic codes are highly identifying). In such case, it is recommended to generate an ID based on a sequential number.

Example: Suppose the unique identification of a household is the combination of variables PROV (Province), DIST (District), EA (Enumeration Area), HHNUM (Household Number). The following options are possible:

Option 1:				Option 2:	Option 3:
Use a combination of four variables				Generate a concatenated ID	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

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.

- Make sure that, in all data files, the identification variable(s) provide a unique identifier. Use the `duplicate` function in SPSS or the `isid` command in Stata to verify this.

- If your dataset is hierarchical, carefully test the file relationships (using the `merge` command in SPSS and Stata). For a household survey for example, verify the all records in the individual-level files have a corresponding household in the household-level master file. Verify also that all households have at least one corresponding record in the household-roster file that lists all individuals.

- Verify that the number of records in each file corresponds to what is expected.

- Do not include string variables if they can be converted into numeric variables. Statistical software is very inefficient when working with string variables. In Stata,

---

[2] See section 3 – *Importing data and establishing relationships* for more information on key variables.

[3] 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 *hhnum*, use the command "`egen hhid=group(province district ea hh_num )`".

the `destring` and the `encode` commands can be used to make such conversions.

- Verify the completeness of your data files by comparing the content of these files with the survey questionnaire. Make sure that data from all sections of the questionnaire are included in the dataset.

- Verify that all variables are labelled (variable labels), and that the codes for all categorical variables are labelled (value labels). The labels should be short but precise; a same label should not be used for two different variables.

- For the convenience of the data users, include the relevant weighting coefficient variable(s) in all data files if possible.

- Avoid repeating a same variable in multiple files unless there is a good reason for it (variables identifying the household, variables such as geographic codes, and weighting variables should be present in all files).

- Remove all unnecessary or temporary variables from the data files (the variables that present no interest for secondary users). To be useful, a derived variable must be documented. For example, the filtering variables ("FILTER$") generated by SPSS, or the _merge variable generated in Stata are not necessary in the files once the analysis is done.

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

- Put the variables in a logic sequence (using the `order` command in Stata if needed or the "/KEEP=" option in SPSS when saving a file).

- Generate descriptive statistics for all variables (frequencies for discrete variables; min/max/mean for continuous variables) and verify that these statistics look reasonable.

- Compress the variables to reduce the file size; use the `compress` command in Stata, or the `compress` option when you save a SPSS data file.

- Sort records by their unique identifier before you save the 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 facilitate this inventory (available at www.surveynetwork.org).

2. 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 supervisors 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.

3. Importing data and establishing relationships

After all data and documentation files are gathered and checked, import the data files in the Toolkit. In the Microdata Management Toolkit, 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 Toolkit 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 Relationships* in the Toolkit. 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*.
- An individual-level file "ind.dat" with variables related to each member of 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*.
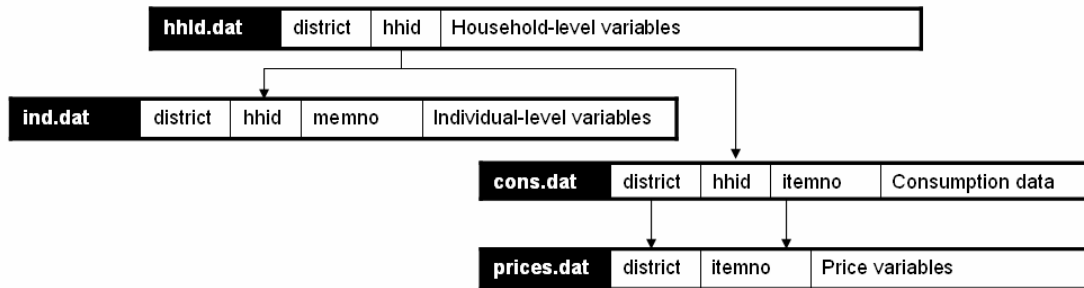
In the IHSN Toolkit, these relationships will be established as follows in the "Key variables and relationships" section of each data file:


If you have imported your data from any format other than fixed ASCII, re-sequence the data using the `Variables > Resequence` option in the Toolkit. 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	StartCol	EndCol	Record	Decimals
v1	hid	Household ID	2	*	*	*	0
v2	hi6	Area	1	*	*	*	0
v3	hi10	Result of household interview	1	*	*	*	0
v4	hl1	Line number	2	*	*	*	0

Variables:

Number	Name	Label	Width	StartCol	EndCol	Record	Decimals
v1	hid	Household ID	2	1	2	*	0
v2	hi6	Area	1	3	3	*	0
v3	hi10	Result of household interview	1	4	4	*	0
v4	hl1	Line number	2	5	6	*	0

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 Toolkit or in the source data files which will then have to be re-imported).
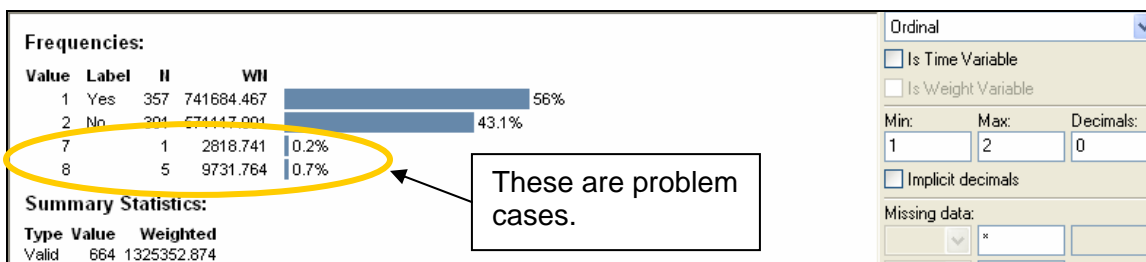
Last, view the content of each data files in the "Data Entry" page. All data should be displayed in black fonts. Data displayed in bold red indicate out of range values. Data in bold blue fonts indicate that value labels are missing for that variable.

Save the file. The Toolkit saves all data files and the information you already added (on key relationships, etc) in one single file (Nesstar format). We recommend you save this file in the survey root directory, using the survey abbreviation, year and version number as file name (e.g., HIES2007_v11). Note that it is good practice to avoid using spaces in a file name (use underscore characters instead).

4. Importing external resources

Before importing your external resources, create folders in the Toolkit as necessary (these are directories in the External Resources section in the Toolkit, 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).
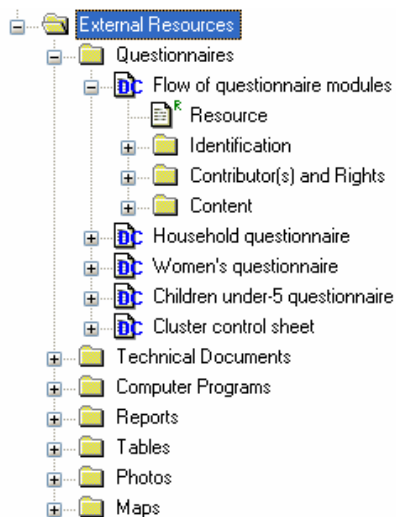
Create an entry for each resource by entering a label in the Resource Information field. This label should be short but explicit. Then identify the resource file in the "Resource" field. The field "Resource" is used to indicate the filename or URL location (website) of the external resource. The resource consists of the filename, and a relative path (relative to the directory where your study Nesstar file is saved). 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.

Example:

Let's assume your study is a household budget survey conducted in 2007. If you followed the recommendations made in the introductory chapter "Before you start – Organizing your files", you will have created a directory like C:\HIES2007. Your nesstar file will be saved as HIES2007.Nesstar in this directory. Suppose also that a document titled Report2007.pdf is saved in a directory C:\HIES2007\DOCS. When you fill the resource field in the External Resources page, do NOT enter "C:\HIES2007\DOCS\Report2007.pdf. Enter the file name as follows:

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.

5. Completing metadata

The IHSN Toolkit (version 1.n) makes use of the Data Documentation Initiative (DDI Version 2) and the Dublin Core (DCMI version X) metadata standards.

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, Data Files Description, Variables Description, and External Resources Description. 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:

- o As an archivist, you may need to seek assistance from key experts involved in some of the technical aspects of the survey.
- o As a general rule, avoid using ALL CAPS when you fill DDI fields. Also, check the spelling of all entries. The Toolkit does not provide (yet) an automatic spell checker.
- o 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.

5.1. 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 5 elements to describe the DDI document.

Study Title	The title is the official name of the survey as it is stated on the questionnaire or as it appears in the design documents. The following items should be noted: o Include the reference year(s) of the survey in the title. o Do not include the abbreviation of the survey name in the title. o As the survey title is a proper noun, the first letter of each word should be capitalized (except for prepositions or other conjunctions). o Including the country name in the title is optional. Examples: *National Household Budget Survey 2002-2003* *Popstan Multiple Indicator Cluster Survey 2002*
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:<table><tr><td>*Name*</td><td>*Role*</td></tr><tr><td>*National Statistics Office (NSO)*</td><td>*Documentation of the study*</td></tr><tr><td>*International Household Survey Network (IHSN)*</td><td>*Review of the metadata*</td></tr></table>
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. Example: *Version 1.1 (July 2006). This version is identical to version 1.0, except for the section on Data Appraisal which was updated.*
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 o *country* is the 3-letter ISO country abbreviation o *producer* is the abbreviation of the producing agency o *survey* is the survey abbreviation o *year* is the reference year (or the year the survey started) o DDI document version number Example: *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.*

5.2. 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.)

Identification	

Title	The title is the official name of the survey as it is stated on the questionnaire or as it appears in the design documents. The following items should be noted:
	   o Include the reference year(s) of the survey in the title.
	   o Do not include the abbreviation of the survey name in the title.
	   o As the survey title is a proper noun, the first letter of each word should be capitalized (except for prepositions or other conjunctions).
	   o Including the country name in the title is optional.
	
	The title will in most cases be identical to the Document Title (see above).
	
	Examples: *National Household Budget Survey 2002-2003*
	                    Popstan Multiple Indicator Cluster Survey 2002

Subtitle	Subtitle is optional and rarely used. A subtitle can be used to add information usually associated with a sequential qualifier for a survey. *Example:* *Title: Welfare Monitoring Survey 2007* *Subtitle: Fifth round*
Abbreviation	The abbreviation of a survey is usually the first letter of each word of the titled survey. The survey reference year(s) may be included. Examples: • *DHS 2000 for "Demographic and Health Survey 2005"* • *HIES 2002-2003 for "Household Income and Expenditure Survey 2003"*
Study type	The study type or *survey type* is the broad category defining the survey. This item has a controlled vocabulary (you may customize the IHSN template to adjust this controlled vocabulary if needed).
Series information	A survey may be repeated at regular intervals (such as an annual labour force survey), or be part of an international survey program (such as the MICS, CWIQ, DHS, LSMS and others). The Series information is a description of this "collection" of surveys. A brief description of the characteristics of the survey, including when it started, how many rounds were already implemented, and who is in charge would be provided here. If the survey does not belong to a series, leave this field empty. Example: *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.*
Translated title	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).

ID Number	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
	o country is the 3-letter ISO country abbreviation
	o producer is the abbreviation of the producing agency
	o survey: is the survey abbreviation
	o year is the reference year (or the year the survey started)
	o version is the number dataset version number (see Version Description below)
	Example: *The Demographic and Health Survey implemented by the Uganda Bureau of Statistics in 2005 could have the following ID: UGA-UBOS-DHS-2005-v01.*

Version

Description	The version description should contain a version number followed by a version label. The version number should follow a standard convention to be adopted by the institute. We recommend 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). Examples:
	• *v0.1: Basic raw data, obtained from data entry (before editing).*
	• *v1.2: Edited data, second version, for internal use only.*
	• *v2.1: Edited, anonymous dataset for public distribution.*
	A brief description of the version should follow the numerical identification.
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.
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.

Overview

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.
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:
	o Household (household survey, census)
	o Person (household survey, census)
	o Enterprise (enterprise survey)
	o Commodity (household survey, price survey)
	o Plots of land (agricultural survey)

Scope

Description of scope	The scope is a description of the themes covered by the survey. It can be viewed as a summary of the modules that are included in the questionnaire. The scope does not deal with geographic coverage.

	Example:
	The scope of the Multiple Indicator Cluster Survey includes:
	• *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.*
	• *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.*
	• *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.*
Topic classifications	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 at www.surveynetwork.org/toolkit).
Keywords	Keywords summarize the content or subject matter of the survey. As topic classifications, these are used to facilitate referencing and searches in electronic survey 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.
Coverage	
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	This filed aims at describing at what geographic level the data are representative. Typical entries will be "National coverage", "Urban (or rural) areas only", "state of …", "Capital city", etc. Note that we do not describe here where the data was collected. For example, as sample survey could be declared as "national coverage" even in cases where some districts where not included in the sample, as long as the sampling strategy was such that the representativity is national.
Universe	We are interested here in the survey universe (not the universe of particular sections of the questionnaires or variables), i.e. in the identification of the population of interest in the survey. The universe will rarely be the entire population of the country. Sample household surveys, for example, usually do not cover homeless, nomads, diplomats, community households. Population censuses do not cover diplomats. Try to provide the most detailed information possible on the population covered by the survey/census. Example: *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.*

Producers and Sponsors

Primary investigator	The primary investigator will in most cases be an institution, but could also be an individual in the case of small-scale academic surveys. The two fields to be completed are the Name and the Affiliation fields. Generally, in a survey, the Primary Investigator will be the institution implementing the survey. If various institutions have been equally involved as main investigators, then all should be mentioned. This only includes the agencies responsible for the implementation of the survey, not its funding or technical assistance. The order in which they are listed is discretionary. It can be alphabetic or by significance of contribution. Individual persons can also be mentioned. If persons are mentioned use the appropriate format of Surname, First name.

Other 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 competed 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:
	    o [Technical assistance in] questionnaire design
	    o [Technical assistance in] sampling methodology / selection
	    o [Technical assistance in] data collection
	    o [Technical assistance in] data processing
	    o [Technical assistance in] data analysis
	Do not include here the financial sponsors.

Funding	List the organizations (national or international) that have contributed, in cash or in kind, to the financing of the survey. The government institution that has provided funding should not be forgotten.
Other acknowledgements	This optional field can be used to acknowledge any other people and institutions that have in some form contributed to the survey.

Sampling

Sampling procedure	This field only applies to sample surveys. Information on sampling procedure is crucial (although not applicable for censuses and administrative datasets). This section should include summary information that includes though is not limited to:
	    o Sample size
	    o Selection process (e.g., probability proportional to size or over sampling)
	    o Stratification (implicit and explicit)
	    o Stages of sample selection
	    o Design omissions in the sample
	    o Level of representation
	    o Strategy for absent respondents/not found/refusals (replacement or not)
	    o Sample frame used, and listing exercise conducted to update it
	
	It is useful also to indicate here what variables in the data files identify the various levels of stratification and the primary sample unit. These are crucial to the data users who want to properly account for the sampling design in their analyses and calculations of sampling errors.
	
	This section accepts only text format; formulae cannot be entered. In most cases, technical documents will exist that describe the sampling strategy in detail. In such cases, include here a reference (title/author/date) to this document, and make sure that the document is provided in the External Resources.
	
	Example:
	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%.
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 rates	Response rate provides that percentage of households (or other sample unit) that participated in the survey based on the original sample size. Omissions may occur due to refusal to participate, impossibility to locate the respondent, or other. Sometimes, a household may be replaced by another by design. Check that the information provided here is consistent with the sample size indicated in the "Sampling procedure field" and the number of records found in the dataset (for example, if the sample design mention a sample of 5,000 households and the data on contain data on 4,500 households, the response rate should not be 100 percent). Provide if possible the response rates by stratum. If information is available on the causes of non-response (refusal/not found/other), provide this information as well. This field can also in some cases be used to describe non-responses in population censuses.
Weighting	This field only applies to sample surveys. Provide here the list of variables used as weighting coefficient. If more than one variable is a weighting variable, describe how these variables differ from each other and what the purpose of each one of them is. Example: *Sample weights were calculated for each of the data files.* *Sample weights for the household data were computed as the inverse of the probability of selection of the household, computed at the sampling domain level (urban/rural within each region). The household weights were adjusted for non-response at the domain level, and were then normalized by a constant factor so that the total weighted number of households equals the total unweighted number of households. The household weight variable is called HHWEIGHT and is used with the HH data and the HL data.* *Sample weights for the women's data used the un-normalized household weights, adjusted for non-response for the women's questionnaire, and were then normalized by a constant factor so that the total weighted number of women's cases equals the total unweighted number of women's cases.* *Sample weights for the children's data followed the same approach as the women's and used the un-normalized household weights, adjusted for non-response for the children's questionnaire, and were then normalized by a constant factor so that the total weighted number of children's cases equals the total unweighted number of children's cases.*
Data Collection	
Dates of data collection	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. 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.

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.
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.
Notes on data collection	This element is provided in order to document any specific observations, occurrences or events during data collection. Consider stating such items like: o Was a training of enumerators held? (elaborate) o Any events that could have a bearing on the data quality? o How long did an interview take on average? o Was there a process of negotiation between households, the community and the implementing agency? o Are anecdotal events recorded? o Have the field teams contributed by supplying information on issues and occurrences during data collection? o In what language was the interview conducted? o Was a pilot survey conducted? o Were there any corrective actions taken by management when problems occurred in the field? Example: *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.* *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.* *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.* *Six staff members of GenCenStat provided overall fieldwork coordination and supervision. The overall field coordinator was Mrs. Doe.*
Data Processing	
Questionnaires	This element is provided to describe the questionnaire(s) used for the data collection. The following should be mentioned: • List of questionnaires and short description of each (all questionnaires must be provided as External Resources) • In what language were the questionnaires published? • Information on the questionnaire design process (based on a previous questionnaire, based on a standard model questionnaire, review by stakeholders). If a document was compiled that contains the comments provided by the stakeholders on the draft questionnaire, or a report prepared on the questionnaire testing, a reference to these documents should be provided here and the documents should be provided as External Resources. Example *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 translators.* *The English and Mumbo-jumbo questionnaires were both piloted as part of the survey pretest.* *All questionnaires and modules are provided as external resources.*

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

Supervision	This element will provide information on the oversight of the data collection. The following should be considered: • Were the enumerators organized in teams that included a controller and a supervisor? With how many controllers/supervisors per interviewer? • What were the main roles of the controllers/supervisors? • Were there visits to the field by upper management? How often? Example: *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).* *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.
	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.
	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.
	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.

Data Processing

Data editing	The data editing should contain information on how the data was treated or controlled for in terms of consistency and coherence. This item does not concern the data entry phase but only the editing of data whether manual or automatic.
	• Was a hot deck or a cold deck technique used to edit the data?
	• Were corrections made automatically (by program), or by visual control of the questionnaire?
	• What software was used?
	If materials are available (specifications for data editing, report on data editing, programs used for data editing), they should be listed here and provided as external resources.
	Example:
	Data editing took place at a number of stages throughout the processing, including:
	a) Office editing and coding
	b) During data entry
	c) Structure checking and completeness
	d) Secondary editing
	e) Structural checking of SPSS data files
	Detailed documentation of the editing of data can be found in the "Data processing guidelines" document provided as an external resource.
Other processing	Use this field to provide as much information as possible on the data entry design. This includes such details as:
	• Mode of data entry (manual or by scanning, in the field/in regions/at headquarters)
	• Computer architecture (laptop computers in the field, desktop computers, scanners, PDA, other; indicate the number of computers used)
	• Software used
	• Use (and rate) of double data entry
	• Average productivity of data entry operators; number of data entry operators involved and their work schedule
	Information on tabulation and analysis can also be provided here.
	All available materials (data entry/tabulation/analysis programs; reports on data entry) should be listed here and provided as external resources.
	Example:
	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:
	1) Questionnaire reception
	2) Office editing and coding

	3) Data entry
	4) Structure and completeness checking
	5) Verification entry
	6) Comparison of verification data
	7) Back up of raw data
	8) Secondary editing
	9) Edited data back up

After all clusters are processed, all data is concatenated together and then the following steps are completed for all data files:

10) Export to SPSS in 4 files (hh - household, hl - household members, wm - women, ch - children under 5)
11) Recoding of variables needed for analysis
12) Adding of sample weights
13) Calculation of wealth quintiles and merging into data
14) Structural checking of SPSS files
15) Data quality tabulations
16) Production of analysis tabulations

Details of each of these steps can be found in the data processing documentation, data editing guidelines, data processing programs in CSPro and SPSS, and tabulation guidelines.

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, CSPro version 2.6.007 was used with a highly structured data entry program, using system controlled approach that controlled entry of each variable. 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 anthropometric Z-scores was also included in the data entry programs for use during analysis. Open-ended responses ("Other" answers) were not entered or coded, except in rare circumstances where the response matched an existing code in the questionnaire.

Structure and completeness checking ensured that all questionnaires for the cluster had been entered, were structurally sound, and that women's and children's questionnaires existed for each eligible woman and child.

100% verification of all variables was performed using independent verification, i.e. double entry of data, with separate comparison of data followed by modification of one or both datasets to correct keying errors by original operators who first keyed the files.

After completion of all processing in CSPro, all individual cluster files were backed up before concatenating data together using the CSPro file concatenate utility.

For tabulation and analysis SPSS versions 10.0 and 14.0 were used. Version 10.0 was originally used for all tabulation programs, except for child mortality. Later version 14.0 was used for child mortality, data quality tabulations and other analysis activities.

After transferring all files to SPSS, certain variables were recoded for use as background characteristics in the tabulation of the data, including grouping age, education, geographic areas as needed for analysis. In the process of recoding ages and dates some random imputation of dates (within calculated constraints) was performed to handle missing or "don't know" ages or dates. Additionally, a wealth (asset) index of household members was calculated using principal components analysis, based on household assets, and both the score and quintiles were included in the datasets for use in tabulations.

Data Appraisal

Estimate of sampling error	For sampling surveys, it is good practice to calculate and publish sampling error. This field is used to provide information on these calculations. This includes: • A list of ratios/indicators for which sampling errors were computed. • Details regarding the software used for computing the sampling error, and reference to the programs used (to be provided as external resources) as the program used to perform the calculations. • Reference to the reports or other document where the results can be found (to be provided as external resources). Example: *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.* *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).* *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.*

Other forms data appraisal	This section can be used to report any other action taken to assess the reliability of the data, or any observations regarding data quality. This item can include:
	• For a population census, information on the post enumeration survey (a report should be provided in external resources and mentioned here).
	• For any survey/census, a comparison with data from another source.
	• Etc.
	
	Example:
	
	A series of data quality tables and graphs are available to review the quality of the data and include the following:
	    - *Age distribution of the household population*
	    - *Age distribution of eligible women and interviewed women*
	    - *Age distribution of eligible children and children for whom the mother or caretaker was interviewed*
	    - *Age distribution of children under age 5 by 3 month groups*
	    - *Age and period ratios at boundaries of eligibility*
	    - *Percent of observations with missing information on selected variables*
	    - *Presence of mother in the household and person interviewed for the under 5 questionnaire*

	- *School attendance by single year age* - *Sex ratio at birth among children ever born, surviving and dead by age of respondent* - *Distribution of women by time since last birth* - *Scatter plot of weight by height, weight by age and height by age* - *Graph of male and female population by single years of age* - *Population pyramid* *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.* *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).*
Data Access	
Access authority	This section is composed of various sections: Name-Affiliation-email-URI. This information provides the contact person or entity to gain authority to access the data. It is advisable to use a generic email contact such as data@popstatsoffice.org whenever possible to avoid tying access to a particular individual whose functions may change over time.

Confidentiality	If the dataset is not anonymized, we may indicate here what Affidavit of Confidentiality must be signed before the data can be accessed. Another option is to include this information in the next element (Access conditions). If there is no confidentiality issue, this field can be left blank.
	
	An example of statement could be the following:
	
	Confidentiality of respondents is guaranteed by Articles N to NN of the National Statistics Act of [date].
	
	Before being granted access to the dataset, all users have to formally agree:
	
	   1. *To make no copies of any files or portions of files to which s/he is granted access except those authorized by the data depositor.*
	
	   2. *Not to use any technique in an attempt to learn the identity of any person, establishment, or sampling unit not identified on public use data files.*
	
	   3. *To hold in strictest confidence the identification of any establishment or individual that may be inadvertently revealed in any documents or discussion, or analysis. Such inadvertent identification revealed in her/his analysis will be immediately brought to the attention of the data depositor.*
	
	This statement does not replace a more comprehensive data agreement (see Access condition).

Access conditions	Each dataset should have an "Access policy" attached to it. The IHSN recommends three levels of accessibility:
	  • Public use files, accessible to all
	  • Licensed datasets, accessible under conditions
	  • Datasets only accessible in a data enclave, for the most sensitive and confidential data.
	
	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.
	
	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.

	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: *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:*
	
	    1. 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].
	    2. 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.
	    3. 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].
	    4. 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.
	    5. 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.
	    6. An electronic copy of all reports and publications based on the requested data will be sent to the [National Data Archive].
	    7. 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 interpretations or inferences based upon such uses.

Citation requirements	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. Example: *"National Statistics Office of Popstan, Multiple Indicators Cluster Survey 2000 (MICS 2000), Version 1.1 of the public use dataset (April 2001), provided by the National Data Archive. www.nda_popstan.org"*
Disclaimer and Copyright	
Disclaimer	A disclaimer limits the liability that the Statistics Office has regarding the use of the data. A standard legal statement should be used for all datasets from a same agency. The IHSN recommends the following formulation: *The user of the data acknowledges that the original collector of the data, the authorized distributor of the data, and the relevant funding agency bear no responsibility for use of the data or for interpretations or inferences based upon such uses.*
Copyright	Include here a copyright statement on the dataset, such as: © 2007, Popstan Central Statistics Agency

Contacts	
Contact persons	Users of the data may need further clarification and information. This section may include the name-affiliation-email-URI of one or multiple contact persons. Avoid putting the name of individuals. The information provided here should be valid for the long term. It is therefore preferable to identify contact persons by a title. The same applies for the email field. Ideally, a "generic" email address should be provided. It is easy to configure a mail server in such a way that all messages sent to the generic email address would be automatically forwarded to some staff members. Example: *Name: Head, Data Processing Division* *Affiliation: National Statistics Office* *Email: dataproc@cso.org* *URI: www.cso.org/databank*

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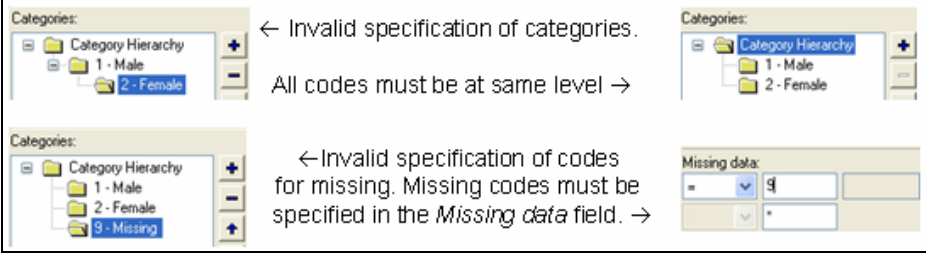
The File Description is the DDI section that aims to provide a detailed description of each data file. The IHSN has selected six of the available DDI elements.

Contents	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. Examples: • *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.* • *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.*
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	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. 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).

Missing data	Missing data can be given certain coding. A common convention is to iterate the number "9" to fill a field. This value needs to be defined as missing in the data set and can be explained in detail in this element.
Notes	This field, aiming to provide information to the user on items not covered elsewhere, will in most cases be left empty.

5.4. Good practices for completing the Variables Description

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

Variable Names	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 TOOLKIT, but make recommendations to the data processor for consideration for future surveys.
Variable Labels	All variables should have a label that • Provides the item or question number in the original data collection instrument (unless item number serves as the variable name) • Provides a clear indication of what the variable contains • Provides an indication of whether the variable is constructed from other items Recommendations: • Do not use ALL CAPS in labels. • Make sure that different variables have different labels (avoid duplicate labels). The IHSN Toolkit provides a tool to check availability and unicity of variable labels (see Tools > Validate Variable). • For expenditure or income: indicating the currency and period of reference is crucial (e.g. "Annual per capita real expenditure in local currency"
Width, StartCol, Endcol	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 > Resequence" command in the Toolkit, for each data file.
Categories	Variable categories are the lists of codes (and their meaning) that apply to the variable. The Toolkit imports categories and their labels from the source data files (SPSS, Stata). If necessary, add/edit the codes. Use the Documentation > Create categories from statistics if the source dataset did not include value labels (e,g., when imported from ASCII). Make sure the categories are not hierarchical, and do not include codes for "Missing". The codes for Missing must be specified in the "Missing data" field. If you fail to do that, the summary statistics (mean, standard deviation, etc) will be calculated including the missing code, which will be considered as a valid value.
Data type	Four types of variables are recognized by the Toolkit: • Numeric: *Numeric variables are used to store any number, integer or floating point (decimals).*

	• Fixed string: *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.* • Dynamic string: *Dynamic string variables can be used to store open-ended questions.* • Date: *date variables stored in ISO format (YYYY-MM-DD?—should specify)* 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 Toolkit. See the section on "Gathering and preparing the dataset" above.

Measure	The Microdata Management Toolkit will allow you to define the measure of a variable as:
	• *Nominal*: variable with numeric assignations for responses; the number assigned to each response does not have a meaning by itself.
	    Example: Variable *sex*: 1 = Male, 2 = Female (the number does not have a meaning by itself; we could as well have assigned Male = 2 and Female = 1). When variables are nominal, we can produce frequency tables by code, but calculating mean or standard deviation of the codes would not make sense.
	• *Ordinal*: variable with numeric assignations and in a logical sequence. The absolute size of the number, or the difference between two numbers has no meaning. But the sequence of the number matters.
	    Example: An example of an ordinal variable would be a variable indicating the level of satisfaction of the respondent, for example on a scale of 1 (very unsatisfied) to 5 (very satisfied).
	• *Scale*: continuous variables that have inherent and not categorical value. Examples of such variables include the age of the person, the amount of income or expenditure, etc.

Time variable	This is a check-box used to tag and identify variables used to define time.
Weight variable	This is a check box that is used to tag the weight variable. It is a good practice to include the weight variable with each data file that is being archived. If it is included, the check box should be ticked.
Min Max	Allows modifying the minimum value of a variable. For each variable where it makes sense, you should check that the Min and Max values are correct. Remember: if a specific value is used for "Missing", this should not be included in the Min-Max range. For example, if codes 1 and 2 are used for Male and Female, and 9 for unknown sex, then the Min will be 1 and the Max will be 2. The code 9 must be listed in the "Missing" codes (see below).
Decimals	Defines the number of decimal places of a numeric variable type.
Implicit decimals	This check box is selected only when a fixed ASCII-type file is imported and the data file includes a decimal character. As the decimal character also requires a space in the variable length assignation, it is important to check this box in order to assure proper alignment of the data.
Missing data	Missing values are those values that are blank in a data file but should have been responses and are within the path or universe of the questionnaire. Missing values should always be coded. Missing values should be differentiated from "not applicable" and zero (0) values.
Statistics Options	Various options exist for displaying and presenting summary information of the variable to the user or the person browsing the output. Summary statistics are saved in the DDI document and become part of the metadata. It is therefore important to select the appropriate ones. • For nominal variables you want to be sure that the categories are well defined and that some of the summary statistics are not displayed (such as means and standard deviations. • For ordinal values, you want to be sure that the categories are displayed if they are required. Not all ordinal values will require a category. In some cases you may want to include some summary statistics such as mean and standard deviation. • For scale values, you do not want to define categories and you may want to include

	some summary statistics such as mean and standard deviation.
	Make sure you do not include "Frequencies" for variables such as the household identification number or enumeration area. This would produce a useless frequency table, that would considerably increase the size of your DDI file (in general, a very large DDI file– 8 to 10Mb or more– indicates such a problem).
	Make sure also that you do not include meaningless summary statistics, such as the mean or standard deviation calculated on the codes used for variable SEX.
	Notes:
	• Summary statistics such as the mean or standard deviation are calculated using all valid values. If special codes are used to indicate missing values, make sure they are declared in the "Missing" section. If not, they will be included in the calculations. For example, if you use code 99999 for indicating missing values in a variable on household expenditure, code 99999 must be listed in the missing section as follows:
	• If you modify information such as the categories or missing values, you must use the "Documentation > Update Statistics" command in the Toolkit to refresh the summary statistics.
Weights	The appropriate weight should be attached to the file and selected in this element. The weight should be well labelled.
Definition	This element provides a space to describe the variable in detail. Not all variables require definition. The following variables should always be defined when available in a questionnaire: • Household (attach this definition to the "household ID" variable • Head of household (attach this definition to the variable "relationship to the head" • Urban/rural
Universe	The universe at the variable level reflects skip patterns within-records in a questionnaire. This information can typically be copy/pasted from the survey questionnaire. Try to be as specific as possible. This information is very useful for the analyst. In many cases, a block of variables will have the same universe (for example, a block of variables on education can all relate to the "Population aged 6 to 24 year). The Toolkit allows you to select multiple variables and enter the universe information to all variables at once.
Source of information	Enter information regarding who provided the information contained within the variable. In most cases, the source will be "Head of household" or "Household member". But it may also be - GPS measure (for geographic position) - Interviewer's visual observation (for type of dwelling) - Best informant in community - Etc.
Concepts	Greater description on the nature of the variable can be placed in this element. For example this element can provide a clearer definition for certain variables (i.e. a variable that provides information on whether a person is a household member). In the case of household membership, a conceptual definition can be provided. Ex: A household member is defined as any person who has been resident in the household for six months or more in a given year and takes meals together OR by default the head of household, infants under 6 months, newly wedded couples etc.

Pre-question text Literal question Post-question text	The *pre-question texts* are the instructions provided to the interviewers and <u>printed in the questionnaire before the literal question</u>. This does not apply to all variables. Do not confuse this with instructions provided in the interviewer's manual. With this and the next two fields, one should be able to understand how the question was asked during the interview. See example below. The *literal question* is the full text of the questionnaire as the enumerator is expected to ask it when conducting the interview. This does not apply to all variables (it does not apply to derived variables). The *post-question texts* are instructions provided to the interviewers, <u>printed in the questionnaire after the literal question</u>. Post-question can be used to enter information on skips provided in the questionnaire. This does not apply to all variables. Do not confuse this with instructions provided in the interviewer's manual. With this and the next two fields, one should be able to understand how the question was asked during the interview. See example above. Example: In the example below (extracted from a UNICEF-MICS standard questionnaire), we find a pre-question, a literal question and a post-question. | 8. *Check age. If child is 3 years old or more, ask*: DOES (*name*) ATTEND ANY ORGANIZED LEARNING OR EARLY CHILDHOOD EDUCATION PROGRAMME, SUCH AS A PRIVATE OR GOVERNMENT FACILITY, INCLUDING KINDERGARTEN OR COMMUNITY CHILD CARE?	Yes... 1 No.. 2 2⇨NEXT MODULE DK .... 9 9⇨NEXT MODULE | • Pre-question: *Check age. If child is 3 years old or more, ask:* • Literal question: *Does (name) attend any organized learning or early childhood education programme, such as private or government facility, including kindergarten or community child care?* • Post-question: *If answer is 2 or 9 > Goto next module*
Interviewer Instruction	Copy/paste the instructions provided to the interviewers <u>in the interviewer's manual</u>. In cases where some instructions relate to multiple variables, repeat the information in all variables. The Toolkit allows you to select multiple variables and enter the information to all these variables at once.
Imputation	The field is provided to record any imputation or replacement technique used to correct inconsistent or unreasonable data. It is recommended that this field provide a summary of what was done and include a reference to a file in the external resources section.
Recoding and derivation	This element applies to data that were obtained by recoding collected variables, or by calculating new variables that were not directly obtained from data collection. It is very important to properly document such variables. Poorly documented variables cannot (or should not) be used by researchers. In cases where the recoding or derivation method was very simple, a full description can be provided here. For example, if variable AGE_GRP was obtained by recoding variable S1Q3, we could simply mention *"Variable obtained by recoding the age in years provided in variable S1Q3 into age groups for years 0-4, 5-9, …, 60-64, 65 and over. Code 99 indicates unknown age."* When the derivation method is more complex, provide here a reference to a document (and/or computer program) to be provided as an External Resource. This will be the case for example for a variable "TOT_EXP" containing the household annual total expenditure, obtained from a household budget survey. In such case, the information provided here could be: *"This variable provides the annual household expenditure. It was obtained by aggregating expenditure data on all goods and services, available in sections 4 to 6 of the household questionnaire. It contains imputed rental values for owner-occupied dwellings. The values have been deflated by a regional price deflator available in variable REG_DEF". All values are in local currency. Outliers have been fixed. Details on the calculations are available in Appendix 2 of the Report on Data Processing, and in the Stata program "aggregates.do" available in external resources."*

Security	This field will be left empty in most cases. It can be used to identify variables that are direct identifiers of the respondents (or highly identifying indirect identifiers), and that should not be released.
Notes	This element is provided in order to record any additional or auxiliary information related to the specific variable.

5.5. 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 IHSN Toolkit 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	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: • Document Administrative: This includes materials such as the survey budget; grant agreement with sponsors; list of staff and interviewers, etc. • Document Analytical: Documents that present analytical output (academic papers, etc. This does not include the descriptive survey report (see below)). • Document Questionnaire: the actual questionnaire(s) used in the field. • 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. • Document Report: Survey reports, studies and other reports that use the data as the basis for their findings. • 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. • Document Other: Miscellaneous items • Audio: audio type files. • Map: Any cartographic information. • Photo: Photos can provide good documentary evidence of a survey. • 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) • Table: Tabulations such as confidence intervals that may not be included in a general report. • Video: video type files provided as additional visual information • 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). • Database: any databases related to the survey (e.g., a Devinfo database providing the aggregated results of the survey).

Title	Full title of the document as it is provided on the cover page.
Subtitle	Subtitle if relevant.
Author(s)	Include all authors that are listed on the report.
Date	Date of the publication of the report or resource (at least month and year). For reports, this is most likely stated on the cover page of the document. For other types of resources, put here the date the resource was produced.
Country	The country (or countries) that are covered by the associated document.
Language	The language of the document. Use the controlled vocabulary (drop-down).
Format	The file format provides information on the kind of electronic document being provided. This includes: PDF, Word, Excel etc. This is a controlled vocabulary. If the controlled vocabulary does not provide the format you need, type it (or add it in the controlled vocabulary using the Toolkit Template Editor). Providing information on the format will inform the user on the software needed to open the file.
ID Number	If there is a unique ID number which references the document (such as a Library of Congress number or a World Bank Publication number) include this as the ID Number.
Contributor(s)	Include the names of all organizations that have been involved or contributed to producing the publication. This included funding sources as well as authoring entities.
Publisher(s)	Include the official organization(s) accredited with producing the report.
Rights	The copyright statement on the publication.
Description	A brief description of the resource.
Abstract	An abstract of the content of the resource.
Table of Contents	Applies to reports only. Include the table of contents if available. It is not necessary to include the page numbers.
Subjects	Optional. Include a summary of the major subjects which have been covered by the report. This list of subjects should be taken from a standard multilingual thesaurus.

6. 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 Toolkit 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.

In the final output of the Toolkit (CD-ROM of website), 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.

7. Running validations and diagnostics

The Microdata Management Toolkit 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.

- *DDI Diagnostic*: this provides a visual display and issues warnings if DDI elements are missing. It also displays information at the file level and identifies any variables with missing labels, discrete variables with missing value or code label, variables with the same name or frequency displays with more than 30 modalities.

- *DDI Diagnostic Detailed*: this provides a more in-depth display as the simpler DDI diagnostic (above). It checks the metadata at the individual variable level and checks: labelling, definitions, universe, source etc.

- *Dublin Core diagnostic*: Checks the metadata provided for the External Resources.

In addition to these validations, it is recommended that you generate the DDI document (in the Toolkit, use the Export DDI" command) 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).

8. Generating the survey documentation in PDF

The Microdata Management Toolkit includes a useful tool for producing a PDF document summarizing all metadata entered in the Toolkit (see *Tools > Study Documentation PDF*). 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 Toolkit.

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).

9. Producing the final output

Once you are confident that all necessary checks have been completed, you may generate the final output using the CD-ROM Builder module of the IHSN Toolkit. This includes the CD-ROM and survey website.

Before you generate the CD-ROM:
- Make sure you have a customized "branding" for the CD-ROM. If you don't, design a branding (instructions are provided in the Toolkit User's Guide).
- Prepare content for the "Home page" of the CD-ROM. Typically, a statement by the Director of the data producing agency, or a brief summary of the objectives and findings of the survey, will be generated.

Generate the CD-ROM with the appropriate options. The IHSN recommends:
- To generate the CD-ROM without data
- To export all datasets to ASCII format, and to include the zipped ASCII files on the CD-ROM, together with the syntax file to export the data to SPSS, Stata and other formats (which shoulod be provided as external resources). The ASCII format is more standard than the Toolkit Nesstar format, and is a guarantee of long-term viability of your data files.
- To include an autorun file.
- To include all external resources.
- To check that the CD-ROM outline does not include any empty pages (use the Toolkit utility to check)

- To name the CD-ROM according to the dataset abbreviation and version.

After you generate the CD-ROM:
- Check all links before you replicate it, in particular the ones to external resources.
- Test the autorun.

If your agency has a website, you may upload the content of the CD-ROM 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.

10. 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 IHSN can assist in identifying external reviewers.

The external review can be based on:
1. The DDI file (xml file, containing no microdata and no external resources)
2. The Nesstar file (containing microdata and DDI/DCMI metadata)
3. The CD-ROM (or website), without microdata
4. The CD-ROM (or website), with microdata

The preferred option is the last one, as it allows a full check of the final output. If data are highly confidential and cannot be shared with the reviewer, option 3 is 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 or the Nesstar file to the reviewer.

International Household Survey Network
DDI Reviewers' Feedback Form

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

Version 1.0 (May 2007) - Form available at www.surveynetwork.org

DOCUMENT DESCRIPTION

Document Description				
DDI Element	**Expected**	**Status**	**Reviewer's comments**	**Action**
Study Title	Proper noun format, years separated by hyphen	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Metadata producer	Name of the person and affiliation.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Date of production	Date in ISO format.	☐ Provided ☐ Not provided		☐ None ☐ Add ☐ Fix ☐ Check
DDI Document Version	Version number based on a standard naming convention.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
DDI Document ID Number	Number coherent with the Study Description ID Number.	☐ Provided ☐ Not provided		☐ None ☐ Add ☐ Fix ☐ 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: *Household Budget Survey 2006-2007*	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Sub-title	In most cases, there will be no subtitle. It there is one, it should provide additional information related to the title.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Abbreviation	All capitalized; includes the reference year. Example: *DHS 2004*	☐ Provided ☐ Not provided ☐ N.A.		☐ None ☐ Add ☐ Fix ☐ Check
Study type	Preferably taken from a controlled vocabulary.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Translated title	Optional (will be empty in most cases). Make sure special characters are readable.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
ID Number	Clear and consistent (based on a standard convention). Should include country abbreviation and year. Example: *UGA-UBOS-DHS-2004*	☐ Provided ☐ Not provided		☐ None ☐ Add ☐ Fix ☐ Check

STUDY DESCRIPTION – Version

DDI Element	Expected	Status	Reviewer's comments	Action
Description	Version number based on naming convention; should include a "label". Examples: *V0 – Raw data, unedited* *V1.1 – Edited non anonymized data* *V2.2 – Public use dataset, 2nd release (Nov. 2007)*	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Production date	Date in ISO format (at least month and year)	☐ Provided ☐ Not provided		☐ None ☐ Add ☐ Fix ☐ Check
Notes	More information on what distinguishes this version from any other.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Kind of data	Taken from the IHSN controlled vocabulary.	☐ Provided ☐ Not provided		☐ None ☐ Add ☐ Fix ☐ Check
Unit of analysis	Usually stated as person, household, enterprise etc. (could be several units)	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check

STUDY DESCRIPTION - Scope

DDI Element	Expected	Status	Reviewer's comments	Action
Description of scope	Typically, list of questionnaire modules.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Keywords	In the future: list of keywords based on an international multilingual thesaurus	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Topics classification	In the future: list of topics based on an international multilingual thesaurus	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check

STUDY DESCRIPTION - Coverage

DDI Element	Expected	Status	Reviewer's comments	Action
Country	Country name in full	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix
Geographic coverage	Clear and concise statement of geographic coverage. Examples: *- National, except province of …* *- Rural only*	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Other producers	Agencies that were not in charge, but participated in the implementation of the study as co-producer.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Funding	List of donors (national and international; having provided cash or in-kind contributions); national government should not be forgotten.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Other acknowledgment	Optional: acknowledgments of technical experts or others (persons or agencies) who contributed to the success of the operation.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ 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.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check

DDI Element	Expected	Status	Reviewer's comments	Action
Response rates	Response rates by stratum and stated as a percentage of the design. For a census, this will be N.A.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Weighting	Information on the weight variables available in the data files. If self-weighted, this must be explicitly stated here.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Provided ☐ Not provided ☐ N.A.		☐ None ☐ Add ☐ Fix ☐ Check
Time periods	In most cases, this will be empty.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Mode of data collection	Taken from IHSN controlled vocabulary.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Notes on data collection	Information on number and profile of interviewers and supervisors; on their training; observations on particular issues during data collection.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Questionnaires	List of questionnaire(s) and their content.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Data collectors	Optional. We do not expect a list of interviewers here.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix

Supervision	Clear description of field supervisory structure including: team size, control mechanisms etc.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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).	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Other forms of data appraisal	Statement is clear and comprehensive.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Confidentiality	Standard statement that serves for all surveys. Can use a customized version of the IHSN recommended statement.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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).	☐ Provided ☐ Not provided	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Copyright	Standard format (Year, copyright statement).	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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).	☐ Provided ☐ Not provided ☐ N.A.	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Yes ☐ No		☐ None ☐ 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)	☐ Yes ☐ No ☐ N.A.		☐ None ☐ Add ☐ Fix
Completeness	Data are available for all sections of all questionnaires; derived files are available as well.	☐ Yes ☐ No		☐ None ☐ Fix ☐ Check
Re-sequence data	Position and length of all variables is available (required for ASCII export).	☐ Yes ☐ No		☐ None ☐ 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.	☐ Yes ☐ No		☐ None

DDI Element	Expected	Status	Reviewer's comments	Action
Content	Short but clear description of the file content. Ideally, should establish the link with questionnaire sections. Example: *Section 3Aof the Household questionnaire: Education.*	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Producer	In most cases, the producer of the file is the producer of the survey.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Version	In most cases, there will be no versioning of individual file (as we have a version of the dataset).	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Processing Checks	Optional. In most cases, information on data editing will be contained at the Study Level.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Missing data	Optional. The variable description provides fields to describe missing values used for each variable.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Notes	Optional.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ All ☐ Some ☐ None		☐ None

Variable Labels	All variables should have a unique, clear label. Use the "Validate Variable" tool in the Toolkit to check.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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).	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check
Weights	The application of weights is optional, but good to have when applicable. The appropriate weight must be applied to each variable.	☐ All ☐ Some ☐ None ☐ N.A.		☐ None ☐ Add ☐ Fix ☐ Check
Data type	Most variables should be numeric. String variables should be recoded, encoded, or de-string when possible.			☐ None ☐ Fix ☐ Check
Measure	Importing data from some formats in the Toolkit will not automatically impute the most appropriate measure.	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check
Is Time variable	Rarely used.	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ 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.	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ 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.	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check

DDI Element	Expected	Status	Reviewer's comments	Action
Decimals		☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check
Implicit decimals	Needed when the files were imported from ASCII.	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ 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.	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ 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 hhid variable.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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*)	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Source of information	Empty in most cases.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Concepts	Empty in most cases, except ones defined in the enumerator's or controller's manuals.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Literal question	All questions in the questionnaires attached to the corresponding variables.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Interviewer Instructions	All instructions available in the interviewer's manual should be provided here.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check

VARIABLES - Imputation and derivation

DDI Element	Expected	Status	Reviewer's comments	Action
Imputation	Empty in most cases.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Recoding and derivation	Information should be provided for all calculated variables.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check

VARIABLES - Others

DDI Element	Expected	Status	Reviewer's comments	Action
Security	Empty in most cases. If available, should indicate a level of confidentiality.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Notes	Optional. Will be empty in most cases.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.	☐ Provided ☐ Not provided		☐ None ☐ Add ☐ Fix ☐ 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)	☐ Yes ☐ No		☐ None ☐ Add ☐ Fix ☐ Check
Resource (links)	All links to external resources should be valid (no broken links). Links should be relative addresses (no absolute paths).	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check

DDI Element	Expected	Status	Reviewer's comments	Action
PDF documentation	Technical document generated using the IHSN Toolkit PDF Generator is provided and documented in the external resources	☐ Yes ☐ No		☐ None ☐ Add ☐ Fix ☐ Check
Programs	Data entry, editing, tabulation and analysis programs should be preserved and provided.	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check
Report	All survey reports and analytical output must be provided in PDF (and in original format if available).	☐ Yes ☐ No		☐ None ☐ Add ☐ Fix ☐ Check
Label	All external resources should have a short but explicit label.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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".	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ 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.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Subtitle	Optional. For documents only. Should correspond to what is displayed on the cover page.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Author(s)	All resources should have an author (person or agency).	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check

Date	At least month and year.	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check
Country	Country to which the resource is related.	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check
Language	For documents only.	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix
Format	Based on the IHSN controlled vocabulary.	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check
ID Number	Optional	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check

EXTERNAL RESOURCES - Contributors and rights

DDI Element	Expected	Status	Reviewer's comments	Action
Contributor(s)	Optional	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check
Publisher(s)	Optional	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check
Rights	Optional but recommended.	☐ All ☐ Some ☐ None		☐ None ☐ Add ☐ Fix ☐ Check

EXTERNAL RESOURCES - Content

DDI Element	Expected	Status	Reviewer's comments	Action
Description	Short description of the resource. Very important for computer programs (must describe the purpose, software needed to run it).	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Abstract	For documents only; optional.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Table of Contents	Optional. No need to include page numbers.	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ Check
Subjects	Optional (should be based on an international multilingual thesaurus)	☐ All ☐ Some ☐ None	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix ☐ 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.			☐ None ☐ Fix
Branding	An agency-specific branding is used, with at least the name of the agency and country if relevant	☐ Yes ☐ No		☐ None ☐ Fix
Home page	Informative message on home page	☐ Yes ☐ No	☐ Typos/spelling errors detected	☐ None ☐ Add ☐ Fix

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

Other comments