I n t e r n a t i o n a l C o m p a r i s o n P r o g r a m

22

Microdata and Metadata Management Toolkit
ICP-Specific Data Documentation Initiative (DDI)

Paper for Session 4

customized for ICP purposes by
Nada Hamadeh & Olga Akcadag

Regional Coordinators Meeting

September 28-30, 2009

Washington DC

	Label	Description	Vocabulary
Document Description			
Metadata Preparation	Survey Title	The title is the official name of the survey as it is stated on the questionnaire or as it appears in the design documents. The following items should be noted: - Include the reference period and year(s) of the survey in the title. - Do not include the abbreviation of the survey name in the title. - As the survey title is a proper noun, the first letter of each word should be capitalized (except for prepositions or other conjunctions). - Including the country name in the title is optional. Examples: - ICP Household Consumption Survey 2011 - ICP Equipment Survey 2011	
	Metadata Producer	Name of the national organization(s) who documented the metadata on the dataset. The metadata producing agency could be different than the data producing agency. Use the "role" attribute to distinguish different stages of involvement in the production process. Examples: Name: National Statistics Office (NSO) Role: Documentation of the study Name: Asian Development Bank (ADB) Role: Documentation of the study	
	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	

	Label	Description	Vocabulary
		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 2011). 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 - country is the 3-letter ISO country abbreviation - producer is the abbreviation of the metadata producing agency - survey is the survey abbreviation - year is the reference year (or the year the survey started) - DDI document version number Example: DDI-UGA-UBOS-HHC-2011-V01 DDI-UGA-AFDB-EQP-2011-V01	

	Label	Description	Vocabulary
Study Description			
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: - Include the reference period and year(s) of the survey in the title. - Do not include the abbreviation of the survey name in the title. - As the survey title is a proper noun, the first letter of each word should be capitalized (except for prepositions or other conjunctions). - Including the country name in the title is optional. The title will in most cases be identical to the Document Title (see above). Examples: - ICP Household Consumption Survey 2011 - ICP Equipment Survey 2011	
	ICP Participation Information	A country may have participated in previous ICP rounds. The ICP Participation Information is a history of the country participation in the ICP. Example: Country X participated in the following rounds of ICP: 1985, 1993-1996, and 2005.	
	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).	

	Label	Description	Vocabulary
	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 - country is the 3-letter ISO country abbreviation - producer is the abbreviation of the data producing agency - survey is the survey abbreviation - year is the reference year (or the year the survey started) - DDI document version number Example: DDI-UGA-UBOS-HHC-2011-V01 DDI-UGA-UBOS-EQP-2011-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 transmission to regional or global coordinators. 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.	

	Label	Description	Vocabulary
	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 10 items but is not limited to them.	HHC data
			Housing data
			Health data
			Education data
			Compensation of employees data
			Construction data
			Equipment data
			National accounts data
			Exchange rates
			Population
	Unit of Analysis	Basic unit(s) of analysis or observation. It is associated with a drop down box providing controlled vocabulary. That controlled vocabulary includes 10 items but is not limited to them.	HHC data: a good or a service
			Housing data: a Housing unit
			Health data: a health service provided
			Education data: an education service
			Compensation of employees data: a Government occupation

	Label	Description	Vocabulary
			Construction data: a Construction input item, a project, a component or a basket of components or inputs
			Equipment data: an equipment item
			National accounts data: the National economy
			Exchange rates: a currency
			Population: a nation/country
Scope	Questionnaire Scope	The scope is a description of the themes covered by the survey questionnaire. It can be viewed as a summary of the themes that are included in the questionnaire. The scope does not deal with geographic coverage. Example: The scope of the ICP Household Consumption Survey Questionnaire includes: - Product Information: Code, Name, etc. - Outlet Information: Name, Type, etc. - Price Information: Observed, Converted, etc.	
	ICP Classifications	It's about: (a) proper classifications such as the ICP Classification of GDP Expenditures or the COICOP, etc.; and (b) lists of items used in price surveys.	
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 field aims at describing at what geographic level the data are representative. Typical entries will be "National coverage", "Urban (or rural) areas only", "State of ...", "Capital city", etc. Note that we do not describe here	

	Label	Description	Vocabulary
		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.	
Producers and Sponsors	Producing agency	The producing agency will in most cases be the institution implementing the survey. The two fields to be completed are the Name and the Affiliation fields. 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.	
	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: - [Technical assistance in] questionnaire design - [Technical assistance in] sampling methodology / selection - [Technical assistance in] data collection	

	Label	Description	Vocabulary
		- [Technical assistance in] data processing - [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 Acknowledgments	This optional field can be used to acknowledge any other people and institutions that have in some form contributed to the survey.	
Sampling	Survey frame	Information on survey frame is crucial to ICP. This section should include summary information that includes though is not limited to: - Sample size - Selection process - Stratification - Sample frame used, and listing exercise conducted to update it Example: The CPI sample frame was the starting point for designing the ICP sample. 500 outlets were selected for the sample. Of these, 200 were supermarkets, 150 were open markets, and 150 were neighborhood shops. 300 were from urban areas, and 200 from rural areas.	

	Label	Description	Vocabulary
	Deviations from Survey Frame	Sometimes the reality of the field requires a deviation from the survey frame (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.	
	Weighting	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.	
Data Collection	Dates of Collection	Enter the dates (at least month and year) of the start and end of the data collection. DATE MUST BE ENTERED IN THE ISO FORMAT YYYY-MM-DD 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 Periods	This field will usually be left empty. Time period differs from the dates of collection as they represent the period for which the data collected are applicable or relevant. NOTE: DATE MUST BE ENTERED IN THE FORMAT YYYY-MM-DD	
	Questionnaires	This element is provided to describe the questionnaire(s) or forms used for the data collection. The following should be mentioned: - List of questionnaires and short description of each - In what language were the questionnaires published? - Information on the questionnaire design process	

	Label	Description	Vocabulary
		(based on a previous questionnaire, based on a standard model questionnaire, review by stakeholders).	
	Data Collectors	This element is provided in order to record information regarding the data collectors. This includes the number of data collectors, their affiliations, their training, etc.	
	Supervision	This element will provide information on the oversight of the data collection. The following should be considered: - Were the collectors organized in teams that included a controller and a supervisor? With how many controllers/supervisors? - What were the main roles of the controllers/supervisors?	
Data Processing	Data Editing	The data editing field should contain information on how the data was treated or controlled for in terms of consistency and coherence. Information should include editing technique used, software used, type of corrections made, etc. If materials are available (specifications for data editing, report on data editing, programs used for data editing), they should be listed here and provided as external resources.	
	Other Processing	Use this field to provide as much information as possible on the data entry design. This includes such details as: - Mode of data entry (manual or by scanning, in the field/in regions/at headquarters) - Computer architecture (laptop computers in the field, desktop computers, scanners, PDA, other;	

	Label	Description	Vocabulary
		indicate the number of computers used) - Software used All available materials (data entry/tabulation/analysis programs; reports on data entry) should be listed here and provided as external resources.	
Data Appraisal	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.	
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 <mailto:data@popstatsoffice.org> whenever possible to avoid tying access to a particular individual whose functions may change over time.	
	Confidentiality	The nature of any confidentiality or non-disclosure agreement signed before data access can be granted.	
	Access Conditions	Description of the terms under which users are allowed to access survey data. Each ICP survey dataset should have an "Access policy" attached to it. The ICP recommends three levels of accessibility: - Public use files, accessible to all - Confidential data, accessible under conditions - Strictly confidential data.	

	Label	Description	Vocabulary
	Citation Requirement	Citation requirement is the way that the dataset should be referenced when cited in any publication. Every dataset should have a citation requirement. This will guarantee that the data producer gets proper credit, and that analytical results can be linked to the proper version of the dataset. Example: "National Statistics Office of X, ICP Survey 2011, Version 1.1 of the public use dataset (April 2012), provided by the National Data Archive. www.nda_X.org"	
Disclaimer and Copyright	Disclaimer	A disclaimer limits the liability that the Statistics Office has regarding the use of the data. The ICP recommends the following formulation: The user of the data acknowledges that the original collector of the data, the authorized distributor of the data, and the relevant funding agency bear no responsibility for use of the data or for interpretations or inferences based upon such uses.	
	Copyright	Include here a copyright statement on the dataset, such as: (c) 2012, National Statistics Office of X	
Contacts	Contact Persons	Users of the data may need further clarification and information. This section may include the name-affiliation-email 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. Example: Name: Head, Data Processing Division	

	Label	Description	Vocabulary
		Affiliation: National Statistics Office Email: dataproc@cso.org	
File Description			
Data Files	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.	
	Producer	Put the name of the agency that produced the data file. Most data files will have been produced by the survey producing agency. 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. It is more important to fill the field identifying the version of the dataset.	

	Label	Description	Vocabulary
	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. data editing, 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.	
Variable Description			
Description	Definition	This element provides a space to describe the variable in detail.	
	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.	
Imputation and Derivation	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.	