



DANE
Para tomar decisiones



Statistical Regulation, Planning, Standardization
and Normalization Division
(DIRPEN)

GUIDE FOR THE DOCUMENTING OF METADATA USING DDI AND DUBLIN CORE STANDARDS

February 2014



NATIONAL ADMINISTRATIVE DEPARTMENT OF STATISTICS

MAURICIO PERFETTI DEL CORRAL
Director

DIEGO SILVA ARDILA
Deputy Director

MARÍA LEONOR VILLAMIZAR GÓMEZ
General Secretary

TECHNICAL DIRECTORS

EDUARDO EFRAÍN FREIRE DELGADO
Methodology and Statistical Production

LILIANA ACEVEDO ARENAS
Censuses and Demography

RAMÓN RICARDO VALENZUELA (e)
Statistical Regulation, Planning, Standardization and Normalization

MIGUEL ÁNGEL CÁRDENAS CONTRERAS
Geostatistics

JUAN FRANCISCO MARTÍNEZ (e)
Synthesis and National Accounts

ÉRIKA MOSQUERA ORTEGA
Dissemination, Marketing and Statistical Culture

Bogotá, D. C., 2014

© DANE, 2015

No reproduction, partial or full, may be undertaken without prior authorization from the National Administrative Department of Statistics, Colombia.

**Statistical Regulation, Planning, Standardization and Normalization
Division (Dirpen)**

Technical Director

Nelcy Araque García

Statistical Regulation Coordinator

Fredy Jahirs Rodríguez

Leader Conceptualization and Design

Grace Andrea Torres Pineda

Technical Team

Paola Fernanda Medina Tovar

Marly Johana Téllez López

Diana Cristina Prieto Peña

Rafael Humberto Zorro

Proofreader Spanish version

Sonia Marcela Naranjo Morales

Translation: Juliana Mosquera Dueñas

Proofreader English version: Ximena Díaz Gómez

CONTENTS

PRESENTATION	8
INTRODUCTION	9
1. OBJECTIVES	10
2. SCOPE	11
3. BASIC CONCEPTS	12
4. IDENTIFICATION OF ACTORS	13
5. DUBLIN CORE AND DDI STANDARDS	14
6. DOCUMENTATION PROCESS	15
6.1 Description of the document.....	17
6.2 Description of the statistical operation	20
6.3 Databases	39
6.4 External reference materials.....	46
BIBLIOGRAPHY	49

LIST OF FIGURES

Figure 1. Document description..... 17

Figure 2. Persons responsible for documentation 18

Figure 3. Single identifier 19

Figure 4. Description of the statistical operation..... 20

Figure 5. Identification..... 21

Figure 6. Title.....21

Figure 7. Subtitle 21

Figure 8. Abbreviation 22

Figure 9. Study type..... 22

Figure 10. Statistical Operation Identifier 22

Figure 11. General description 23

Figure 12. Types of data..... 25

Figure 13. Thematic coverage 26

Figure 14. Keywords..... 27

Figure 15. Classification of Topics 28

Figure 16. Geographic Coverage 29

Figure 17. Country 29

Figure 18. Producers and Sponsors..... 30

Figure 19. Sampling.....	32
Figure 20. Data Collection.....	33
Figure 21. Method of Data Collection	34
Figure 22. Notes on data collection	35
Figure 23. Database access.....	37
Figure 24. Contact persons	38
Figure 25. File description	39
Figure 26. Variables	40
Figure 27. Documentation.....	42
Figure 28. Questions	43
Figure 29. Imputation and derivation.....	44
Figure 30. Others.....	45
Figure 31. External reference materials	46
Figure 32. Example of presentation of external reference material.....	47
Figure 33. Example of presentation of documents' names	47
Figure 34. Type of reference material.....	47

LIST OF TABLES

Table 1. Contributors 31

Table 2. Funding..... 32

Table 3. Collection dates..... 34

Table 4. Reference period 35

Table 5. Data Collection 36

Table 6. Access authority 37

PRESENTATION

The National Administrative Department of Statistics (DANE), as the coordinator entity of the National Statistical System (NSS) and within the framework of the Statistical Harmonization and Planning project, works towards the strengthening and consolidation of the NSS through the following processes: the strategic production of statistics; the generation, adaptation, adoption and dissemination of standards; the consolidation and harmonization of statistical information and the coordination of instruments, stakeholders, initiatives and products. These actions are carried out in order to improve the quality of strategic statistic information, and its availability and accessibility to respond to users demand.

In this context DANE, aware of the need and obligation to provide better products for its users, developed a standard presentation guide for methodologies which contributes to the visualization and understanding of the statistical processes. With this instrument the entity was able to develop the methodological documents of its statistical researches and operations which are made available to specialized users and the general public. The documents are presented in a standard and comprehensive manner, thus facilitating the understanding of the main technical characteristics involved in the processes and sub-processes of each research, making them available for both specialized users and the general public.

These series of guides promote the transparency, trust and credibility of the technical expertise of DANE, for a better understanding and use of statistical information. This information is produced according to the principles of coherence, comparability, integrality and quality of the statistics.

INTRODUCTION

DANE, aware of the need to provide greater clarity, transparency and technical trust to users regarding the process of statistical operations developed in the entity and within those institutions that are part of the NSS, adopted in 2009 an initiative based on an international statistical standard called Accelerated Data Program (ADP). This program aims to document, disseminate and preserve metadata and microdata in accordance with international standards and practices through the Nesstar Publisher tool. Similarly, this program enables the development of parameters for the presentation of the technical and methodological documents used by each statistical operation.

The need for information systems that capture and compile the information, presenting it in a friendly way to the user, and guide it according to the principles of accessibility, transparency, reliability and timeliness has increased over time. Thus the need arose of documenting with structured and standardized metadata and microdata, being of great importance the organization and maintenance of the institutional memory for the preservation of information and the data accompanying it.

This guide provides the necessary guidelines for documenting metadata and microdata through DDI and Dublin Core standards to private, academic and public entities, in order to facilitate its implementation and have a standardized resource with a common vocabulary.

1. OBJECTIVES

General objective

To provide guidelines for the documentation of statistical operations produced by the entities that are part of the NSS by means of the DDI and Dublin Core standards. This is done, taking into account the context, quality, conditions and characteristics of data, microdata and related material for the purpose of using an international statistical standard and thus obtaining harmonized and standardized metadata.

Specific objectives

- To promote the use of international statistical standards for the documentation of statistical operations in the NSS.
- To define the steps to be taken when preparing the documentation using the DDI and Dublin Core Standards.
- To establish the information required for documenting a statistical operation.

2. SCOPE

Metadata documentation relies on the standardization of methodological documents where technical content is specified in accordance with statistical processes. This seeks to facilitate compliance with the fundamental principles for the use and integration of statistics and thus to contribute to national, regional and international comparison. This process applies to all entities producing statistics that are part of the NSS and it should be developed by the entities' officials.

This document provides the necessary guidelines for documenting metadata in the context of their creation, utilization and conservation, considering that the objectives of the DDI and Dublin Core Standards are:

- To provide a better understanding of statistical operations produced in the NSS.
- To contribute to the institutional memory.
- To provide policy makers and society in general with the characteristics of statistical operations that are produced in the country.
- To contribute to the coordination between the various producers and users.

3. BASIC CONCEPTS

Statistical Standard: “A statistical standard provides a comprehensive set of guidelines for surveys and administrative sources collecting information on a particular topic. (...) The use of statistical standards permits the repeated collection of statistics on a consistent basis. They also enable the integration of data over time and across different data sources, allowing the use of data beyond the immediate purpose for which it was produced. Standards also reduce the resource requirements associated with many aspects of survey development and maintenance” (OECD).

International Statistical Standard: “The comprehensive body of international statistical guidelines and recommendations that have been developed by international organizations working with national agencies. The standards cover almost every field of statistical endeavour from data collection, processing and dissemination. Such standards also include international statistical classifications” (OECD).

Metadata: it refers to information necessary for the use and interpretation of statistics. Metadata describe the conceptualization, quality, generation, calculation and characteristics of a set of statistical data (DANE, 2012e).

Microdata: data pertaining to the characteristics of population units under study (individuals, households and establishments, among others.) which constitute a unit of information in a database, and that are collected by means of a statistical operation (DANE, 2012e).

Statistical operation: set of processes and activities, which starting from the systematic collection of data lead to the production of aggregate results (DANE, 2012e).

4. IDENTIFICATION OF ACTORS

The thematic team of the statistical operation or operations, which are subject to documentation of the different NSS entities, is involved in the process of metadata documentation. The thematic team is formed by the persons who have all the knowledge about the statistical activity. Once the documentation is completed, a representative of the entity must validate and approve all the information included in the Nesstar Publisher.

5. DUBLIN CORE AND DDI STANDARDS

The Accelerated Data Program - ADP has as a basic tool named Nesstar Publisher, which is an editor for the documentation and preparation of metadata and data for publishing in the online catalog named National Data Archive (ANDA for its acronym in Spanish) which was developed by the Data Group for the International Household Survey Network (IHSN).

Nesstar is designed to promote the adoption of international standards for documentation, dissemination and preservation of metadata and microdata. This tool allows having a documentation process containing two international standards with the aim of having organized information. These are the DDI and Dublin Core Metadata Initiative.

- The Data Documentation Initiative (DDI): is an international effort to establish an XML-based standard for microdata documentation. Its objective is to provide a simple tool for recording and reporting all the important characteristics of microdata. These aspects are used in the study template in Nesstar Publisher.
- The Dublin Core Metadata Initiative (DCMI): is a set of elements for the description of digital resources. This initiative is particularly useful for describing resources related to microdata, such as: questionnaires, reports, manuals, scripts and data processing programs, etc. These items are identified in the description template of the reference material in Nesstar Publisher.

6. DOCUMENTATION PROCESS

This starts with the classification of the information that will be documented and then entered in the Nesstar Publisher where the DDI and DCMI standards are configured in order to begin the documentation process. It is important to have the following information with respect to the statistical operation subject to documentation: its methodology, databases and reference material used.

Methodology

The basic requirements that a methodology must contain can be found in the “Guía para la elaboración de documentos metodológicos estándar de las operaciones estadísticas”¹ (*Guide for the preparation of standard methodological documents of statistical operations*).

Database

The following information is required:

Databases (with or without microdata): they contain the variables and records of a statistical operation. Databases must be anonymized in order to upload them to an international statistical standard and disseminate them for user consultation.

Data Dictionary: it contains information on the database such as: the name of the file or table from where the variable comes from, field name, field description, data type and extent and field length.

Form Completion Manual: it contains a description of each of the questions of the form and how to answer each one of them, which enables the documentation of the variables that will be part of the database of the statistical operation.

Validation and Consistency Manual: it contains information on the database such as: the name of the table or file where the variable comes from, the field name, field description, data type, field length, and the database validation and consistency rules.

¹ Available at:

http://www.dane.gov.co/files/planificacion/fortalecimiento/cuadernillo/Elaboracion_documentos_metodologicos.pdf

Reference material

It refers to those documents that constitute the thematic support of the operation's statistical process. These include:

- Manuals
- Design documents
- Specifications of indicators
- Guides
- Instructions
- Presentations

Documentation is done in the Nesstar Publisher following document classification, considering it covers four sections (description of the document, description of the statistical operation, database and reference material). Each section is identified below.

6.1 DESCRIPTION OF THE DOCUMENT

The statistical operation is not always documented and disseminated by the same agency that produced the data. It is important to provide in the metadatum information on individuals and/or entities involved in the operation and during the process of its documentation. The following fields are included in this section:

Figure 1. Document description

The screenshot shows a software interface for document description. On the left, a tree view under 'My Projects' shows a 'New Study 1' folder containing 'Document description', 'Responsible for documentation', 'Date of documentation', 'Version', and 'Single identifier'. Below these are 'General description of the statistical ope', 'Datasets', 'Variable Groups', and 'External Resources'. The main area displays a form with the following sections:

- Responsible for documentation**: A table with four columns: Name, Abbreviation, Affiliation, and Role. There are two empty rows below the header.
- Date of documentation**: Three input fields for Year, Month, and Day.
- Version**: A single input field.
- Single identifier**: A single input field.

Source: DANE. Screenshot taken from Nesstar Publisher²

Persons responsible for documentation: it contains the name, affiliation and role of individuals and organizations involved in the creation of documentation of the statistical operation (they are not necessarily the producers of the information).

² The screenshot was translated for convenience of the reader.

Figure 2. Persons responsible for documentation

Responsible for documentation			
Name	Abbreviation	Affiliation	Role
Diana Cristina Prieto Peña	dcprieto@dane.gov.co	Dirección de Regulación, Planeación, Estandariz.	Documentador PAD
Date of documentation			
Year	Month	Day	
<input type="text"/>	<input type="text"/>	<input type="text"/>	
Version			
<input type="text"/>			
Single identifier			
<input type="text"/>			

Source: DANE. Screenshot taken from Nesstar Publisher³

Date of documentation: it corresponds to the date in which the documentation of the statistical operation was done. This box must be updated every time the document is revised and modified.

Version: this option allows tracking the version of the document, which provides users the ability to determine if they have the latest version. It also includes a list of changes made in each revision. This information can help users determine whether errors in earlier versions were the source of errors in the analysis.

E.g.: Version 1 (January 2011)

Version 2 (February 2011). This version replaces version 1 (January 2011) since it includes a more detailed description of the sample.

Single identifier: it corresponds to the single identification code of the document in a file. It defines and uses the scheme country-producer-statistical operation-year in which:

- Country: it corresponds to the 3-letter ISO abbreviation.
- Producer: it corresponds to the abbreviation of the producing agency.
- Statistical operation: it corresponds to the abbreviation of the operation.
- Year: it corresponds to the year in which the metadata document was created.

³ Idem.

Figure 3. Single identifier

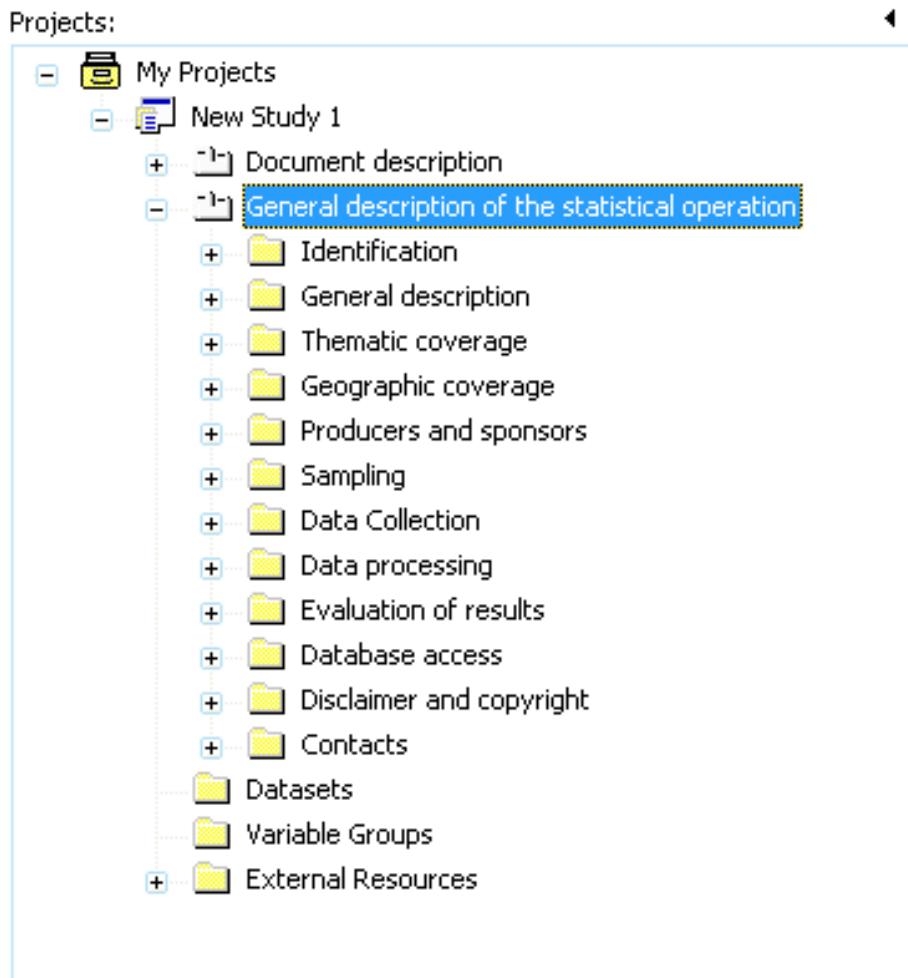
Single identifier
COL-DANE-GEIH-2012
Description of field: Definir un identificador numérico o alfanumérico único para el metadato. El campo del identificador sólo debe contener letras, números, guiones al medio (-) y puntos (.). Usar el siguiente esquema: País-Productor-Operación estadística-Año, donde: - País: abreviación de 3 letras tipo ISO. - Productor: abreviación de la agencia productora. - Operación estadística: abreviación que identifica la operación estadística. - Año: año al que corresponde la información documentada en el metadato (año del periodo de referencia). Ejemplos: COL-DANE-GEIH-2012 COL-Profamilia-ENDS-2010 COL-ICBF-ENSIN-2010

Source: DANE. Screenshot taken from Nesstar Publisher

6.2 DESCRIPTION OF THE STATISTICAL OPERATION

This includes general information on the statistical operation such as: how information should be cited; who collected, compiled and distributed data; data content; data collection and processing methods, etc.

Figure 4. Description of the statistical operation



Source: DANE. Screenshot taken from Nesstar Publisher

The following aspects should be documented within the description of the statistical operation:

Identification: the elements of this group refer to the statistical operation as follows:

Figure 5. Identification

Title
Subtitle
Abbreviation
Study type
:-2-3 Encuesta, fase I [h\123-:]
Translated title
Statistical operation identifier

Source: DANE. Screenshot taken from Nesstar Publisher

Title: it refers to the full name of the statistical operation, including the year of implementation (if relevant). Abbreviations should not be used in this box. The name of the statistical operation should be capitalized and include the year.

Figure 6. Title

Title
Gran Encuesta Integrada de Hogares 2012

Source: DANE. Screenshot taken from Nesstar Publisher

Subtitle: it corresponds to the secondary title of the statistical operation. It usually corresponds to the stage, the quarter in which the statistical operation was carried out or other similar information.

Figure 7. Subtitle

Subtitle

Source: DANE. Screenshot taken from Nesstar Publisher

Abbreviation: it corresponds to the abbreviation or known acronym that is most used to refer to the statistical operation. It should be capitalized without including the year.

Figure 8. Abbreviation

A screenshot of a web form with a light yellow header labeled 'Abbreviation'. Below the header is a white input field containing the text 'GEIH'.

Source: DANE. Screenshot taken from Nesstar Publisher

Study Type: the type of study, which appears in the drop down list is selected according to the subject and the collection or gathering method. The template uses a controlled vocabulary for this element. If none of the options listed is appropriate, the relevant category can be incorporated with a new entry in the editor.

Figure 9. Study type

A screenshot of a web form with a light yellow header labeled 'Study type'. Below the header is a dropdown menu with a blue border. The menu is open, showing a list of options: '1-2-3 Encuesta, fase 1 [hh/123-1]', '1-2-3 Encuesta, fase 2 [hh/123-1]', '1-2-3 Encuesta, fase 3 [hh/123-1]', 'Censo agropecuario [ag/census]', 'Censo de empresas [en/census]', 'Censo de población y vivienda [hh/popcen]', 'Cuestionario de indicadores básicos de bienestar [hh/cwiq]', and 'Encuesta agropecuaria [ag/oth]'. The first option is highlighted in blue.

Source: DANE. Screenshot taken from Nesstar Publisher

Translated title: if the title is in a language other than English, enter the English translation of the title. If the statistical operation was developed in a country with more than one official language, the title can be translated into the second official language, instead of English. Capitalizing the first letter is suggested, for example: Monthly Hotel Sample.

Identification number: it corresponds to the number or single text that identifies the statistical operation. One can use the identification number of one's choice or the abbreviation of the entity (DANE), the abbreviation of the responsible division (DIMPE), the abbreviation of the operation (GEIH) and the date of the statistical operation (2010).

Figure 10. Statistical Operation Identifier

A screenshot of a web form with a light yellow header labeled 'Statistical operation identifier'. Below the header is a white input field containing the text 'DANE-DIMPE-GEIH-2007'.

Source: DANE. Screenshot taken from Nesstar Publisher

General description: the elements in this group provide a general overview of the statistical operation.

Figure 11. General description

Summary
Statistical operation background and international benchmarks
Objectives
Reference framework
Statistical units

Source: DANE. Screenshot taken from Nesstar Publisher

Summary: this section provides a general idea, but specific of the various aspects that comprise the statistical operation, i.e., it must present an overview of what the operation or statistical research is about in a clear and summarized manner. This means making a clear and orderly exposition of the subject and the importance of its implications, as well as the manner in which the different elements forming it have been addressed.

Background of the statistical operation and international benchmarks: this section outlines the origin and historical development of the statistical operation, indicating its major milestones as well as its most relevant changes. This is vital to contextualize users on the progress made and the experience gained in conducting the statistical operation. In addition, it enables building a comparative view on the methodological changes that may affect the collection and analysis of its results. With respect to benchmarks, this section mentions international organizations that are leading agencies in the subject (the UN and its agencies, EUROSTAT, and national statistical institutes recognized internationally, among others) and presents the main recommendations adopted and / or adapted by the

statistical research or operation. In the case of new operations, it is important to contextualize the user on the origin of such operations and the benchmarks of studies that address similar themes or have similar characteristics as well as presenting the main conclusions of the pilot tests conducted in the design of the statistical operation.

Objectives: they establish the aim of the statistical operation and the purpose of the research topic.

Reference Framework: the framework consists of the theoretical framework, the conceptual framework, the legal framework and those subjects that may be deemed necessary to include, with the purpose of contextualizing the statistical operation in the best possible manner.

- a. *Theoretical framework.* It contains a summary of the review of the literature on the subject of the statistical operation. It describes the state of the art and the contributions made in the thematic field. This process allows obtaining the necessary arguments to define the research problem; knowing the theories that help locate the subject; better interpretation of obtained results; and seeking the generation of new approaches on how to address the problems.
- b. *Conceptual Framework.* It describes, explains and establishes the relationships between the fundamental concepts of the statistical operation. It corresponds to the basic concepts or technical terms used in it. The rest of the terms that are used are listed in a glossary in the “key word(s)” field. It is important that the concepts used in the statistical operation are standardized or harmonized jointly with DANE taking into account international benchmarks of organizations such as: the UN, Eurostat, OECD, or other statistical offices of leading countries in the subject. This exercise aims to achieve comparability, integration and interoperability of the statistical information.
- c. *Legal framework.* It describes the regulations in which the statistical operation is circumscribed.

Statistical units: this element describes the basic unit of analysis (individuals, households, establishments, enterprises, etc.). These include: the observation unit, the unit of analysis and the sampling unit. Each of the units of the study to be documented must be identified with their corresponding title written in capital letters.

E.g.: OBSERVATION UNIT

Farms producing flowers under greenhouse and open sky.

UNIT OF ANALYSIS

Farms and plots dedicated to the cultivation of flowers under greenhouse and open sky.

SAMPLING UNIT

Farms producing flowers under greenhouse and open sky.

Types of data: it describes the type of data (sample survey, census, statistical operation based on administrative records for statistical purposes, etc.) collected during the study. The template includes a controlled vocabulary for this element.

Figure 12. Types of data

Type of data
Datos clínicos (cli)
Operación estadística basada en registros administrativos (adm)
Datos agregados (agg)
Series históricas
Datos clínicos (cli)
Datos de transacción o evento (evn)
Datos obtenidos de la observación (obs)
Datos obtenidos por un proceso (pro)
Datos de manejo de tiempo (tbd)

Source: DANE. Screenshot taken from Nesstar Publisher

Thematic coverage: this element enables the documentation of the topics covered by the statistical operation.

Figure 13. Thematic coverage

Universe		
Target population		
Thematic content		
Thematic content		
Questionnaires		
keywords		
Text	Vocabulary	Vocabulary URI
Topic classifications		
Text	Vocabulary	Vocabulary URI

Source: DANE. Screenshot taken from Nesstar Publisher

Universe: it describes the set of units or individuals to which the statistical operation relates to or that constitute the group of interest and that meet a common definition.

Target Population: it describes the set of units or individuals to which the statistical operation relates. These units are defined in terms of content, space and time. This description includes a list of all groups excluded from the operation, and if relevant, the reason why they were excluded. As in the above elements, titles should be capitalized.

Thematic Content: it mentions the most important variables (or blocks of variables) of the statistical operation. In the case of researches having a large number of variables, it is necessary to classify them according to the major themes of the research. The titles should be capitalized to differentiate each one of the themes.

Questionnaires: this field identifies the name and type of questionnaire (not the content of the questionnaire) as well as the main modules. One of the following three terms must be used to describe the type of data collection instrument used:

- **Structured:** it indicates an instrument in which the same questions / tests, possibly with pre-coded answers were applied to all respondents. If a small portion of the questionnaire included open questions, appropriate comments should be provided.
- **Semi-structured:** it indicates that the research document contains mainly open questions.
- **Non-structured:** it indicates that in-depth interviews were conducted.

The titles should be capitalized. In the methodological document one can find the title as *Conceptual framework* and / or *Conceptual base* but the title should be: **THEMATIC CONTENT**. This also applies to questionnaires. If in the methodological document the title appears as *Design of Instruments, questionnaires or forms* the title should be: **QUESTIONNAIRES**.

Keyword (s): their purpose is to help individuals and organizations wishing to search and classify archived projects. Although there is no controlled vocabulary for this element in the DDI template, organizations will be able to create a unified keyword list to ensure consistency between projects.

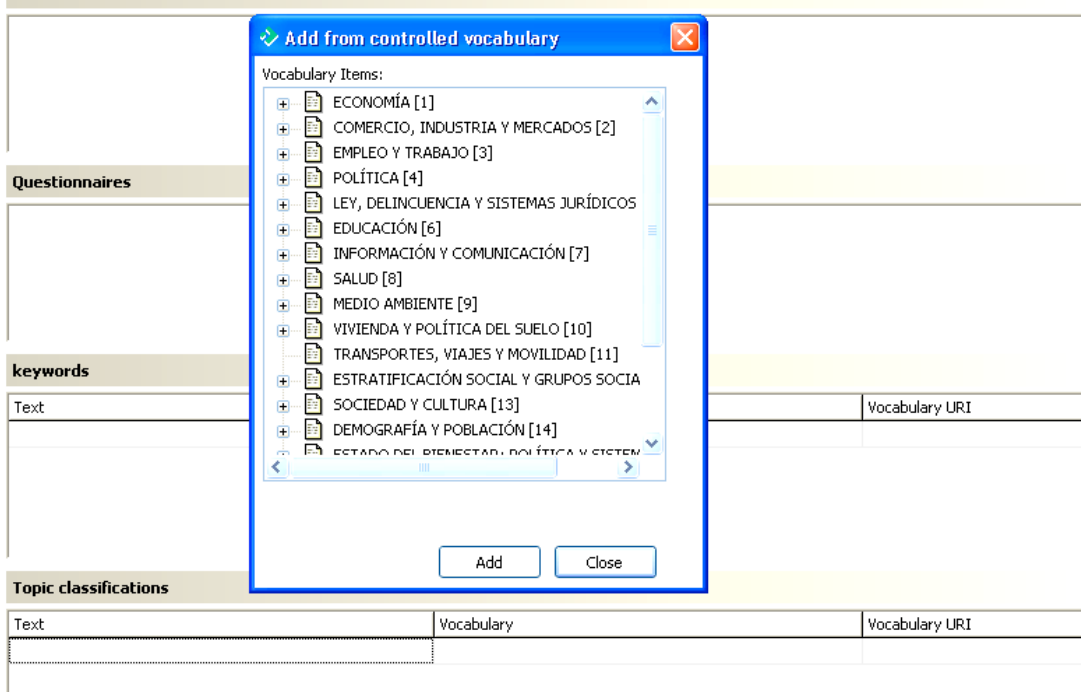
Figure 14. Keywords

keywords		
Text	Vocabulary	Vocabulary URI

Source: DANE. Screenshot taken from Nesstar Publisher

Topic classification: this box displays a list of all topics covered by the data. The list of terms provides users a list of default values to choose from. The “URI vocabulary” attribute specifies the location of the controlled vocabulary.

Figure 15. Classification of Topics



Source: DANE. Screenshot taken from Nesstar Publisher

Geographic coverage: the items in this group can be used to provide a description of the geographical area and population covered by the statistical operation.

Figure 16. Geographic Coverage

Countries	
Name	Abbreviation

Study domains	

Geographic unit	

Source: DANE. Screenshot taken from Nesstar Publisher

Country: all the countries included in the statistical operation must be listed. The DDI template uses a controlled vocabulary for this element. The country name is capitalized as well as the corresponding ISO code abbreviation which corresponds to the first three letters.

Figure 17. Country

Countries	
Name	Abbreviation
COLOMBIA	COL

Source: DANE. Screenshot taken from Nesstar Publisher

Study domains: this refers to a subgroup of the target population for which results are presented. These can be in a geographic area such as a region or a population center.

These could also include a category of a specific population, as a large national or ethnic group.

In sample surveys the number of domains has an important influence on the size and distribution of the sample. Normally, statistics are presented for different subgroups of the population, called study domains.

Chosen study domains can match the stratified sampling stratum or groups that are in them. These domains can be geographic or non-geographic. In general, these subgroups are associated with a classification (for example, the territorial units, economic activity, etc.) (Eurostat, “Assessment of the quality of statistics: Glossary”, Working Group, Luxembourg, October 2003).

Geographic unit: this element is used to provide information about the areas that were covered and to state the reasons the survey did not cover the total geographical area of the country. It aims to specify the lowest level of aggregation covered by the data.

Producers and Sponsors: this group contains elements that can be used to acknowledge the work of individuals and organizations responsible for the design, implementation and financing of the statistical operation. This group has four elements:

Figure 18. Producers and Sponsors

Primary investigator	
Name	Affiliation

Contributors			
Name	Abbreviation	Affiliation	Role

Fundings			
Agency	Abbreviation	Grant Number	Role

Acknowledgements		
Name	Affiliation	Role

Source: DANE. Screenshot taken from Nesstar Publisher

Primary investigator: this person is responsible for designing and conducting the statistical operation. It is important to note that the study may have more than one primary investigator. The name of the area producing the research can also be entered.

For example:

Methodology and Statistical Production Division – DIMPE

Contributors: persons or organizations involved in the various stages of the statistical operation (design, data collection on field, processing and analysis). Information should be entered as follows:

Name: division name - the name of the statistical operation (First letters of each word must be capitalized).

Abbreviation: the abbreviation of the division – statistical operation must be written in capital letters.

Affiliation: it refers to the entity and its acronym. The first letters of the name of the entity should be capitalized.

Role: it refers to the Technical Team. The first letters should be capitalized. (See Table 1).

Table 1. Contributors

Name	Abbreviation	Affiliation	Role
Methodology and Statistical Production Division – DIMPE – Great Integrated Household Survey	DIMPE - GEIH	National Administrative Department of Statistics - DANE	Technical Team

Source: DANE

Funding: those organizations responsible for financing the statistical operation should be listed in this field. If possible, the grant or contract number should be reported.

Agency: name of the entity.

Abbreviation: acronym of the entity.

Role: it would be *executant* if one is the producer of the operation with their own resources (See Table 2).

Table 2. Funding

Agency	Abbreviation	Grant number	Role
National Administrative Department of Statistics – DANE	DANE		Executant

Source: DANE

Acknowledgements: individuals and / or institutions that contributed to the production of the statistical operation: technical assistance, baseline studies, etc.

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

Figure 19. Sampling

Sample design and size definition
Major Deviations from the Sample Design
Estimation procedure
Calculation of the accuracy of results

Source: DANE. Screenshot taken from Nesstar Publisher

Sample design and size definition: this element provides information on the sampling frame and the methods and procedures used to select respondents. The desired sample size should also be mentioned. The titles of each of the themes should be capitalized

Major deviations from the sample design: this element is used to describe the correspondence between the units that were successfully surveyed and the planned sample. Any significant deviation should be mentioned here.

Estimation procedure: it explains the estimators that are used to obtain indicators of the statistical operation. It defines and justifies the selected methodology and its components. It indicates the calculation of the expansion factors and elements that determine the expansion. Each one of the titles used in the documentation of this field should be capitalized.

Calculation of the accuracy of results: it describes the methodology used to estimate sampling errors and their presentation, in order to determine the level of confidence. It reviews the design of the variance estimation method. It presents the formulas for calculating the standard error and / or the coefficient of variation of the estimators. Subtitles used in the documentation of this field should be capitalized.

Data collection: a statistical operation properly documented will be accompanied by a full report on the activities of data collection and a copy of all the questionnaires used during its execution. This group of elements is intended to provide a brief summary of the activities on data collection.

Figure 20. Data Collection

Organization and preparation		
Dates of Collection		
Start	End	Cycle
<input type="text"/>	<input type="text"/>	<input type="text"/>
Reference period		
Start	End	Cycle
<input type="text"/>	<input type="text"/>	<input type="text"/>
Method of Data Collection		
Face-to-face [f2f]		
Notes on Data Collection		

Source: DANE. Screenshot taken from Nesstar Publisher

Organization and preparation: this field must incorporate the elements associated with the processes of awareness-raising, training and the operational scheme. The subtitles should be capitalized.

Collection dates: the start and end dates of data collection indicating the collection cycle (monthly, quarterly, weekly, etc.) should be entered. If the month and the day are not available, the year can be entered. Dates must be entered in ISO format, e.g. YYYY-MM-DD (or YYYY-MM or YYYY). It is important to take into account that this date should be subsequent than the one reported in the reference period.

Table 3. Collection dates

Start	End	Cycle
2009-11-01	2009-12-30	Bi-monthly

Source: DANE.

Method of data collection: select the main mode of data collection from the list provided on this element.

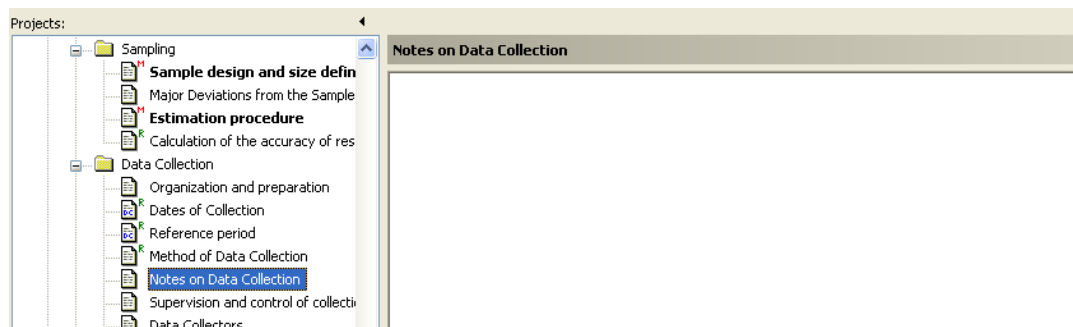
Figure 21. Method of Data Collection

The screenshot shows a dropdown menu titled "Method of Data Collection". The menu is open, displaying a list of options. The first option, "Otro método", is highlighted in blue. The other options listed are: "Entrevista personal asistida por computador", "Entrevista por convocatoria al informante", "Entrevista telefónica", "Autodiligenciamiento por correo", "Autodiligenciamiento de formulario electrónico vía página web (por selección; por ejemplo en encuestas por muestreo o censos)", "Autodiligenciamiento de formulario electrónico vía página web (libre; por ejemplo encuestas de satisfacción u opinión)", and "Grupo focal".

Source: DANE. Screenshot taken from Nesstar Publisher

Notes on data collection: this field must specify aspects regarding the data collection method, especially if the “other method” option was selected. Important aspects that may have arisen during data collection may also be described. Factors such as respondent cooperation, interview length, number of visits and other events and occurrences should be included.

Figure 22. Notes on data collection



Source: DANE. Screenshot taken from Nesstar Publisher

Supervision and control of the data collection operation: it presents the controls applied on field to staff under the field supervisor’s or coordinator’s charge, in aspects such as: routes, coverage of geographical areas and allocated units, equipment and devices, security of completed questionnaires. It also includes controls to reduce bias in situations such as: total or partial rejection; no response; lack of coverage; revisits; errors in the frame; nonexistence of the source; address error; address change; filling out coverage surveys and verification of the veracity of the information collected with the most relevant variables for quality assessment. It will also include a summary of the procedures followed to minimize errors, reduce information loss and prevent forgery. These procedures may involve reviews by supervisors of completed questionnaires on field, re-interviewing respondents and other similar activities.

Reference period: it indicates and justifies the time interval (years, months, weeks, days) to which the information refers to, indicating its cycle (monthly, quarterly, weekly, etc.) If only the year is available, this can be entered without the month and the day. Dates must be entered in ISO format, e.g. YYYY-MM-DD (or YYYY-MM or YYYY). It is important to take into account that this date should be previous to the one reported in the collection period.

Table 4. Reference period

Start	End	Cycle
2009-07-01	2009-09-30	Quaterly

Source: DANE.

Data Collector: it provides a list of persons and organizations responsible for managing the questionnaires and collect data. This refers to entities that collect the data, not to the ones that produce the documentation. The first letters of the name of the entity must be capitalized as well as the abbreviation and the affiliation. The affiliation corresponds to the national government for public entities at the national level; for the case of departmental secretaries, their affiliation will be departmental; and for the case of professional associations the affiliation will be association. See Table 5.

Table 5. Data Collection

Name	Abbreviation	Affiliation
National Administrative Department of Statistics	DANE	National Government

Source: DANE.

Data processing: data validation is the only element in this group. An overview of the procedures used to identify and correct errors in the data is entered in this field. Among the topics included are:

- File consolidation.
- Validation and consistency rules.
- Verification of the internal consistency of the data and adjustments.
- Imputation and / or coverage adjustments.

Evaluation of results: it describes data accuracy estimators of the statistical operation (sampling error, coefficient of variation, confidence intervals, etc.).

Other forms of data validation: it refers to other observations related to the validation of data. For example, the variance of the responses; sample bias tests; Interviewer or response bias; confidence intervals, etc.

Database access: the items in this group provide a summary of the conditions under which a database study can be accessed.

Figure 23. Database access

Access Authority			
Name	Affiliation	E-Mail	URI
Confidentiality			
Access and use conditions			
Citation Requirement			

Source: DANE. Screenshot taken from Nesstar Publisher

Access authority: it must contain the names of persons or organizations responsible for granting access to data and documentation.

Table 6. Access authority

Name	Affiliation	E-mail	URL
National Administrative Department of Statistics	National Government	dane@dane.gov.co	www.dane.gov.co

Source: DANE

Confidentiality: it describes the norms establishing the anonymity of the informant and the commitments of the institution in order to ensure the confidentiality of the results.

Access and use conditions: it provides a description of the terms under which users are allowed access to the data of the statistical operation. For example, some databases can

be obtained free of charge on the Internet, while others can only be accessed from designated computers in special facilities.

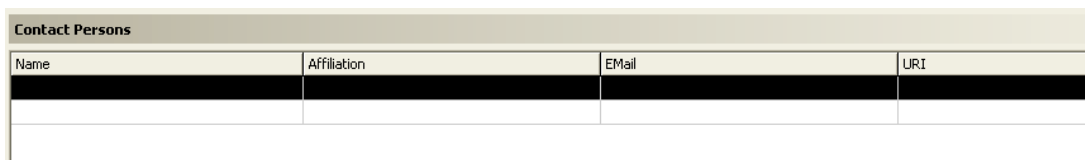
Citation Requirements: it presents the accreditation of the data source in accordance with norms on citation procedures of the corresponding country.

Disclaimer: it provides information related to the responsibility of users of documentation and databases.

Copyright: it refers to the copyright statement of the results with the purpose of respecting the authorship and intellectual property, providing protection to the sources of information as well as to the entity generating the indicators.

Contacts: this group can be used to provide users with information on who can answer additional questions about the statistical operation.

Figure 24. Contact persons



Contact Persons			
Name	Affiliation	EMail	URI

Source: DANE. Screenshot taken from Nesstar Publisher

Contact persons: this field must contain the name of the persons or agencies responsible for the statistical operation and how to contact them. These persons are usually the principal investigators, but other persons who could answer questions on the design of the operation and data collection could be included.

6.3 DATABASES

This section contains all the data files associated with the statistical operation, along with a detailed description of each variable in the data files. The elements in this section are obtained from the DDI specifications. Each data file has four groups of elements: file description, key variables and relationships, variables and data entry.

File description: The elements in this group are used to provide basic information about each file. The DDI template has the following elements in this group:

Figure 25. File description



Source: DANE. Screenshot taken from Nesstar Publisher

Contents: this element must refer only to the selected data file and not the statistical operation. It provides a description of databases including thematic coverage, special characteristics of their content and main variables.

Producer: this field must contain the name of the person or organization responsible for the creation of the database. It should be noted that the producer of the data file is not necessarily the agency that collected or processed the data.

Example: National Administrative Department of Statistics (DANE)

Version: this element shows a description of each version of the data file. The information should include the date, origin and type of data (“validated”, “partly validated”, “original”, etc.). Ideally, a version number should be assigned to each data file. Using a formal numbering is recommended to identify different publications of the database. The word version and the year of the statistical operation must be entered, as follows: Version 2005.

Missing data: this element is used to describe the causes for the existence of missing data in the data file. The particular lost data are defined when documenting each variable.

Notes: any additional comments about the data file can be added in this field.

Key variables and relationships: most of statistical operations include multiple files of related data. This group of elements is designed with the purpose of defining the relationship between these files. There are two types of files used to define the relationships between data files: base key variables and external key variables.

- **Base key variables**: these are variables that identify individually, each observation in the data file. For example, a key base variable in the household file would be the variable that contains a single identification number for each household. If the data file does not contain this variable, a new one can be created through a single combination of variables (e.g. the combination of variables such as region, area and the identification number of the household in the area).
- **External key variables**: it refers to those cases in which the variables that can be used to link databases do not include a variable with a single identification number for each observation. These variables allow merged files not to contain duplicate values.

It is important to create key variables and define the relationships between databases in a project in a proper manner since the key base variables will be used by data analysts to merge the files in STATA, SPSS, SAS or any other statistical software.

Variables: this group of elements is used to enter metadata related to each variable.

List of variables: this field must contain the following information on all the variables in the databases:

Figure 26. Variables

Variables							
Number	Name	Label	Width	StartCol	EndCol	Record	Decimals

Source: DANE. Screenshot taken from Nesstar Publisher

- The **name** is the primary means to refer to a variable. It must not have more than 8 characters; it must not start with a number; and must not contain blank spaces.
- The **label** is the description of the variable. This description should be brief but detailed.
- The **width** is the maximum number of characters that can be included in a variable.
- The **decimals** correspond to the number of places used by decimals for each variable.

Variable description: this option has fields to define value or category labels of a variable and to enter other information, such as missing data or measurement level. Categories enable the definition of values labels of a variable.

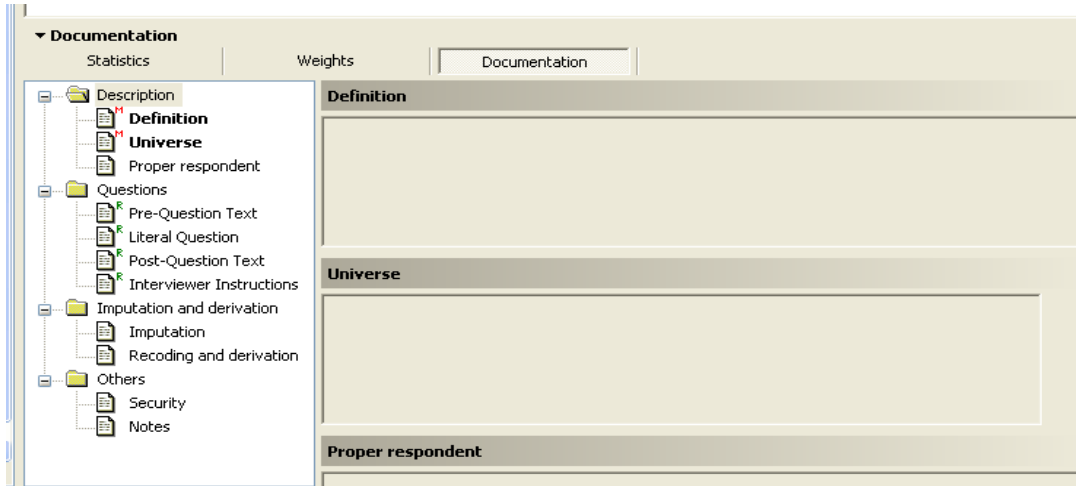
Documentation: it provides information about the database and allows creating and saving statistical summaries such as metadata, which facilitates the entry of more detailed information on the variables that require it.

Statistics: Nesstar Publisher calculates and stores various common statistical measures such as: number of valid weighted and unweighted cases; minimum and maximum values; mean and standard deviation weighted and unweighted. These statistics are stored as metadata and this enables frequencies and other summary measures to be reported without publishing microdata.

Weights: when documenting the statistical operations, variables can be specified as weights and this weighting can be applied to individual variables when appropriate. Thus, weighted statistics can be calculated and shown in the statistics summary.

Documentation: the documentation of the variable included in the statistical standards of data files is often limited to the variable label and value labels. Nesstar Publisher provides more elements for the documentation to be complete, as follows:

Figure 27. Documentation

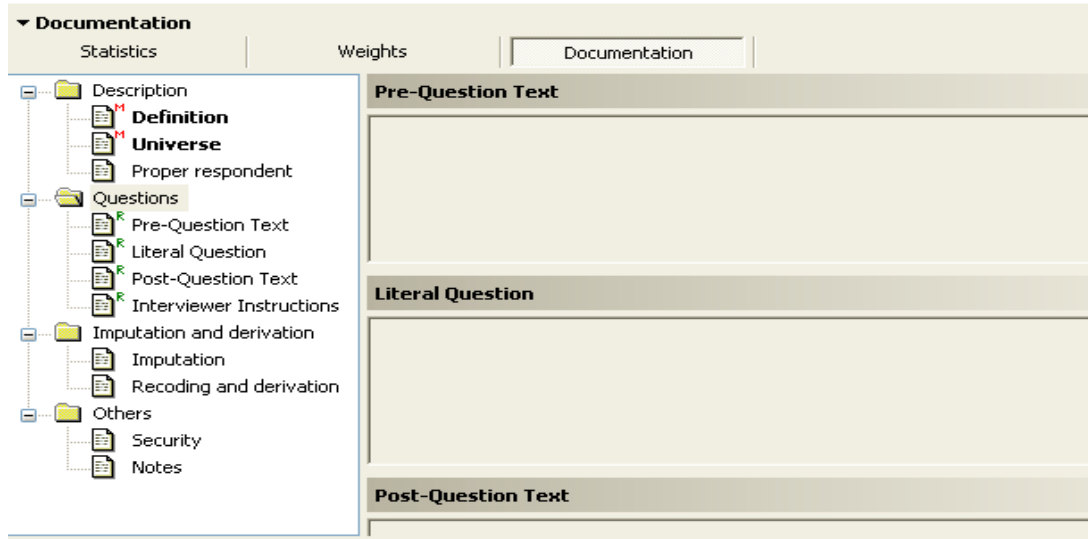


Source: DANE. Screenshot taken from Nesstar Publisher

Description

- The **definition** element allows the variable to be described in greater detail than the variable label.
- The **universe** element allows users to specify the exact population to which the variable applies.
- The **proper respondent** element must document information regarding the person / entity providing the information on the variable. In household surveys, the source may be the head of household or a household member, but it can also be the interviewer through visual observations, the manager or legal representative of a company, etc.

Figure 28. Questions



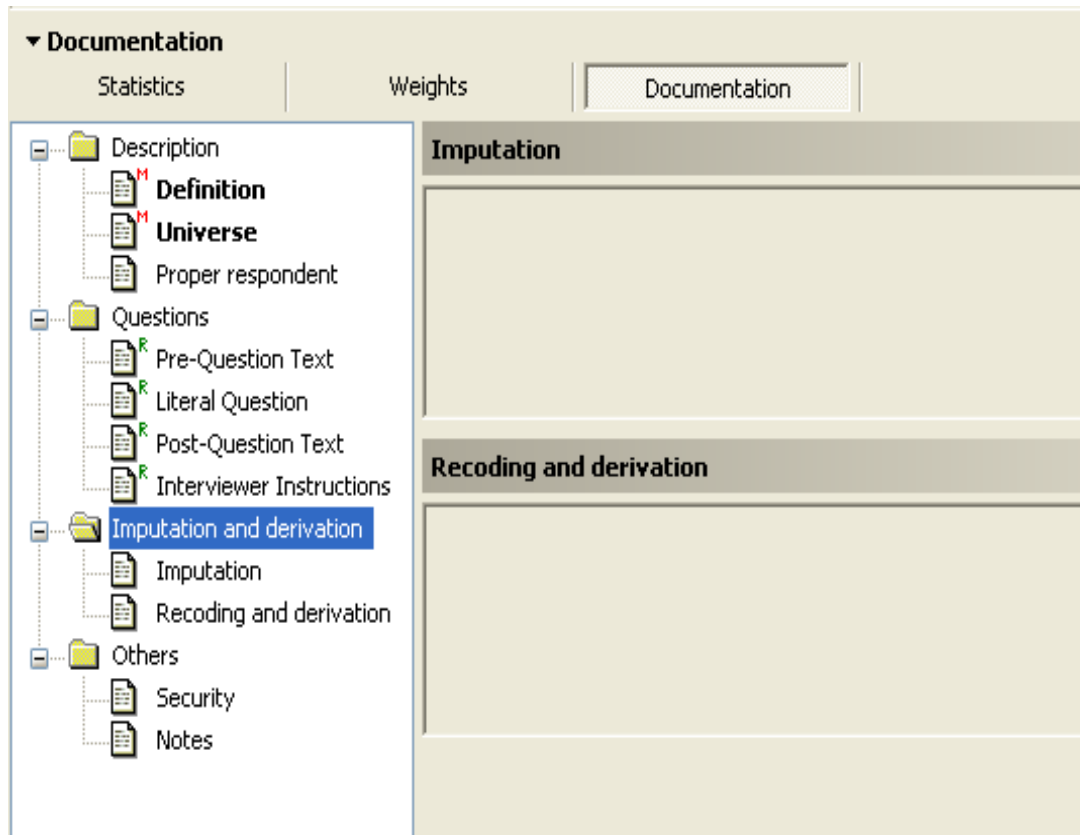
Source: DANE. Screenshot taken from Nesstar Publisher

Questions:

- The **Pre-question Text** field can be used to describe the immediately preceding question, according to the conditions of the form.
- Enter the exact wording of the question in the **Literal Question** element. This is the most important element in the Description section of variable documentation.
- A text that follows the question asked in the questionnaire and serves as a guide to the interviewer is entered in the **Post-questionText** field.
- The **interviewer Instructions** element must contain a guide for the interviewer. This information will typically be copied from the interviewer manual.

Imputation and derivation:

Figure 29. Imputation and derivation



Source: DANE. Screenshot taken from Nesstar Publisher

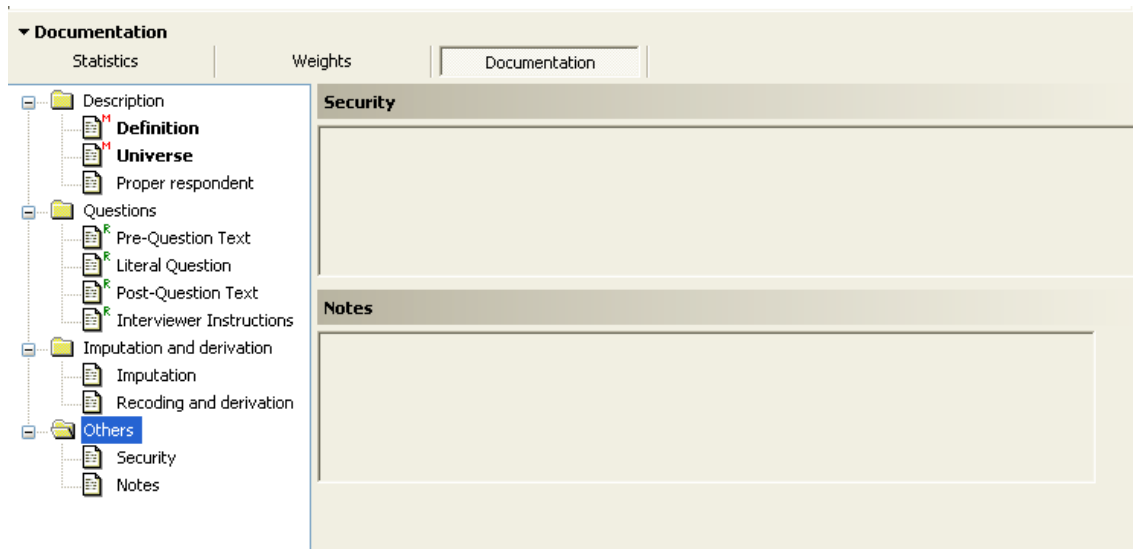
Many data files shall include derived or generated variables, in addition to variables with collected data. Additionally, some variables may include imputed values. Documenting those imputations is crucial to build trust and ensure that users can reproduce the data construction.

If missing data have been replaced with estimates, then the process used to make these estimates should be described in as much detail as possible on the imputation element. This element may include a reference to a more detailed technical document.

Variables can also be obtained by recoding or combining other variables. In such situation, the recoding and derivation element must contain a clear and complete description of all actions carried out to prepare the variable.

Others:

Figure 30. Others



Source: DANE. Screenshot taken from Nesstar Publisher

- Use the **Security** element to describe the level of appropriate access for a variable.
- The **Notes** element may be used to indicate any other information about the variable not mentioned anywhere else.

Group of variables: Data files may include hundreds of variables. Nesstar Publisher provides a tool for organizing variables into groups. Variable grouping allows users to browse through variable listings quickly and helps to control the analysis providing an indication of which items in a database are conceptually connected. A variable can belong to more than one group and a group of variables can include variables from more than one data file.

The definition of a group of variables is a two-step process.

Step 1: the group is created.

Step 2: the variables are added to the group.

6.4 EXTERNAL REFERENCE MATERIALS

This section is used to provide a listing and description of materials such as documents (manuals, questionnaires, technical and analytical reports), computer softwares (data entry, editing, tabulation, analysis), and photos and maps related to the statistical operation. DCMI specification is used in this section and, unlike data files, external reference materials are not stored in the Nesstar Publisher file.

Only metadata describing these resources are saved with the project.

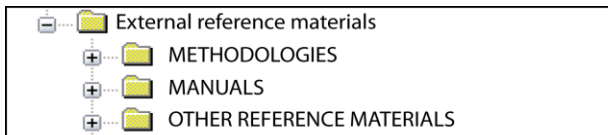
Figure 31. External reference materials

Types
technical
Title
Subtitles
Authors
Date
Year: <input type="text"/> Month: <input type="text"/> Day: <input type="text"/> <input type="text"/> Hours: <input type="text"/> Minutes: <input type="text"/>
Country
Languages
Format
Identification number

Source: DANE. Screenshot taken from Nesstar Publisher

Reference material can be classified into different folders. These should be named according to the types of documents that they will contain, as shown in the following graph:

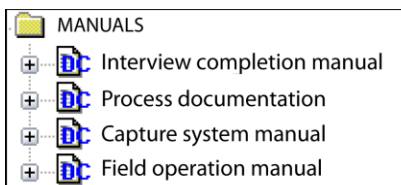
Figure 32. Example of presentation of external reference material



Source: DANE. Screenshot taken from Nesstar Publisher

According to the above, the name of the documents contained in each one of the folders must have their first letter capitalized and it must be representative, as shown below.

Figure 33. Example of presentation of documents' names

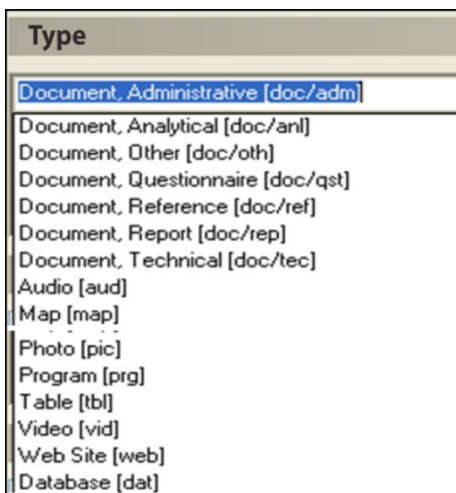


Source: DANE. Screenshot taken from Nesstar Publisher

Identification: this group consists of eight elements that are set forth below:

1. *Type:* this element is used to indicate the type of reference material that is being documented.

Figure 34. Type of reference material



Source: DANE. Screenshot taken from Nesstar Publisher

2. *Title and subtitle*: a formal name and an optional name that is secondary to the reference material are entered in these fields.
3. *Author*: this element shows the name of individual persons or organizations responsible for creating the reference material. Full names must be used and writing abbreviations or acronyms in parenthesis after the name of the organization is recommended.
4. *Date (optional)*: this element shows the full or partial date in which the reference material was created or was last modified.
5. *Country*: this element lists all countries within the scope of the reference material.
6. *Language*: this element lists all languages in which reference materials appear.
7. *Format*: an object from the list is selected in this field in order to identify the file format of the reference material.
8. *Identification number*: if it exists, the identifier of the reference material must be presented.

Contributors and legal rights:

Contributors: in this field it is necessary to cite the persons or organizations who have supported or contributed to the development of the reference material.

Rights: this field is used to provide information about the legal rights of reference materials.

Content:

Description: this field is used to briefly describe the content of the reference material.

Summary: it provides summary information about each of the main aspects of the reference material.

Table of Contents: this field is used to list all the sections of the report, questionnaire and other documents.

BIBLIOGRAPHY⁴

Departamento Administrativo Nacional de Estadística (National Administrative Department of Statistics) (DANE). (2012a). Guía para documentar y codificar documentación técnica. Sistema Integrado de Gestión Institucional. (Guide for documenting and coding technical documentation. Institutional Management Integrated System). Bogotá: DANE.

_____. (2012b). Guía para la elaboración de protocolos para la actividad estadística. (Guide for the development of protocols for the statistical activity) DIRPEN. Version 2.0. Bogotá: DANE.

_____. (2010c). Instructivo para la elaboración de guías (Manual for the development of guides). Sistema Integrado de Gestión Institucional (Institutional Management Integrated System). Bogotá: DANE.

_____. (2012d). Instructivo para la elaboración de guías (Manual for the development of guides). Sistema Integrado de Gestión Institucional (Institutional Management Integrated System). Bogotá: DANE.

_____. (2012e). Resolution 1503 of 2011. Bogotá: DANE.

Organization for Economic Cooperation and Development (OECD) (sf). Glossary of statistical terms. Retrieved in 2012 from: <http://stats.oecd.org/glossary/detail.asp?ID=532>

⁴ The translation of the bibliographic titles is for information purposes only.