Food and Agriculture Organization
of the United Nations

METADATA AND MICRODATA CURATION AND DISSEMINATION PROTOCOL

Contents

Figures

Acknowledgements

This document was originally prepared by Aliou Mballo, Michael Rahija, and Anidi Oluwakayode during 2019 in the Office of Chief Statistician. It was develop in order to define the processes and procedures for populating the Food and Agriculture Microdata (FAM) catalogue. Financial support for the work of Aliou Mballo and Anidi Oluwakayode was provided by the Bill and Melinda Gates Foundation, and United States Agency for International Development.

The World Bank microdata library team especially, Matthew Welch, provided a lot of help by sharing similar documents which govern the use and management of the World Bank microdata library.

1. Background, concepts, and definitions

1.1 Background

Data collected through surveys, and administrative systems form the foundation of official statistics, and are an invaluable source for research. They are aggregated to generate national estimates by official statisticians, and analyzed by researchers and policy analysts to gain scientific insights which can be translated into policy. These data are commonly referred to as microdata defined as to unit-level information on individual people or entities (such as individuals, households, business enterprises, farms, or even geographic areas).

The power of microdata stems from its granularity. Because microdata contain individual level information, they allow an analyst to investigate the unique ways a certain phenomenon may effect sub-populations. For example, a particular agricultural policy may effect male and female agricultural holders differently. Likewise, a social protection scheme may benefit a particular demographic and disadvantage another. This type of analysis is impossible without highly granular datasets which allow for the analyst to stratify a dataset by a one or more variables.

It is important to note here that this protocol incorporates the FAO Personal Data Protection Principles (AC No. 2021/01) and it is in accordance with the principles.

1.2 Terms, concepts and supporting platform

This document will use the following terms with the corresponding definitions.

Data provider is the technical unit, officer, or any individual that submits a micro dataset to be published through the data deposit system**.**

Data user refers to an individual or institution which accesses the micro datasets disseminated in the FAM catalogue or consults the metadata shared on the platform.

Data curator is the officer or unit that will process and prepare the information submitted by the data provider for dissemination. In most all cases, the data curator will be an officer of OCS.

Study refers to the main topic of interest; that can be for example, the name of a survey or data collection operation.

Micro dataset is a dataset which contains unit-level information associated with an individual, group of individuals, or legal entities collected for statistical, research, or scientific purposes. In the case of FAO, microdata come almost exclusively from household or farm surveys. However, microdata can come from administrative reporting systems, or other types of registers as well. This is not to be confused with national level time series aggregated datasets which represent the majority of other datasets FAO publishes (e.g. FAOSTAT, Aquastat, Fisheries and Aquaculture, etc.).

Metadata are data about data, or information which describes a particular dataset. In the context of microdata, metadata describes important information about how the microdata were collected. The Data Documentation Initiative (DDI) is the metadata standard which FAO will use for archiving microdata. The DDI items and elements are described in Section 2.

The **FAM catalogue** is the IT platform, which FAO will use to disseminate microdata accessible at https://microdata.fao.org. Many of the processes described in this document will actually be

implemented using FAM. FAM is based on the National Data Archiving (NADA) [6] platform developed by the World Bank in order to assist countries in publishing household survey datasets, and provide a global web-based cataloguing system where users can search all country level NADA instances for survey data. The DDI XML based metadata standard, and the Resource Development Framework (RDF) are built-in to NADA to maximize interoperability.

The **data deposit system** is a back-end component of the FAM catalogue providing a user interface which allows data providers to submit datasets, along with required metadata, and related materials (i.e. reports, questionnaires, manuals, etc.).

The **Internal archive** is the place where all of the microdata, related materials, and programs for processing will be stored.

The **metadata publisher** is the software used to edit, or create the DDI metadata document in XML.

DDI XML Document is a file generated by the metadata publisher that contains information related to the metadata and that will be translated to XML format.

Related materials are all additional documents such as questionnaires, technical reports etc. that correspond to a study which can help users analyze the corresponding micro datasets.

External contributor is a country, NGO, or any other institution willing to share their metadata in the FAM catalogue and/or link their micro datasets to FAM. This is different from a data provider, as no microdata is deposited in FAM but only accessible via the platform of the contributor. This could also be a country that publishes microdata under an open data license which allows free redistribution.

1.3 Motivation

FAO and member countries increasingly rely on micro datasets collected through agricultural surveys and censuses for monitoring and evaluation, tailor programming and policy interventions, conducting research, and monitoring important development trends and indicators such as small holder resilience and animal disease. Furthermore, the demand for highly granular and disaggregated microdata is exacerbated by the SDG indicators related to hunger (2.1.2), small productivity (2.3.1, 2.3.2), agricultural sustainability (2.4.1), livestock (2.5.2), women's ownership of agricultural land (5.a.1), food loss (12.3.1) and many others. In this regard, FAO has ongoing donor and regular program funded activities which collect farm and household survey data.

These activities yield rich micro datasets that can and should be leveraged by stakeholders inside and outside FAO for a variety of uses. For example, a micro dataset on decent work in Kenya produced by the Statistics Division (ESS) could be used by the Social Protection Division (ESP) to design policy recommendations on decent work, child labor, etc. Another example from the ESS, comes from a survey of vegetable producers in Ghana. This micro dataset contains detailed farm level data on horticultural producers which could be used by the Technical Cooperation and Investment Division (TCI) to design targeted interventions to increase productivity or income for Ghanaian vegetable producers. The potential value of insights that could be gained by outside stakeholders such as researchers and policy analysts are unknowable, and as a producer of global public goods, FAO has a responsibility to put them in the public domain.

[6] http://www.ihsn.org/nada

Due to the importance of disseminating microdata, the Office of Chief Statistician (OCS) advocates for the dissemination of all microdata that FAO collects directly, and promote the dissemination of micro datasets for which it provides assistance, supports (i.e. technical, financial, or otherwise), and/or is relevant to FAO's mandate. In this regard, during 2019, OCS coordinated the development of the Food and Agriculture Microdata (FAM) Catalogue to provide officers with a platform to disseminate microdata and a Statistical Disclosure Control (SDC) Protocol which define the process of anonymizing microdata.

The purpose of this document is to define all the steps that will be undertaken to publish micro datasets, roles and responsibilities, and quality standards. Technical details of specific methods will not be described in this document, but literature references will be provided as needed. SDC is mentioned as a step in the overall dissemination process, but due to its complexity, it has its own document.

This document is divided as follows:

- Section 1: Background, concepts and definitions
- Section 2: Metadata Standards
- Section 3: Microdata Formats and Quality Assurance
- Section 4: User Access
- Section 5: Metadata and Microdata Workflow

2. Metadata standards

Metadata is concerned with the accessibility principle of the FAO Statistical Quality Assurance Framework[7]. This section will describe what constitutes high quality metadata, provide an overview of the DDI standard, define different acquisition processes, and finally show how the DDI standard is implemented in the metadata production process for the FAM.

2.1 What is metadata?

Metadata should contain all the information users need to analyze a dataset, and draw conclusions. It increases data accessibility by summarizing the most important information (i.e. methodology, sampling design, interview mode, etc.) required for analyzing a dataset which alleviates the need for users to search for supporting documents and reports. Furthermore, good metadata clearly articulates the potential uses for a dataset, preventing potential misuses. Metadata is also a tool for rendering complex microdata structures into something meaningful, navigable, and user-friendly. Finally, the adoption of well-known metadata schemas and vocabularies allows for semantic interoperability.

At a minimum, metadata should answer the following questions:

- What were the objectives of the study?
- Where and why did the study take place?
- What were the main tools used to implement the study?
- What were the outputs of the study: datasets, reports etc.?
- What was the workflow of the data production process?
- What were the different institutional bodies involved in the overall process?

The Metadata process is fully integrated in the Generic Statistical Business Process Model (GSBPM)[8] which has metadata as one of the key element in the version 5.1.

Figure 1: Generic Statistical Business Process Model (GSBPM)

The arrows forming the outside of the cycle describe each step in the GSBPM. One can see from Figure 1 that the metadata records information about every step of the statistical process. Accordingly, a good practice developing metadata is to fill all the elements at each stage. In that case, the time required for compiling it will be minimized and important information will be less likely forgotten.

2.2 The Data Documentation Initiative (DDI)

To achieve a clearer understanding and interpretation of different metadata among diverse set of users (including humans and machines), it is important to use the same terminology (i.e. vocabulary). Hence, the Data Documentary Initiative (DDI) was selected as the metadata standard for FAO's microdata dissemination process.

2.2.1 Benefits of DDI

The most important benefit of adopting DDI is **semantic interoperability**. Semantic interoperability allows all the platforms which use the DDI standard to communicate and share information. This improves the discoverability, and visibility of datasets across platforms. It also facilitates the harvesting of metadata as described in Section 2.3.2.

In addition to interoperability DDI offers the following attractive features:

Repurposing: DDI provides a core document from which different types of outputs can be generated.

Support for online analysis: DDI standardizes variable structure, facilitating import into online analysis and sub-setting systems.

[8] UNECE : United Nations Economic Commission for Europe, https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

Enhanced data discovery: DDI enables field-specific, granular searches.

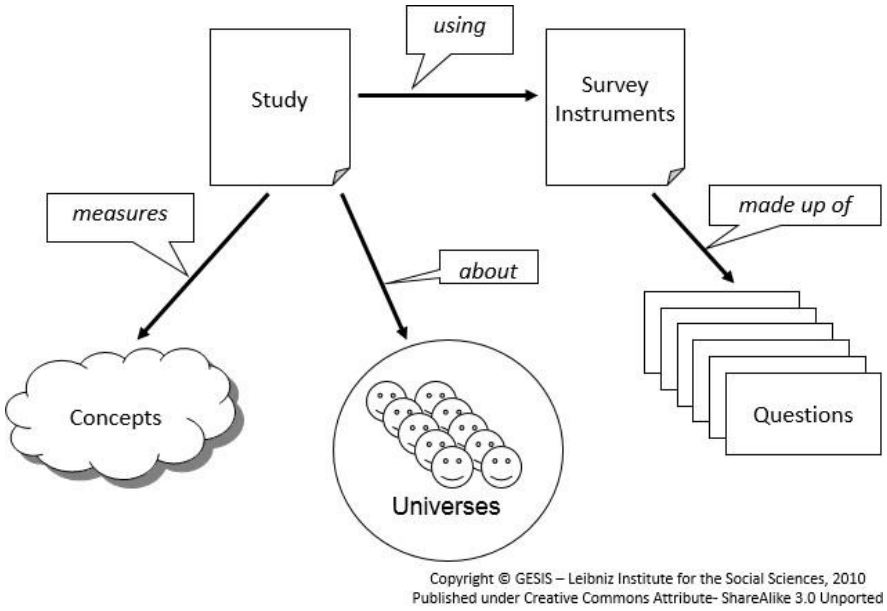
Multiple presentation formats: XML-based DDI can be rendered as ASCII, PDF, HTML, or RTF.

Preservation: XML documents are nonproprietary, containing embedded "intelligence" and optimal for long-term preservation

2.2.2 DDI Structure (version 2.5)

The structure of the DDI (illustrated in Figure 2) allows the description of all the aspects of the study to give the user all the information he/she needs. This includes but is not limited to methodology, dataset structures and formats, contact and copyright information, etc.

Figure 2: What is the Data Documentation Initiative trying to answer?



In this regard, the DDI standard contains a list of hundreds of elements and attributes each of which correspond to a specific piece of information about the study. For example, "Abstract" includes a high level description of the study. "Mode of Interview" indicates how the interview was performed (i.e. face-to-face, telephone, etc.). Each element can be defined as mandatory and optional. Some elements can be repeated/multi-answers (e.g. institutions involved in the process). Also a controlled vocabulary can be defined to avoid wrong entries and facilitate searching in the platforms (a list of controlled vocabularies can be found here: http://www.ddialliance.org/controlled-vocabularies).

Each metadata file has four sections which follow DDI elements and one which follows the Dublin Core Metadata Initiative (DCMI). The DCMI is a metadata standard for archiving related materials such as questionnaires, technical reports, manuals, scripts, and/or programs. It was initiated the same year as the DDI initiative by the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) in Dublin, Ohio. The use of the DCMI is supported by the fact that it makes the documentation of related materials easy because of its simplified design and can be used by non-specialist (Dupriez and Boyko, 2010).

The sections of every metadata document are as follows:

1. ***Document Description*** **-** This section is commonly called "metadata about metadata". It describes the metadata on the documentation process by providing information on the DDI XML document itself.
2. ***Study Description*** **-** This section includes DDI elements providing an overview of the study such as scope, data collection methods, abstract of the study, keywords, citation, data processing etc.
3. ***Data file Description*** **-** This section describes the contents of the microdata files including record and variable counts, etc.
4. ***Variable Description*** **-** This section provides details for each variable including the question text, universe, variable and value labels, derivation, imputation methods, summary statistics, etc.
5. ***Related Materials*** **-** This section is used to describe external resources which may vary across studies such as technical reports, questionnaires, interviewers' manual, photos, programs etc. It does not follow the DDI standards, but use the DCMI standard to describe digital resources on the web and is approved as an ISO standard.

2.3 Acquisition of metadata

There are two ways of acquiring metadata: (i) through the data deposit system; (ii) harvesting metadata from external sources e.g. catalogs, statistical websites and study reports. In both cases, a folder corresponding to the study is created in the internal archive containing the DDI XML document, and all related materials. Details on the organization, naming schemes, and structure of the internal archive are described in Section 5.1.2.

2.3.1 Receiving metadata through the data deposit system

The data provider should prepare the metadata for his/her study by completing all of the information specified in the data deposit system. In this regard, the data deposit system includes fields corresponding to each DDI element with a user friendly description and a classification of mandatory or not. Furthermore, the data provider can upload related materials, and provide information related to Statistical Disclosure Control (see SDC Protocol). Annex 3 provides screenshots and examples of how the data deposit works.

2.3.2 Harvesting metadata from external sources

Many National Statistical Offices (NSOs) and institutions publish micro datasets containing information within the scope of FAO's areas of work. If these datasets are published according to the DDI standard, FAO can download the DDI file, and re-publish the metadata in the FAM Catalogue. In other words, thanks to the interoperability of the DDI, the XML file can be extracted from the external catalogue (herein referred to as the "parent platform") and loaded into the FAM catalogue (in this case the "child platform"). This practice is advantageous for both FAO, and the original publisher by making the datasets easier to find (i.e. enhancing discoverability).

The first step in the process is asking permission to re-publish metadata to the custodian of the parent platform. If permission is granted, then the data curator downloads the DDI XML file, and opens it in the metadata publisher to review for accuracy, compliance and completeness. To determine the later, the data curator reviews published reports, and any materials available on the website of the custodian of the platform. This procedure of fact-finding from reports and materials is also done in

cases where the XML file is not available for download or when there is no metadata available for the data. For the later, the metadata has to be created from scratch. The data curator then sends a draft version of the newly created/edited metadata in .pdf, and XML to the custodian asking for confirmation/validation; especially if major edits were introduced to the metadata. Lastly, the curator downloads all of the related materials, and creates a folder related to the study in the internal archive for storage.

Notably, in most cases during this process, only the metadata is shared in the FAM, not the microdata. NSO's and other institutions which disseminate microdata frequently have strict policies which govern the terms of access. Accordingly, for data download, the FAM provides a link to the national platform. In this way, FAO can improve the findability of national datasets without assuming the risk of disseminating them. However, a risk to this approach occurs when a parent platform is not maintained. The user may find the metadata in the FAM, but the link to download the microdata is not functional. Nevertheless, in this case, the metadata should remain in the FAM because at least users will know that the datasets exists, and may be able to procure it using a different means (e.g. following directly with the host of the parent platform). At regular intervals, the data curator will check the links to external catalogues contained in FAM.

There are however few exceptions where access to the data is granted when harvesting metadata. Consequently, data curator will download this data for preservation in the internal archive (original folder), but will not disseminate through the FAM. Lastly, it is important to note that metadata is usually harvested in its original language and then translated to English for easy access/comprehension to a wide range of users. However, certain keywords/abbreviations will remain in the original language for easy findability of the study in the parent catalogue/source. Also, due to resource constraints, all data dictionaries will remain in the original language.

2.4 Metadata required for the FAM catalogue

OCS has created a DDI template containing the fields which will be used for the FAM. As in original standard described in 2.2.2, each DDI document has the 5 sections. However, the template is modified to include the fields most relevant for agriculture, nutrition, and food security studies. Descriptions and examples of the fields have also been modified in this regard. Finally, the minimum set of DDI elements have been defined as mandatory.

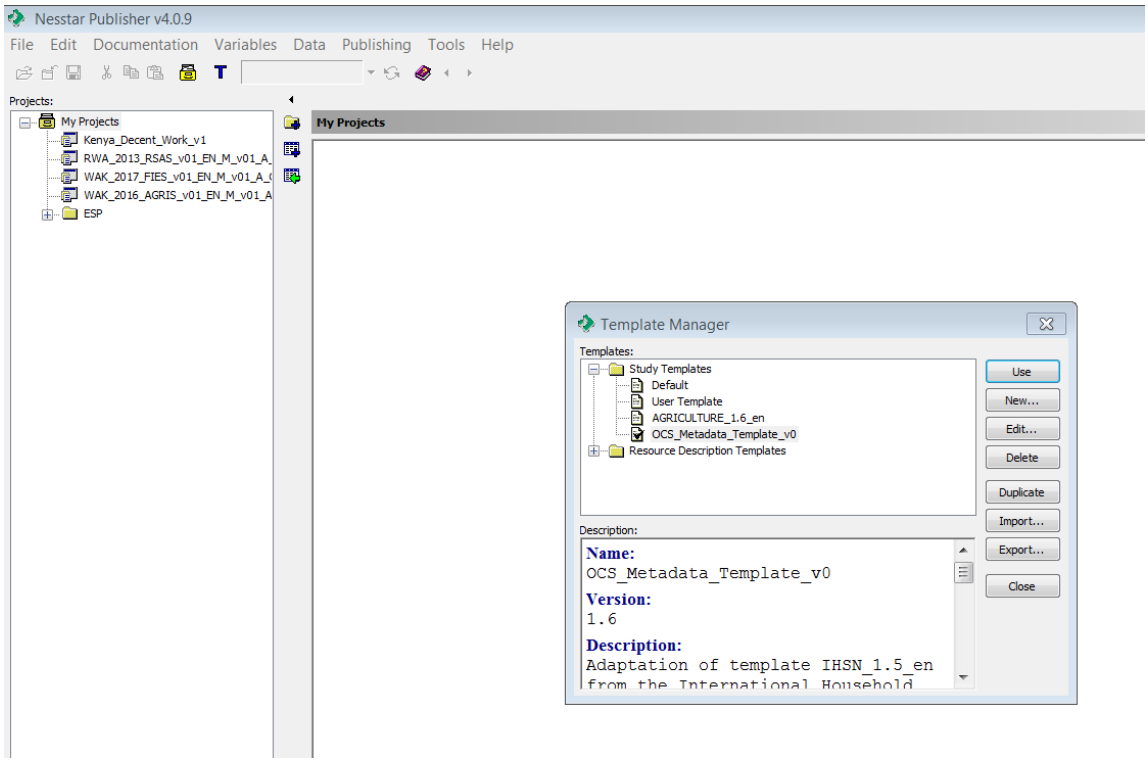
This section will give an overview of how to use the metadata publisher and DDI template using screen shots from the Bangladesh Food Insecurity Experience Scale (FIES) as an example. Screen shots showing how the template appears corresponding to some elements are displayed directly under their descriptions. Also, for each element included in the template, it is noted whether the element is optional or mandatory.

2.4.1 The metadata publisher

The purpose of the metadata publisher is to provide a user friendly interface to produce the DDI XML document. Currently, the Nesstar publisher is available at http://www.nesstar.com/software/download.html. The metadata publisher uses templates that define which DDI and DCMI elements are to be used to document datasets. Also, the template provides descriptions and examples of each DDI element to help users.

In order to open a specific template, click on "Documentation" in the top menu bar, and then click on "Templates". In our case, select "OCS_Metadata_Template_v0" as seen in the screen shot below. This template was created by OCS for the FAM catalogue. You have different options on the right:

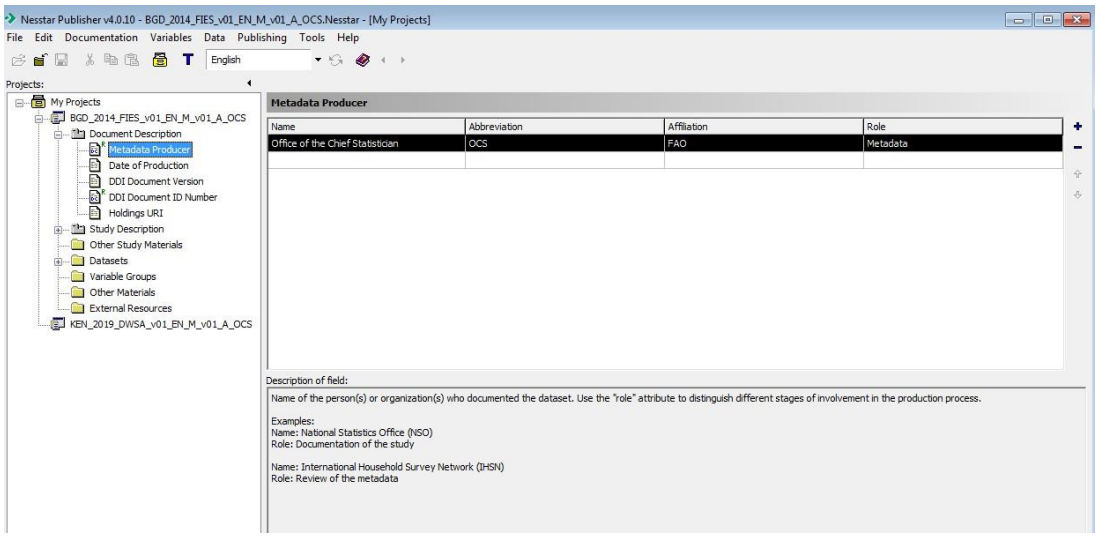
- “Use”: to use the template to fill DDI elements
- “New”: to create a template from scratch
- “Edit”: to open a template in edit mode and make changes
- “Delete” to remove the template from Nesstar
- “Duplicate”: to make an additional template similar to the original one
- “Import”: to open a template from a folder
- “Export”: to save a template from Nesstar to a folder
- “Close”: to close the Template Manager Box



2.4.2 Document description

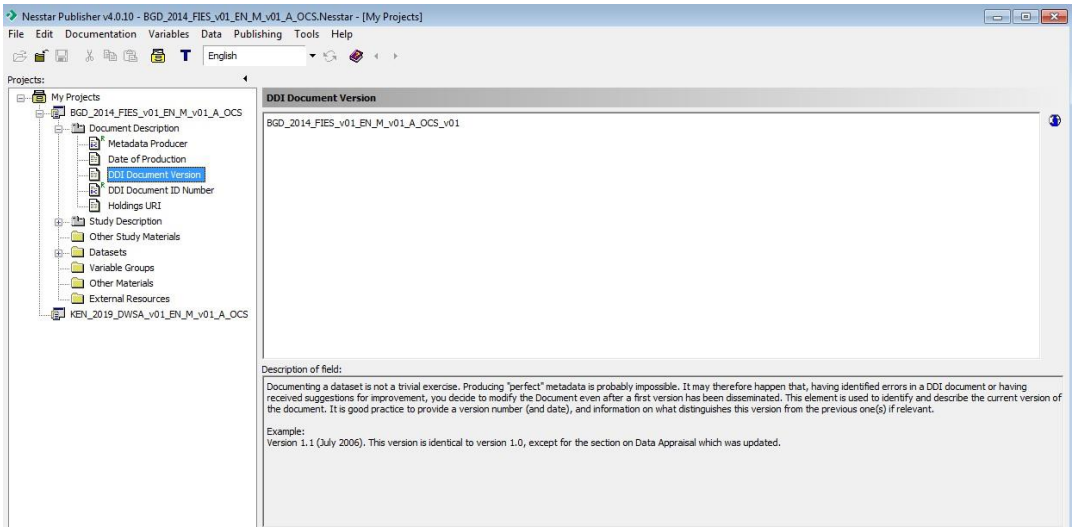
The first section, Document Description, includes information about the DDI XML document itself, such as:

Metadata producer (mandatory) captures information about who created the metadata. For harvested metadata, the creator will be the designator from the data provider; while OCS will adapt the metadata for FAM. However, OCS will be the producer if the metadata if created from scratch.

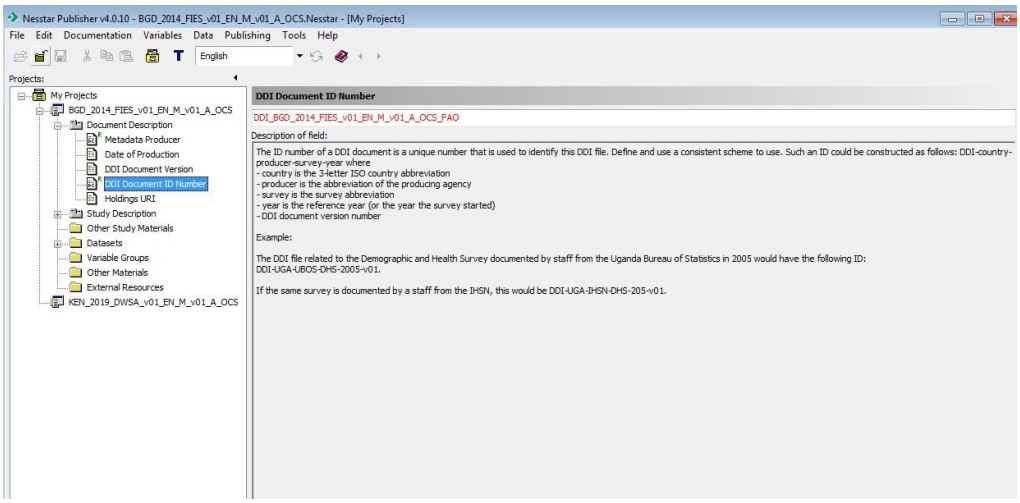


Date of production (mandatory) is automatically generated by the system to be the last day on which the file was saved.

DDI Document version (mandatory) identifies and describes the version of the metadata. In case some changes are made in the metadata, a new version should be created and this field should be updated. The DDI Document version naming convention is described in Annex 1.



DDI Document ID number (mandatory) follows the naming convention defined in Annex 1. In the example, the field is red because Nesstar does not recognize certain characters. However, it is not considered a mistake.

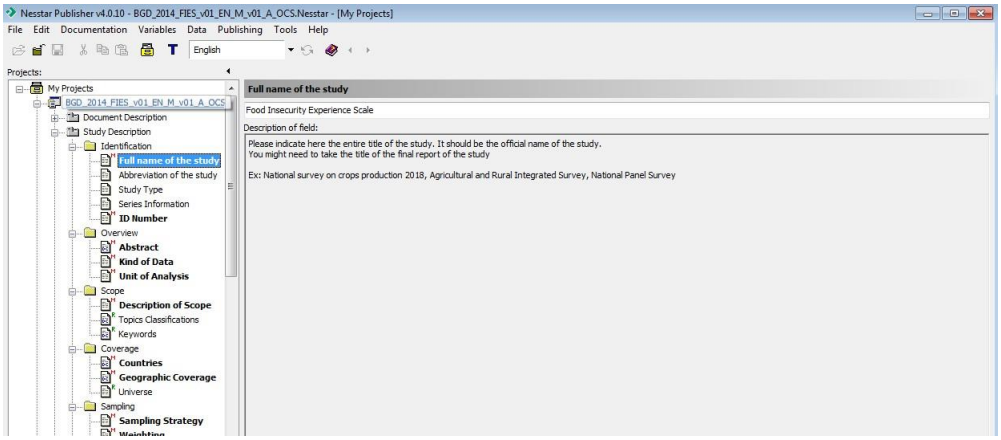


2.4.3 Study description

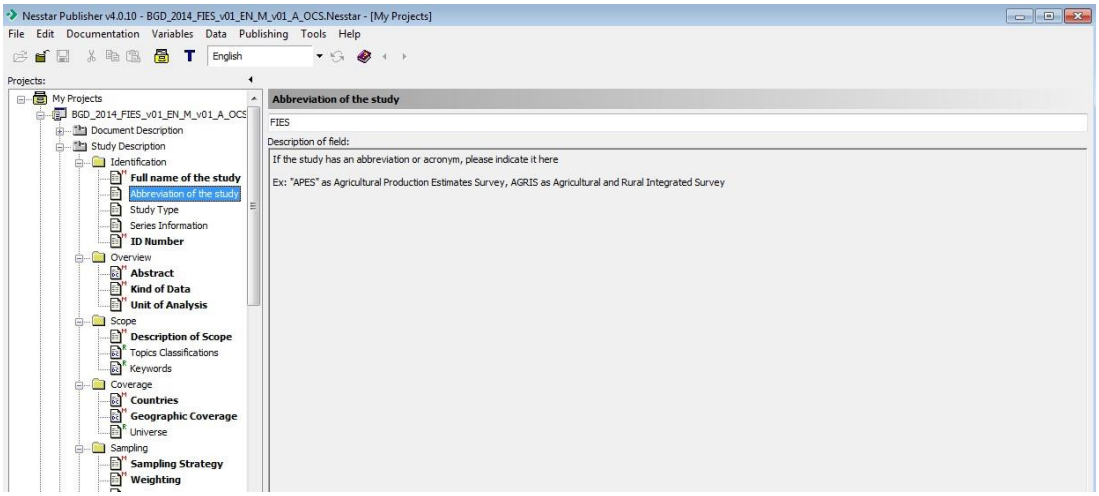
This section contains 12 sub-sections: Identification of the study, Overview, Scope, Coverage, Sampling, Data collection, Data processing, Data appraisal, Producers and Sponsors, Data access, Disclaimer and Copyright, Contacts which correspond to information about the study.

Identification

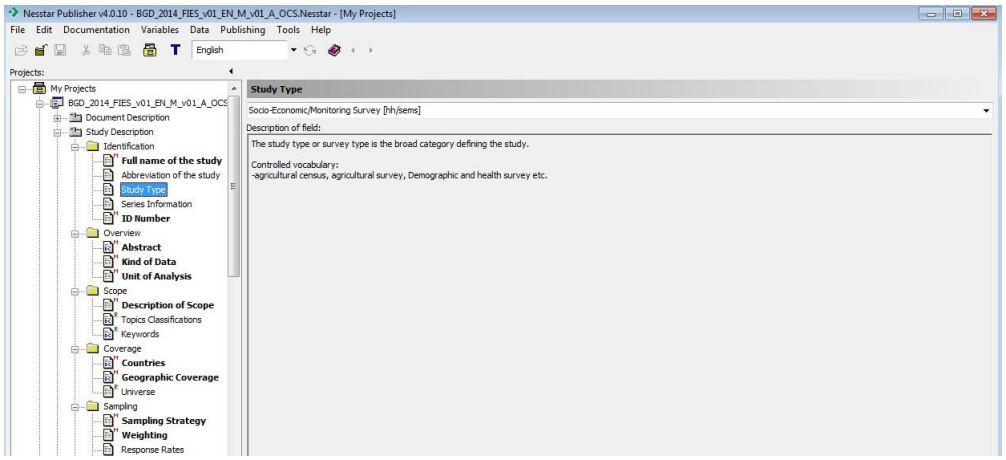
Full name of the study (Mandatory) is complete title of the study. It should be the official name, and the year the study was conducted.



Abbreviation of the study (optional) is most often an acronym, plus the year of the study for example Annual Agricultural Survey 2018 is abbreviated AAS, 2018.

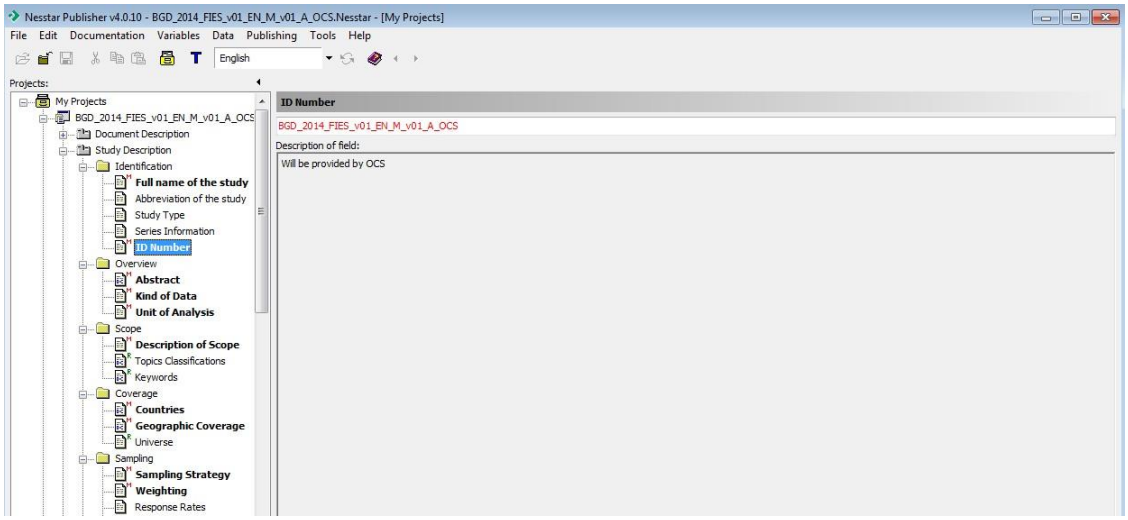


Study type (optional) is a broad category defining the study. There is a controlled vocabulary.



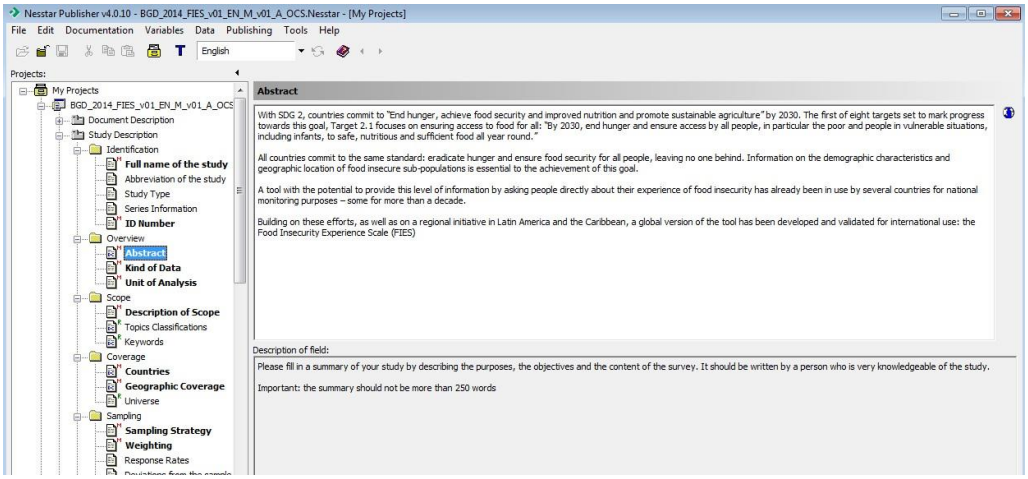
Series of information (optional) is a group of related studies in the past that have the same objectives and could be compared to the current one. A common example could be previous waves in a longitudinal study.

ID Number (mandatory) follows the naming convention rules and will be provided by OCS.

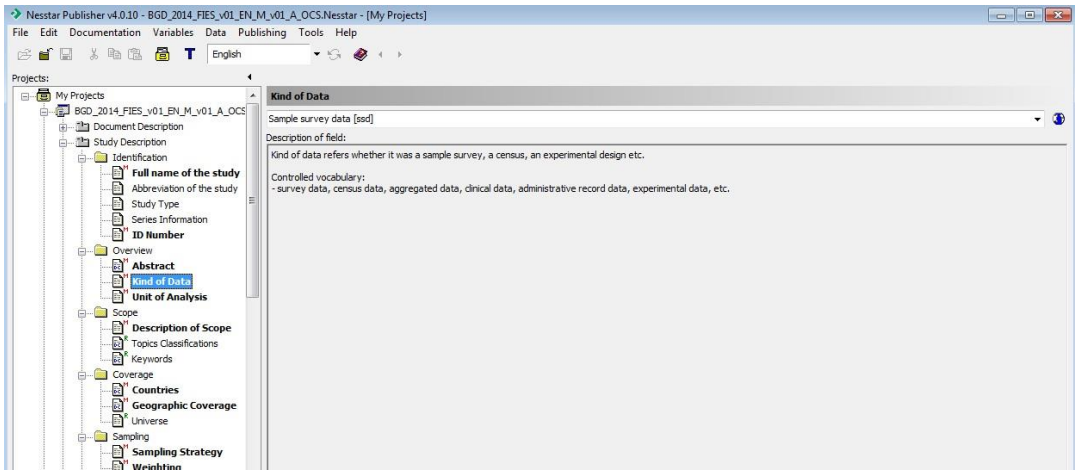


Overview

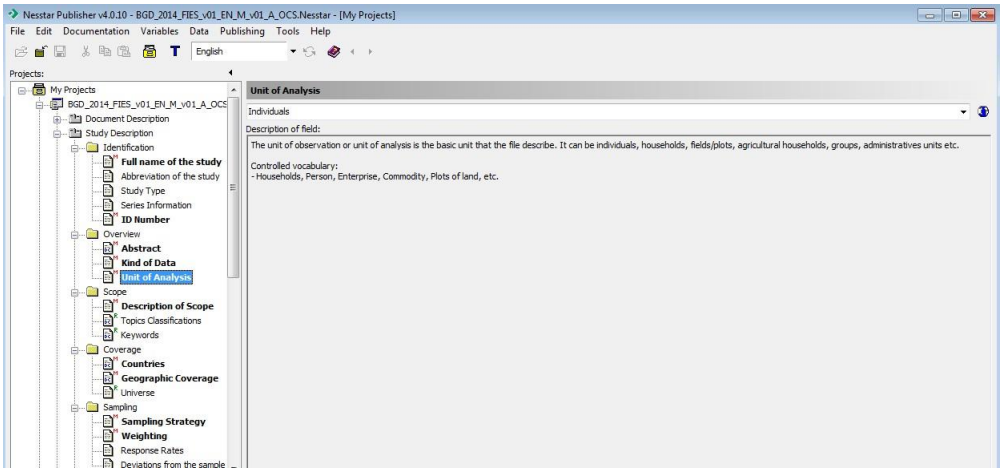
Abstract (mandatory) is a summary of the study including purpose, objectives, and content of the survey. It must be written by a person who is very knowledgeable of the study, and is akin to an abstract written for any type of research paper.



Kind of data (mandatory) specifies whether it was a sample survey, a census, an experimental design etc. There is a controlled vocabulary.

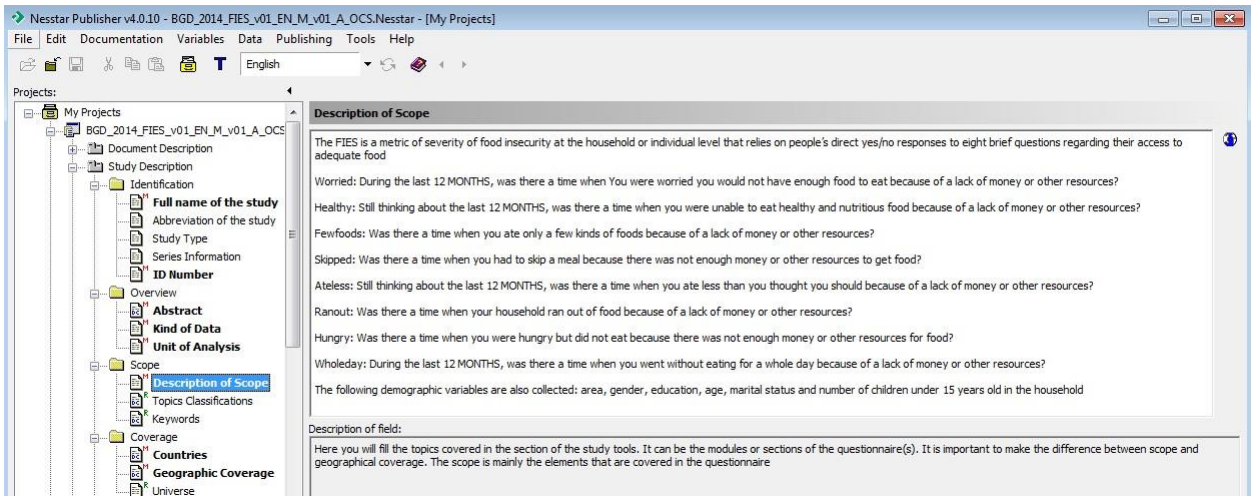


Unit of analysis (mandatory) is the basic unit that the data describe. It can be individuals, households, fields/plots, agricultural households, groups, administrative units etc. A controlled vocabulary is prepared.

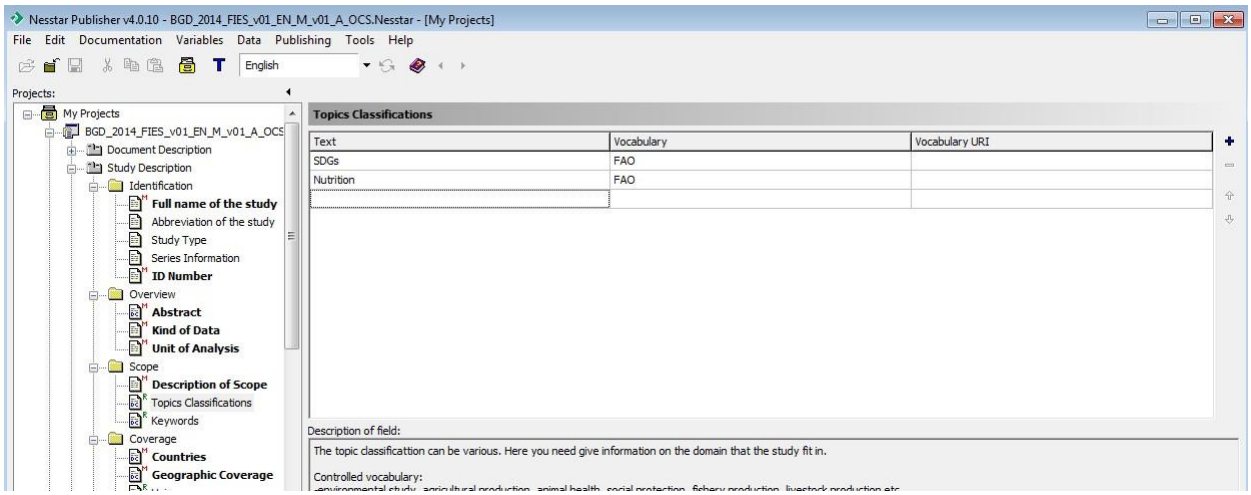


Scope

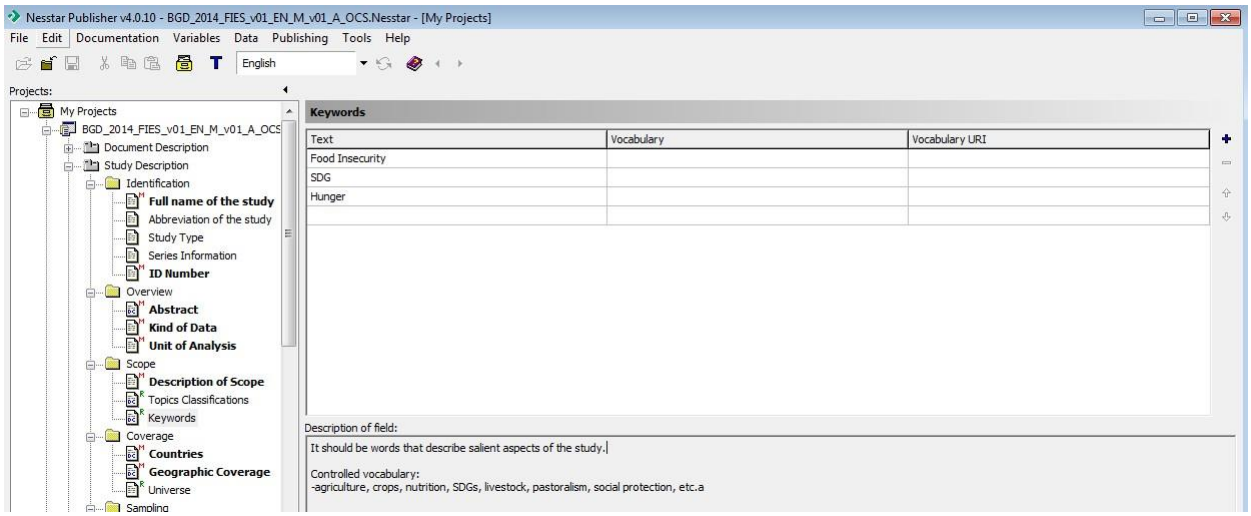
Description of the scope (mandatory) refers to the topics covered in the section of the study tools. It can be the modules or sections of the questionnaire(s). It is important to clarify the difference between scope and geographical coverage. The scope is mainly the topics that are covered.



Topics classification (optional) provides information on the domain that the study belongs to. There is a controlled vocabulary.

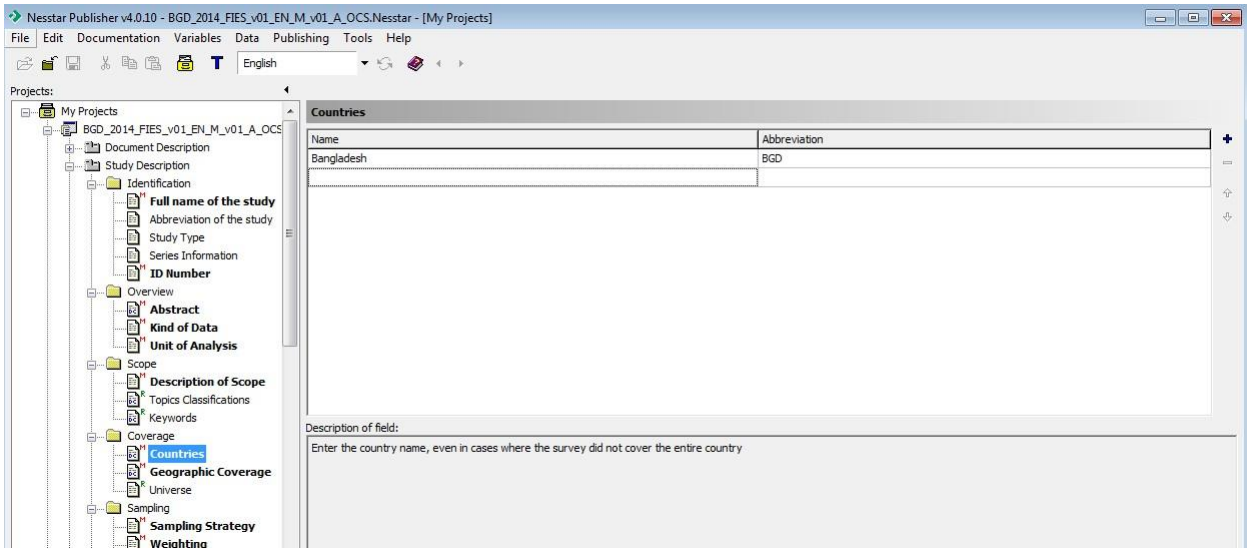


Keywords (optional) describe salient aspects of the study.

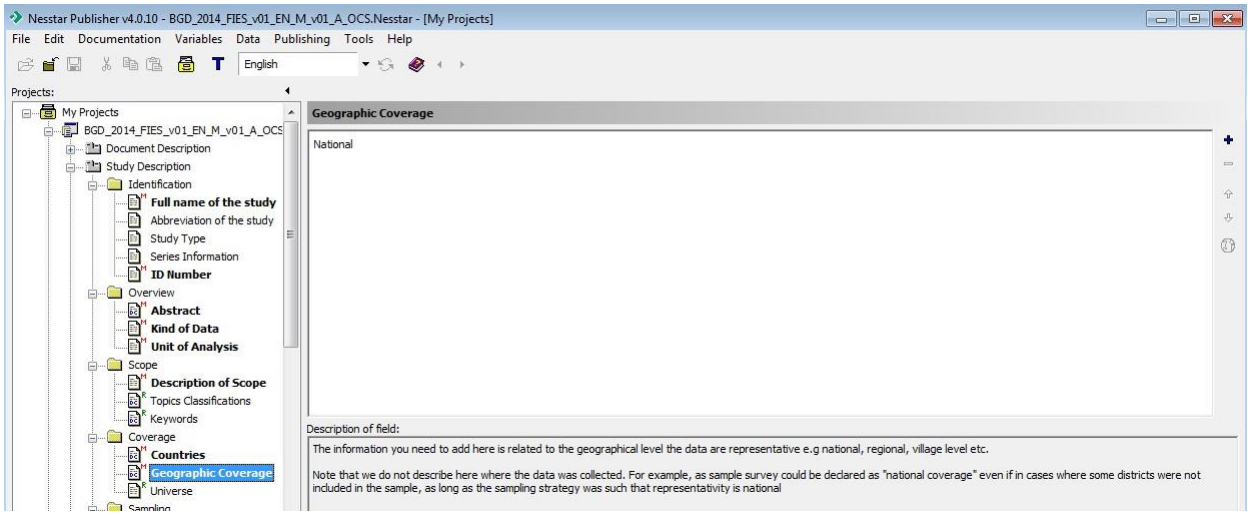


Coverage

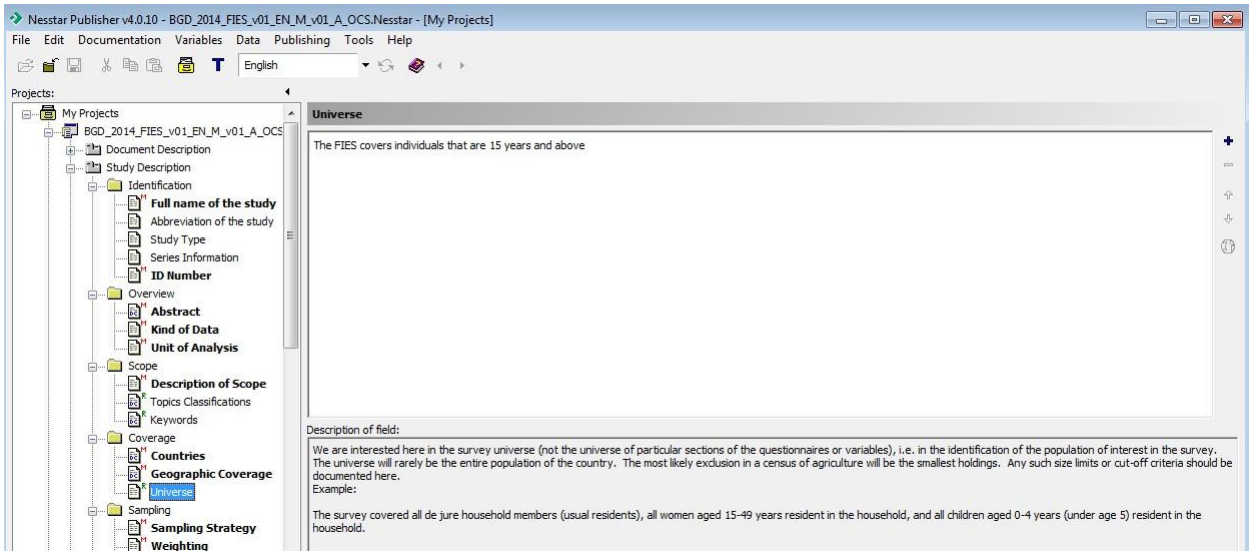
Countries (mandatory) is the list of all the countries that the study is referring to.



Geographic coverage (mandatory) is the geographic/administrative level for which the data provide representative estimates (e.g. national, regional, village etc.).



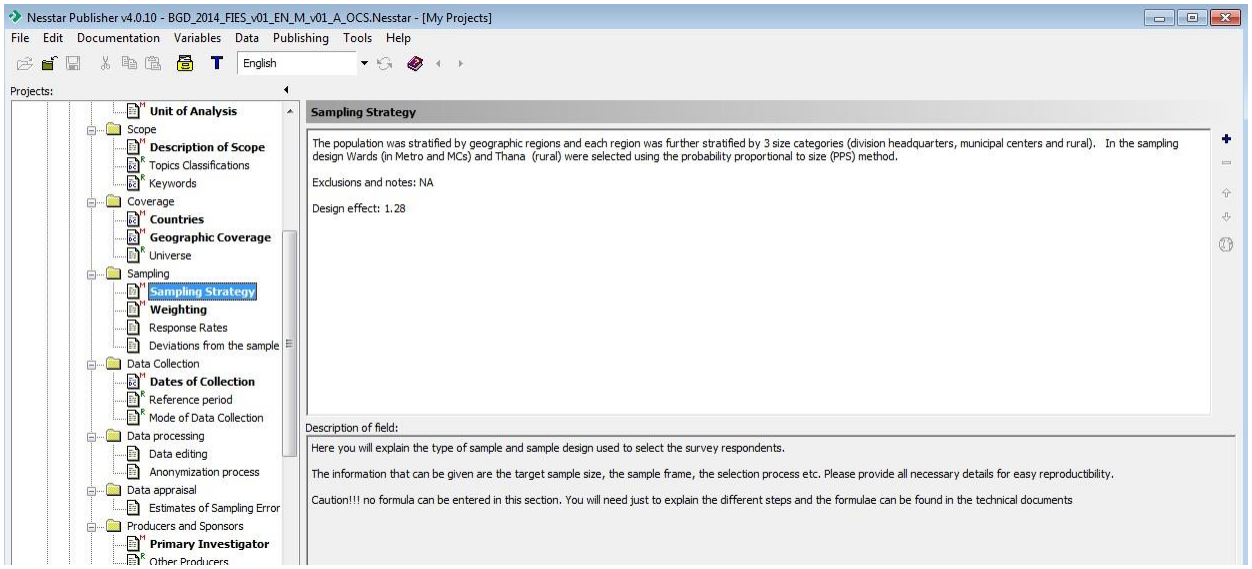
Universe (optional) is the population of interest in the study e.g. female 15 years and above, children under 5 years etc. In other words, it is the universe of the sample.



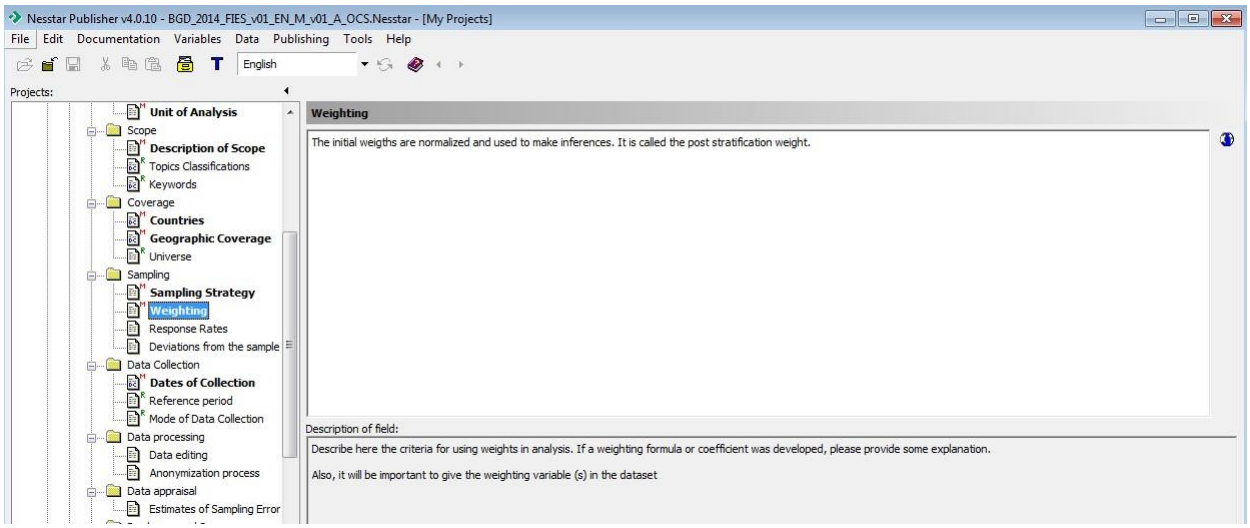
Sampling

In addition to providing the information below, if a detailed document pertaining to the sampling and estimation procedures is available, it should be uploaded with the study in the other materials.

Sampling strategy (mandatory) is open text which should contain the target sample size, frame and selection process. Due to the use of XML, no formulas can be entered in this section.



Weighting (mandatory) describes the criteria for using the weights in the analysis. It can also give explanation on the weighting formula or coefficient that was developed.

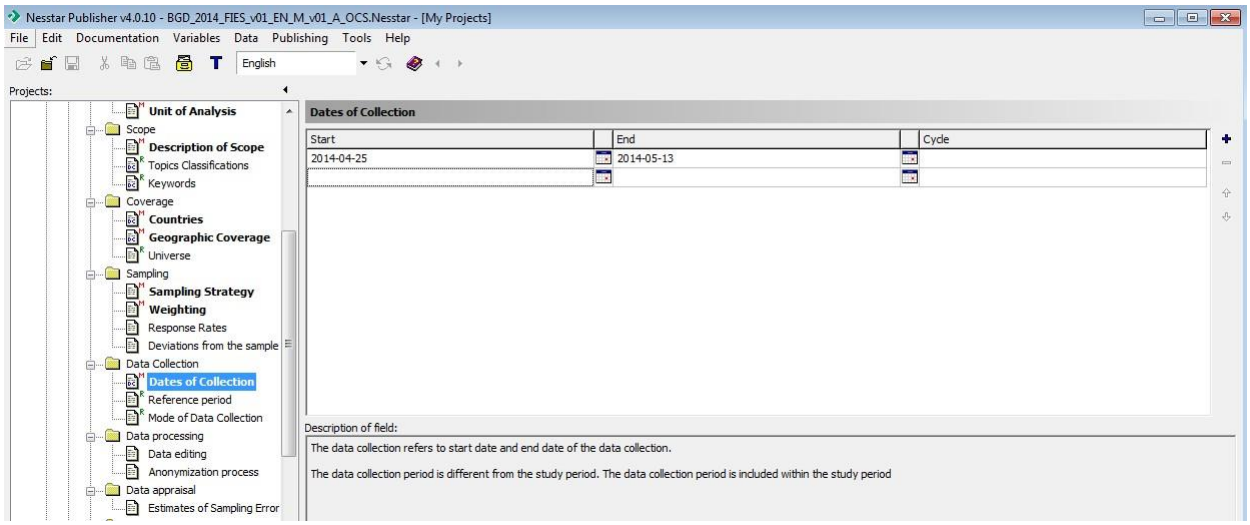


Response rates (optional) provides the percentage of sample units that participated in the study based on the original sample size.

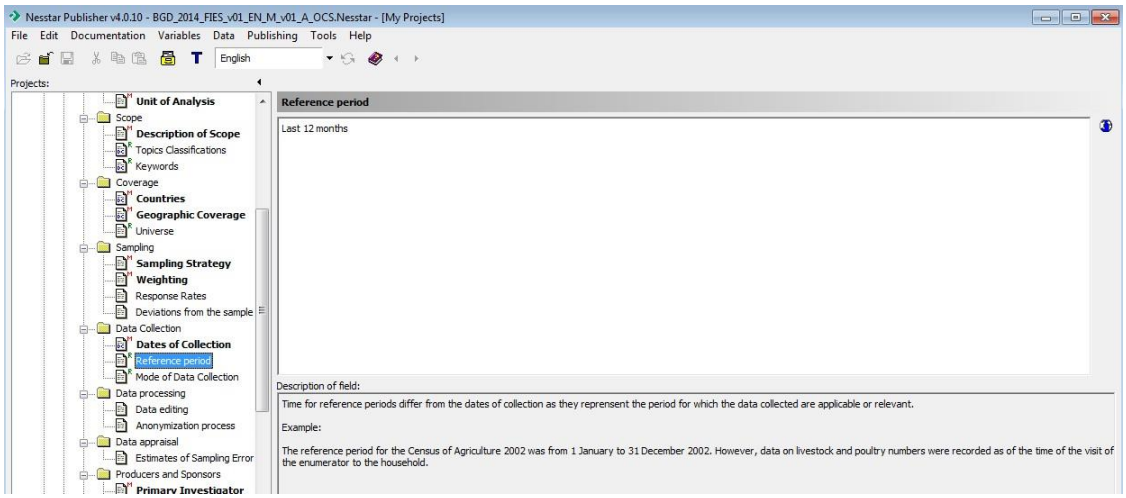
Deviations from the sample design (optional) explains any deviations from the planned sampling design.

Data collection

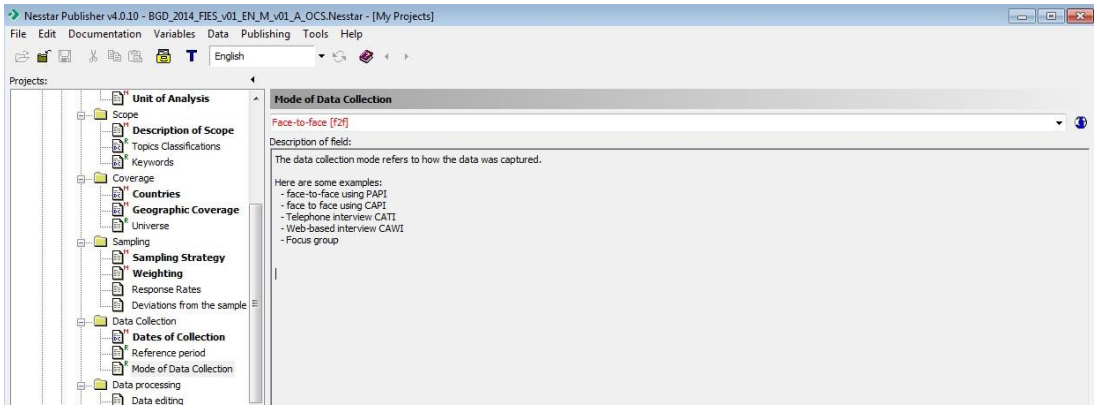
Dates of collection (mandatory) refers to the start date and end date of the data collection.



Reference period (optional) is the period for which the data collected are applicable and relevant. Reference periods within a study may vary according to the variable. In this case, only the longest reference period has to be mentioned and the sub-reference period should be mentioned in the variable description section.

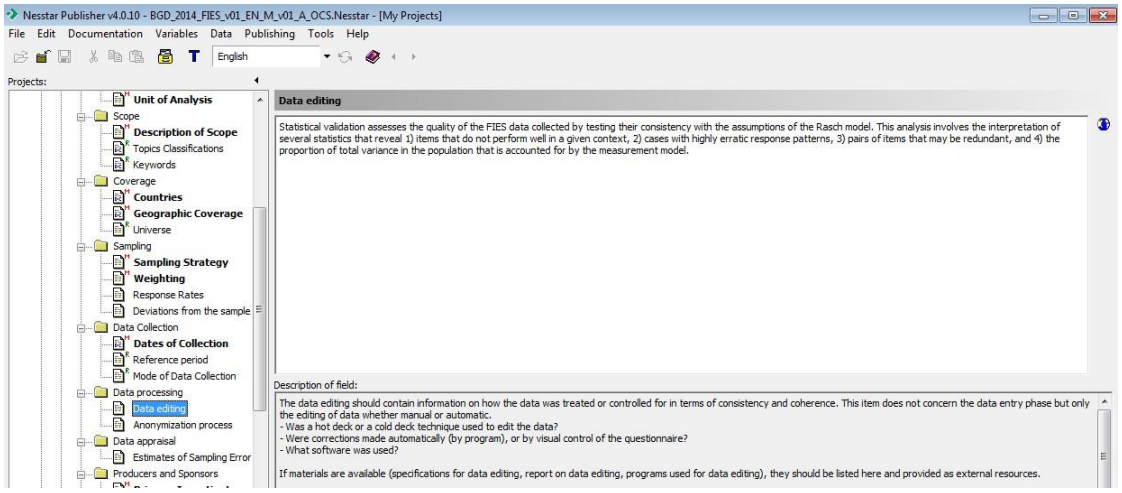


Mode of data collection (optional) refers to how the data were captured. There is a controlled vocabulary.



Data processing

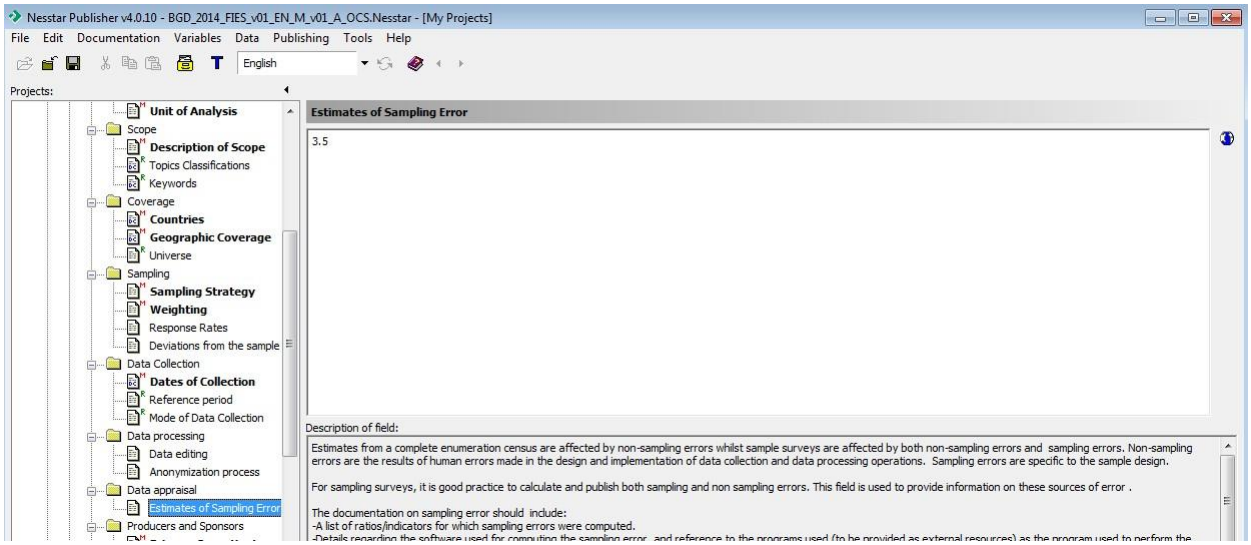
Data editing (optional) contains information on how the data were treated or controlled for, in terms of consistency and coherence. The data entry operation should not be mentioned here.



Anonymization process (optional) indicates the relevant information about the anonymization of the micro datasets. This information is described in the SDC protocol.

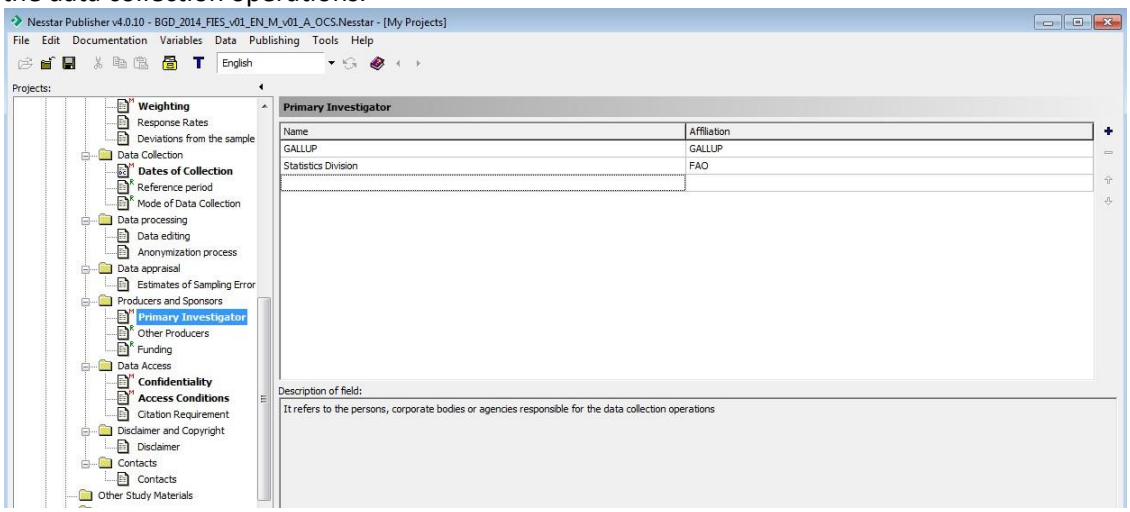
Data appraisal

Estimates of sampling errors (optional) is self-explanatory, and should not be confused with the non-sampling errors that are the results of human errors made in the design and implementation of data collection and data processing operations.



Producers and sponsors

Primary investigators (mandatory) refers to the persons, corporate bodies or agencies responsible for the data collection operations.



Other producers (optional) is the list of other interested parties and persons that have played a significant but not the leading technical role in conducting the study.

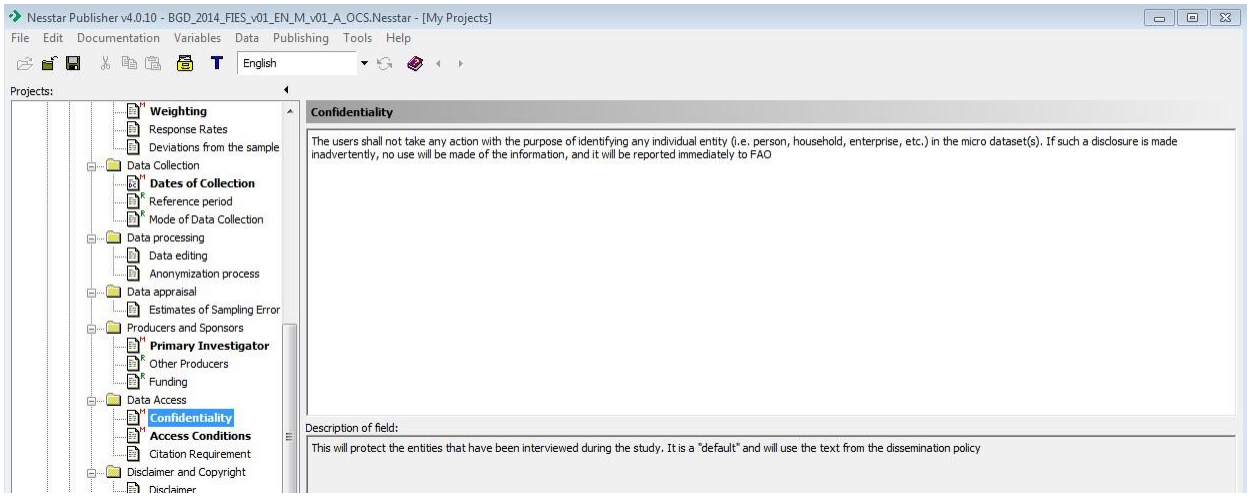
Funding (optional) contains the list of organizations (national or international) that have contributed, in cash or in kind, to the financing of the survey.

Data access

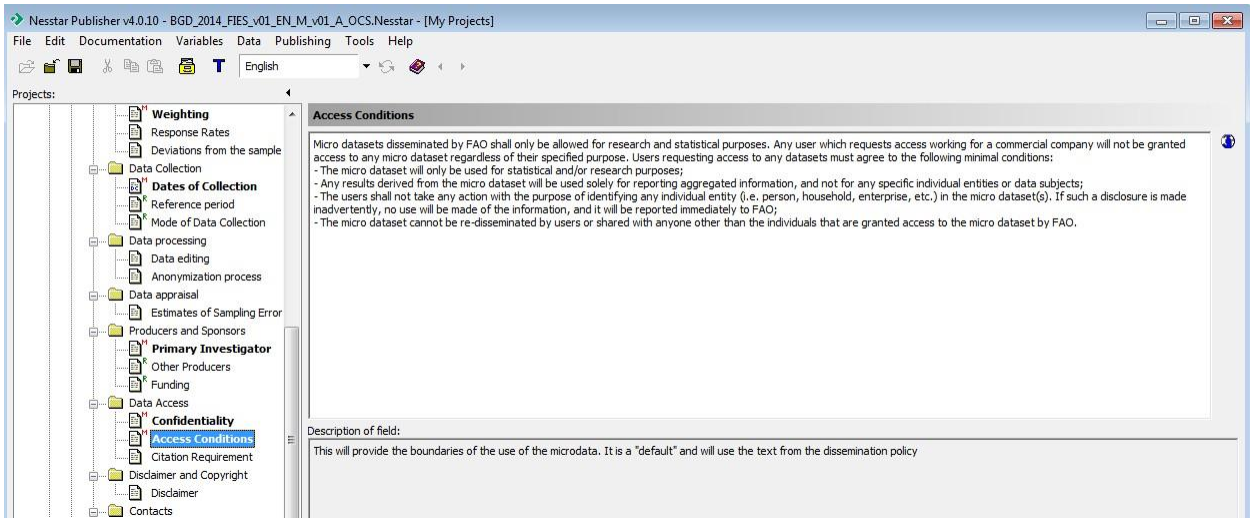
Confidentiality (mandatory) is a "default" in most cases and will use the defined text specified below.

"The users shall not take any action with the purpose of identifying any individual entity (i.e. person, household, enterprise, etc.) in the micro dataset(s). If such a disclosure is made inadvertently, no use will be made of the information, and it will be reported immediately to FAO."

However, confidentiality clauses will be different for harvested metadata and will be based on the conditions of the external contributor.



Access conditions (mandatory) provides the boundaries of the use of the microdata. It is a default in most cases and we will use the text developed by OCS. Access conditions only differ for harvested metadata and will be based on the condition of the external contributor.

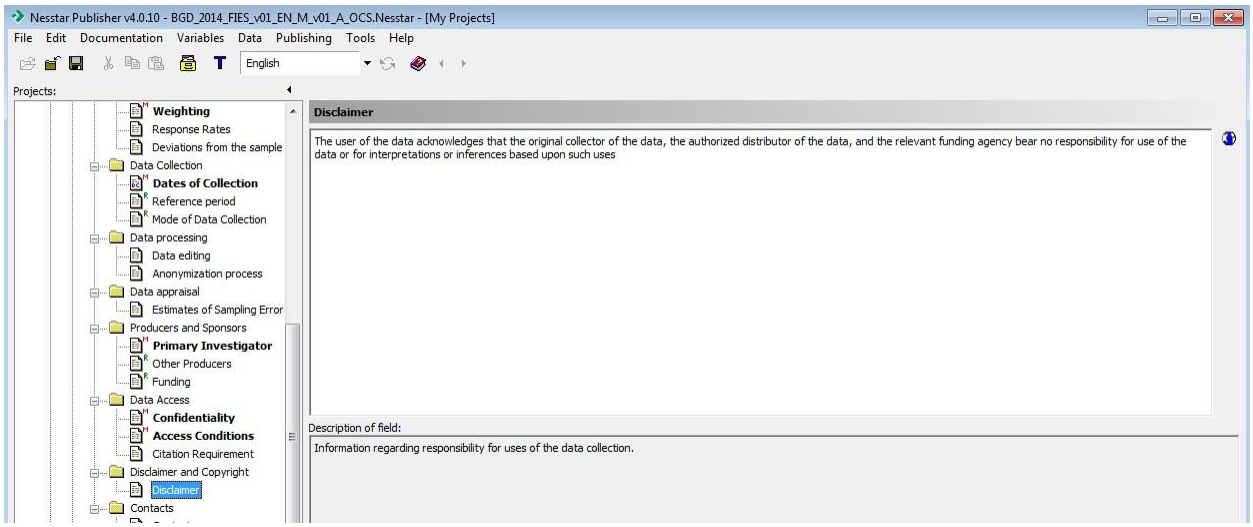


Citation requirement (optional) indicates how the dataset should be referenced when cited in any publication. It will guarantee that the data producer gets proper credit. The citation will also differ based on the source of the data (Internal or External). External datasets will follow the given citation of the external contributor while internal datasets will follow FAO guidelines. The advised format is:

Full last name, Capital letter of first letter of the first names of the authors. (Year). *Title of the article*. Journal. Version. Number of pages. Link if it exists.

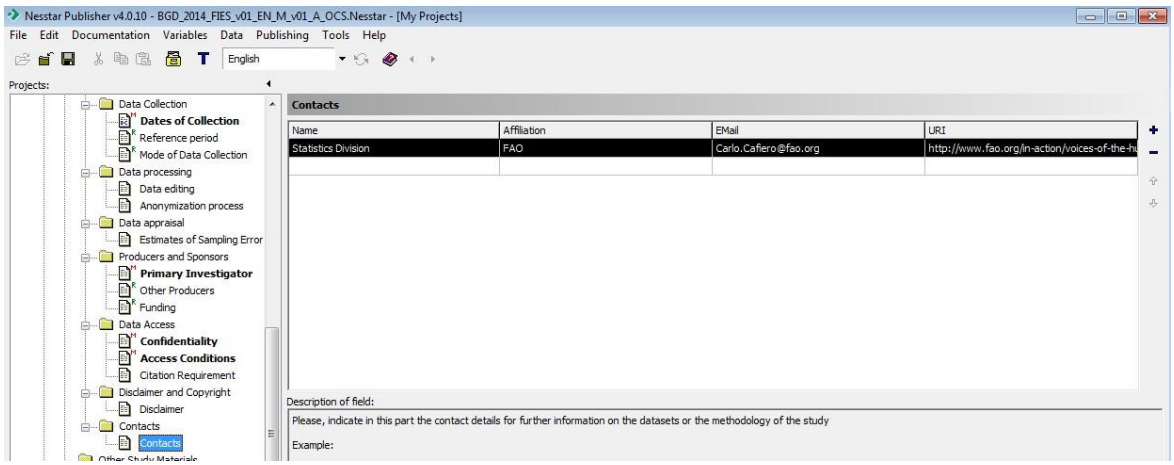
Disclaimer and copyright

Disclaimer (optional) contains information regarding responsibility for uses of the data. For internal data, it is going to be a default and we will use the text developed by OCS. However, disclaimers for external datasets will be based on the conditions given by the external contributor.



Contacts

Contacts (optional) provides the contact details for further information on the datasets or the methodology of the study.



2.4.4 Data file description

The data file description contains the description of the content of the dataset, producer, version, processing checks, missing data and notes. They are all optional.

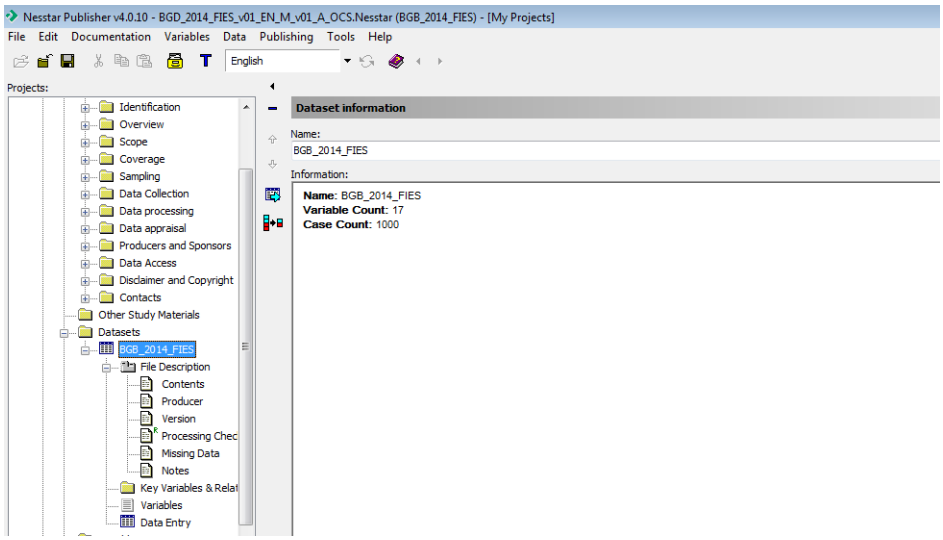
Contents describes the contents of the individual data file including sections of the questionnaire, etc.

Producer is the agency that produced the data file. If there is only one agency that produced all the data files, it might not be needed to fill this field and provide the information in the producers and sponsors sections of the DDI.

Version is useful if the data file undergoes various changes and modifications to keep track.

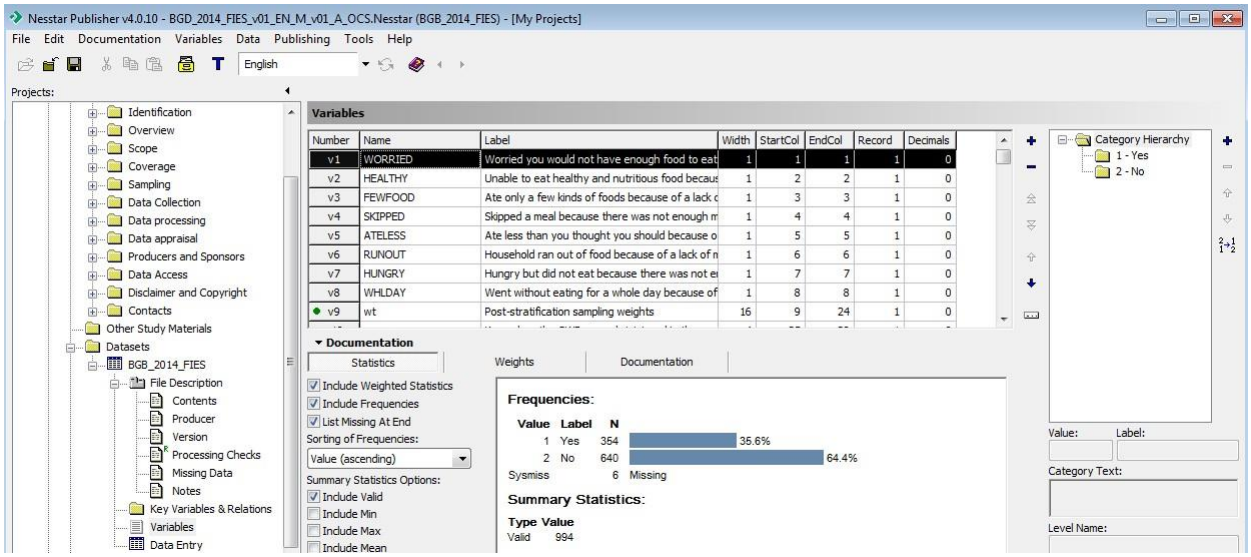
Processing checks provides information about the type of checks and operations that have been performed on the data file.

Missing data defines the codes that are assigned to missing records.

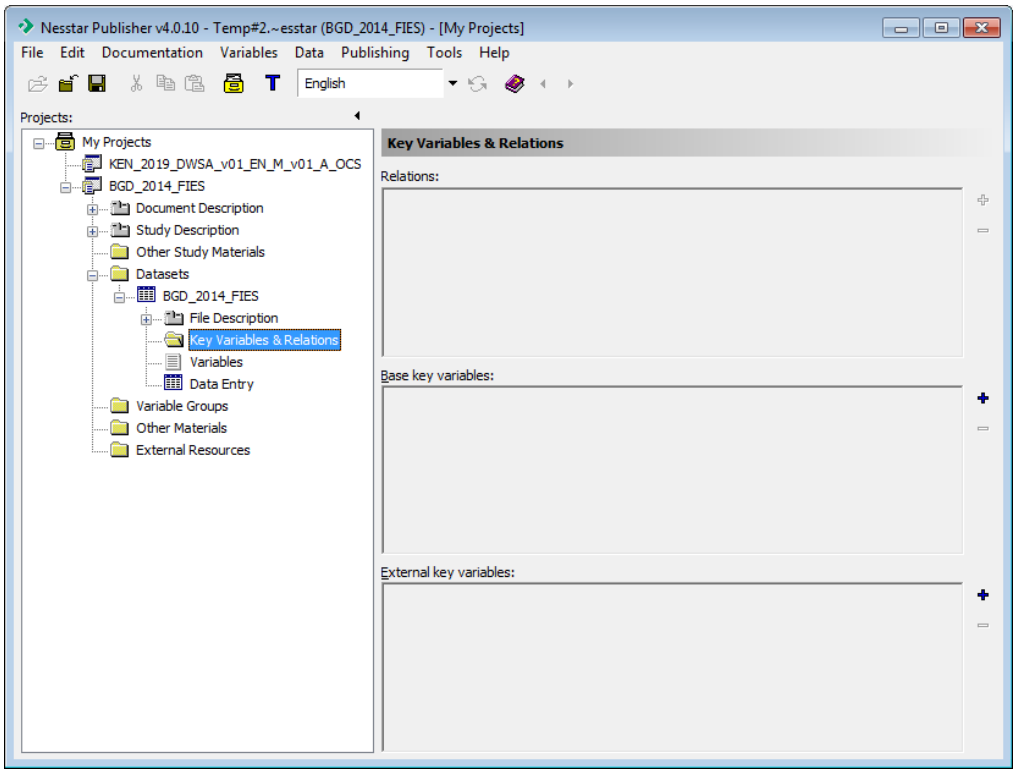


2.4.5 Variable description

Provides basic information for each variable such as definition, universe, label, literal questions as asked during the survey, etc. It has also options to edit/transform the variable, input labels and questions etc.



In case there are many datasets and are linked with variables called “keys”, the relations between those datasets should be provided by adding the keys to each dataset. To do so, the data curator should go to “Key Variables & Relations” under each datasets and provide the variables and other datasets related.



2.4.6 Related materials

Related materials are documents which help users interpret and analyze the microdata (e.g. questionnaires, enumerator's manual, technical reports, etc.). Studies which are linked, and related publications may be included. For each document, the following information must be provided: title, URI (if available), description (if necessary) and notes for additional information that need to be shared about the document.

2.4.7 Validation of the metadata in the metadata publisher

Once all the elements of the DDI and DCMI are filled, and the datasets verified and the links between datasets validated, the metadata can be validated by using the tools provided in Nesstar. This validation is mainly for the data curator to check if all the mandatory fields are filled and the links between datasets are set properly.

To do so, the data curator should go to "**Tools**" then click on "**Validate Metadata**". In case some mandatory fields are not filled, a warning will be given and the fields concerned will be highlighted in "red". However, the data curator might know the mandatory fields in the template are not necessarily required information for all types of studies. Some studies may not have these information; hence the data curator can proceed without filling the related fields as Nesstar Publisher allows it.

In a nutshell, the validation of the metadata in Nesstar is important in terms of verifying that all important (mandatory) elements are completed, but the data curator has flexibility in case some elements classified as mandatory are not available or important for a particular study.

2.4.8 Creating the XML and RDF file

Once the metadata has been validated, the DDI XML document and DCMI can be generated.

"Documentation" -> "Export" -> "Export DDI".

The DDI XML file and DCMI file should then be saved in the proper study folder in the internal archive.

2.4.9 Metadata quality assurance

Prior to publication of any metadata in the FAM, there must be complete and clear information at least for the following DDI elements:

- Full name of the study
- Abstract
- Kind of data
- Unit of analysis
- Description of the scope
- Countries
- Geographical coverage
- Sampling Strategy **No formulas allowed, if formulas needed, a related external document should be provided.*
- Weighting
- Dates of collection
- Primary investigators
- Confidentiality
- Access conditions
- Contacts

If information is available for other DDI elements, then as long as it is complete and clear, those additional DDI elements will be published in the FAM. Data providers are strongly encouraged to fill in as many of the additional optional fields as possible. The more information that is provided the less likely there will be additional clarifications from the data curator or eventual users. Data providers are required to submit the metadata in English.

The data curator will review the information submitted by data provider to be sure that it is clear, complete, and easy to understand. The curator may consult published reports or any publicly available material in order to determine the accuracy and adequacy of the metadata. Furthermore, the metadata text should not contain typographical, spelling, or grammatical errors.

3. Microdata Formats and Quality Assurance

This section will cover microdata standards including formats, and quality control which covers general data quality aspects as well as the application of statistical disclosure control. Finally, quality checks of supporting documentation will also be discussed.

3.1 Microdata Formats

The data deposit system will accept datasets in the following formats: .dta (STATA files), .sav (SPSS files) .rdata (R data files), and .csv (comma separated value)/Excel files for datasets with less than 50 variables. The dataset must be well-structured where each row in the data files should refer to a unique record, and variables/categorical values should be clearly labelled. As mentioned, in case of datasets with less than 50 variables, .csv, .xls and .xlsx are accepted as long as a separate sheet containing the complete codebook is provided.

If the dataset has a hierarchical structure, there must be clear keys, and identifying variables which facilitate easy understanding of the relationships and merging. The number of datasets should be reduced as much as possible, but maintain a natural division. For example, a farm survey may have a holding, employee, parcel, plot, and crop level datasets. The data provider should consider merging at least two of the datasets to reduce the number (i.e. probably parcel and plot, or holding and employee).

All datasets will be published in .dta, .sav, or .rdata format since they are the most common formats used in research. In collaboration with the data provider, the format for the microdata to be disseminated will be decided according to the utility for the users. In fact, disseminating microdata files in certain formats will jeopardize the goal of making them available and usable which is why large micro datasets will not be published in .csv or Excel.

3.2 Quality Assurance

OCS has a mandate to ensure quality of statistical outputs. In the case of microdata, quality also extends to protecting the confidentiality of data subjects. The following will define the quality checks that the data curator will perform, followed by the basic workflow for confidentiality protection using statistical disclosure control.

3.2.1 Data quality checks

The following quality controls are checked for every dataset by the data curator prior to dissemination. If any of the following criteria are not met, the data provider will have to resubmit the dataset until it reaches full compliance.

- The microdata files must not contain any variables which can directly identify a data subject (i.e. name, phone number, ID number, address, geo-reference of dwelling or farm, etc.). There are referred to as direct identifiers in the SDC Protocol.
- The data file cannot contain any extremely sensitive information which if disclosed, could cause harm to a data subject. Extremely sensitive information is more fully described in the SDC protocol.
- All the variables and values for categorical variables should be labelled. In case of a small dataset in a .csv, .xls or .xlsx format, a code book must be provided.
- Each dataset must contain a unique ID or a combination of variables that uniquely identifies every record.
- Missing values should be clearly coded, and labeled.
- In case of a sample survey, a weighting factor for every record should be provided. In case they are missing, explanation should be provided, and the appropriate uses of the dataset should be defined.
- Numerical variable ranges must meet realistic thresholds (e.g. age cannot be greater than 120, and area planted cannot be negative).
- The micro datasets should not contain any variables with all missing values.
- Relationships between hierarchical datasets should be clear, and contain unique identification variables for merger.

3.2.2 Confidentiality and Statistical Disclosure Control (SDC)

Confidentiality is one of the Fundamental Principles of Official Statistics, and included in Principle 10 of FAO's Statistical Quality Assurance Framework (SQAF). As a result, OCS developed legal tools (e.g. application for accessing licensed dataset and terms of use, license to redistribute datasets) and technical documents (i.e. SDC Protocol) to define strict rules on how the microdata should be anonymized, and shared to minimize the risk of any type of disclosure. The specific statistical procedures are described in detailed in the SDC protocol. The following are the five main steps with corresponding roles:

1. Removal of direct identifiers and extremely sensitive variables – Data provider
2. Definition of key variables, disclosure scenarios, preferred terms of access [9], and published statistics – Data provider
3. Measure risk and apply disclosure limitation methods – Data curator (OCS)
4. Evaluate protected dataset and document – Data Curator (OCS) and Data Provider
5. Approval by the Chief Statistician and release of the anonymized microdata file – OCS

Step 1 requires that the data provider remove any variables which can be used to directly identify a data subject. This includes names, phone numbers, identification numbers, etc. It requires that all extremely sensitive variables are removed which are defined as variables which if disclosed, would cause the data subject significant harm. Some examples may be certain health information, migratory status, political affiliation, etc., depending on the local context.

Step 2 requires that the data provider provide a list of key variables, disclosure scenarios, terms of access, and published statistics. Key variables are variables which may be used to link to external registers to disclose data subjects. Disclosure scenarios incorporate the key variables and describe the way that disclosures may be made. The data provider also can note whether they prefer licensed access, or public use, and finally provide any published statistics derived from the dataset. All of this information is needed for the data curator to apply SDC methods.

Step 3 and 4 are an iterative process wherein the data curator measures risk, applies SDC methods, and then evaluates the amount of protection, and resulting information loss. When the data curator believes that an appropriate equilibrium has been reached, then the data provider will be consulted for validation. The final approve is provided by the Chief Statistician.
Please refer to the SDC protocol for a detailed explanation of each step.

3.2.3 Adequacy of the supporting documents and the microdata files

Data providers are encouraged to submit supporting documents such as methodological reports, enumerator manuals, or any other information which may help user. If the following criteria are not held, the data provider will be asked to correct and resubmit:

- ☐ The questionnaires provided correspond to the dataset, and any labels, codes, etc. are consistent.
- ☐ The language used in the data files is the same as the questionnaire.
- ☐ All the other supporting documents (technical reports, etc.) pertain to the dataset.

[9] For the foreseeable future, all datasets will be distributed in the same way requiring users to submit an application prior to access.

4. User access

All information defined in the DDI are publically accessible without registration. It is also exposed for machines to harvest in XML, and JSON format through an open Application Programming Interface (API). However, if a user wants to download a micro dataset, he/she must complete an application for access to licensed dataset described below and attached as Annex 3. The applications are evaluated one-by-one by data curators in OCS. For external datasets, users will be taken to the parent platform's site, and required to follow the access terms of the parent platform.

4.1 Application to access licensed dataset

The Application for Access to a Licensed Dataset will capture the following information about the "Lead Researcher" who is defined as the specific user requesting access:

- First Name
- Last Name
- Organization
- Email
- Dataset requested
- Receiving organization
- Telephone number
- Intended use
- Expected outputs
- Expected completion date
- Other research team members which will have access
- Whole dataset, or specific subset of variables required

Then, the user must agree to the following terms of use:

- The Lead Researcher's organization and other researchers who will be involved in using the data at that organization must be identified. The Lead Researcher certifies that he/she is authorized to sign on behalf of the Lead Researcher's organization. If not, a suitable representative must be identified. Any violation of the Terms of Use of this agreement will be considered to have occurred on behalf of the Lead Researcher's organization, and FAO will take appropriate measures to sanction such misconduct, which may include denying any further and future access by the Lead Researcher organization or any other researchers involved in using the data ;
- The micro dataset will only be used for the stated statistical and/or research purpose. It shall not in any way be used for other purposes, including any administrative, commercial or law enforcement purposes.
- Any results derived from the micro dataset will be used solely for reporting aggregated information, and not for any specific individual entities or data subjects;
- The Lead Researcher nor anyone else authorized in this agreement shall take any action with the purpose of identifying any individual entity (i.e. person, household, enterprise, etc.) in the micro dataset(s). If such a disclosure is made inadvertently, no use will be made of the information, and it will be reported immediately to FAO;
- The micro dataset and other materials provided by the Food and Agriculture

Microdata Catalogue will not be re-disseminated, or sold or otherwise shared with anyone other than the Lead Researcher and anyone else authorized in this agreement without the written agreement of FAO.

- No attempt will be made to produce links or matching among datasets provided by FAO, and any other datasets that could identify individuals or organizations.
- Any books, articles, conference papers, theses, dissertations, reports, or other publications that employ data obtained from FAO will cite the source of data in accordance with the Citation Requirement provided with each dataset.
- An electronic copy of all reports and publications based on the requested data will be sent to FAO.
- The Lead Researcher shall implement security measures to prevent unauthorized access to this micro dataset. The micro dataset must be destroyed upon the completion of the research, unless FAO obtains satisfactory guarantee that the micro dataset(s) can be secured and provides written authorization to the Lead Researcher in this respect.
- FAO may monitor, at its discretion, use of datasets obtained from the Food and Agricultural Microdata Catalogue, and decide whether an abuse has taken place under these Terms of Use. In such case, FAO may sanction users for violations. Penalties may include restrictions or denial of further access to the Food and Agriculture Microdata Catalogue.
- The designations employed in the dataset do not imply the expression of any opinion whatsoever on the part of FAO concerning the legal or development status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Information contained in the dataset is provided on an "as is" and "as available" basis. No guarantee is given that the information is correct, complete or up-to-date. FAO does not represent or endorse the accuracy, completeness, authenticity or reliability of any information contained in the dataset. FAO shall not be held liable for any loss or damage arising from, or directly or indirectly connected to, the use of, reference to, or reliance on any dataset, including, but not limited to, any liability arising from any interpretation or inferences based upon the data, nor from any intentional or negligent misuse, errors, disclosure, undue transfer, loss or destruction of data that may occur
- The Lead Researcher will ensure that the Terms of Use are shared with any individual authorized to download datasets.
- FAO reserves the right to amend these Terms of Use at its own discretion. Any amendment affecting the conditions agreed upon under these Terms of Use will be notified to the Lead Researcher;
- Nothing contained in or related to these Terms of Use shall constitute or be interpreted as a waiver, express or implied, of the privileges and immunities of FAO.

4.2 Application review process

The completed Application for Access to Licensed Dataset will be sent automatically by the user to OCS. The following criteria will be used to determine whether or not access is granted:

3. Is the requester qualified and/or employed by an institution with a reputation in the proposed domain of study?

4. Is the dataset fit-for-purpose for the research or statistical project proposed?

If OCS is not able to determine the answers to those two questions, OCS will consult the data user.

If a request is rejected, the requester will be advised by email, and a justification will be provided. In case the request is approved, an email will be sent to the requester providing a link to download the micro datasets.

5. Metadata and microdata workflow

This section integrates the steps described in the previous sections into an integrated workflow. The reader should note that this workflow is subject to change as more experience is gained.

The first subsection describes the workflow of curation and dissemination when microdata and metadata are disseminated through FAM. The second subsection provides the workflow for studies wherein only the metadata is published, and the microdata is disseminated through an external contributing catalogue or statistical website.

5.1 Dissemination of both microdata and metadata

Microdata disseminated directly through FAM is likely to most frequently come from technical officers inside FAO. However, it may be the case that sometimes other institutions may request a dataset of theirs to be disseminated in FAM. The overall standards and processing is the same, but modality of curation is different.

5.1.1 Microdata and Metadata Curation

As the OCS does not have a comprehensive list of micro datasets, the responsibility of identifying micro datasets collected by FAO is the responsibility of technical units. In this regard, FAO staff members will have access to a data deposit system when they login into FAM using their corporate credentials. The deposit system allows FAO staff to create a request to publish a dataset otherwise referred to as a "study". This request requires submission of all the metadata and relevant documentation defined in the data deposit interface.

Only FAO staff members will have access to the data deposit system. Accordingly, any other institution which wishes to disseminate microdata through FAM will have to send a request to fam-catalogue@fao.org with a general description of the dataset(s) and request to disseminate through FAM. Then if approved, the OCS will send a legal template for the data provider to sign granting FAO the right to disseminate the dataset though FAM. Next, the data provider must complete an Excel based template capturing all the metadata, and provide the datasets as well as related materials.

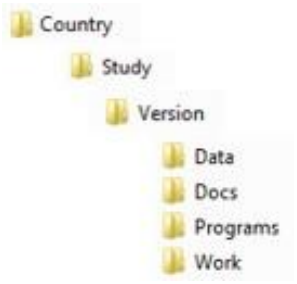
5.1.2 Internal Archiving

When a dataset is received, a folder will be created for it in the internal archive. The naming of the folder, relevant documents, and organization are described in detail with examples in Annex 1. A brief overview of the structure is described below:

1) Country folder: Full country name as in the M49 list
2) Study (subfolder of Country folder): study/survey name. The name of this folder should reflect, the country (ISO 3 alpha code), the year of the survey and the acronym of the survey.

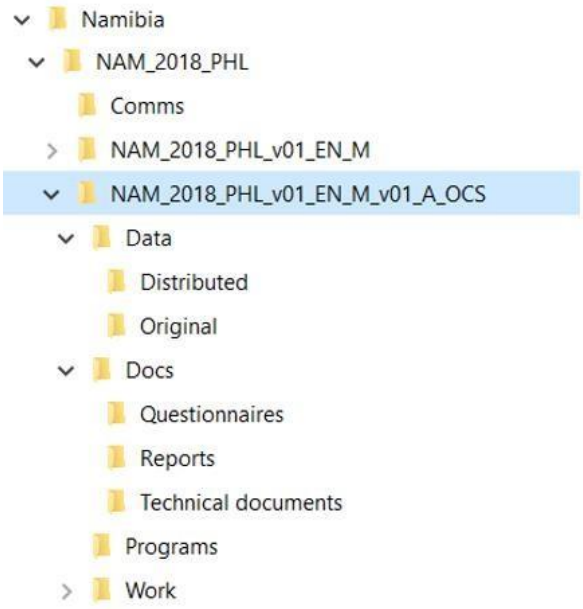
3) Version (subfolder of Study folder): version of the study that should follow the naming convention described in Annex 1. The DDI XML metadata document is stored in this folder.
4) Various sub-folders: The study folders comprises of a “Data” folder, a “Docs” folder, a “Programs” folder and a “Work” folder. The content of each folder is provided in the naming convention document.

Below is a generic example of a country folder for an internal archive.



The next diagram shows an actual folder structure in the internal archive for Namibia with a study folder expanded. In the diagram, the country is Namibia, the study is NAM_2018_PHL because it is a post-harvest loss study that took place in 2018 in Namibia. There is a "Comms" folder, in some cases, to archive conversations with the partner (National Statistical Agency of Namibia) showing their review and approval for publishing the study. Then, the folder NAM_2018_PHL_v01_EN_M_v01_A_OCS is the disseminated study with sub folders containing material indicated by the titles of the folders. The naming scheme of the folder is described in detail in Annex 1.

All of the work and materials performed by the data curator related to this study should be done in this internal archive. The idea is that at any point in the future, someone should be able to retrace the actions of the data curator, and access all the materials which were developed, or procured during the archiving process.



5.1.3 Quality assurance

After creating the folder in the internal archive, the next step is to perform quality assurance by reviewing the metadata elements, microdata, and related materials.

The metadata is imported into the Nesstar publisher if it has been received through the data deposit system, otherwise, it is copy and pasted from the Excel template into Nesstar. The metadata quality is checked by the data curator according to the standards described in 2.4.9.

Then the data curator uses an Rmarkdown template which integrates microdata quality, as well as the application of SDC methods. There is a standard code in the template, and plenty of space to annotate

each step for the data curator to explain his/her logic. This renders the quality assurance processes transparent and re-producible. This Rmarkdown document is stored in the "Work" subfolder, and when finalized generated and saved as .pdf.

Finally, the related materials will be reviewed and validated.

5.1.4 Validation by data provider

After completing the previous steps, OCS will provide the metadata template, and anonymized micro dataset to the data provider for validation. The Rmarkdown file can also be provided if requested. The data provider will be asked to validate the following:

- ☐ Accuracy of DDI XML document (i.e. metadata)
- ☐ Anonymization (i.e. no highly valuable analytical information was removed)

5.1.5 Approval and publication

After the validation, the data curator will ask the Chief Statistician for final approval to publish in the FAM catalogue. Once approved, the study will be published in the FAM catalogue, and an email will be sent notifying the data provider with the URL in the FAM.

5.1.6 Review access request for licensed datasets

For micro datasets disseminated directly by the FAM, data users can request access by completing an Application for Access to a Licensed Dataset attached as Annex 2. The approval process is also described in Section 4.

5.1.7 User support, evaluation, and feedback

Data users may need to have some clarifications about the micro datasets. An email address will be created to receive these types of request: fam-catalogue@fao.org. Once the request is received, the OCS team will judge the need to share it with the data provider. Indeed, some queries can be processed directly by OCS team. In case they are complex and need the expertise of the data provider, a delay of 5 days will be given to the data provider to answer the request. The OCS team will coordinate the entire process and send the answers to the user.

Every so often, the OCS team will initiate an evaluation process on the FAM catalog in accordance with Statistical Standard Series 12: User Consultations (UC). The UCs will include both data providers and users and feedback will be collected from both parties.

In addition to UCs, FAM users will always be invited to give feedback at user support email address. Feedback will also be generated in form of reports for data providers to provide insights on the use of their data in FAM.

5.1.8 Summary of workflow

The following are the main steps in the typical workflow:

Step 1: Request to disseminate microdata by a data provider. We identify two categories of data provider: internal (FAO employees) and external (non-FAO employees).

Step 2: Create workspace and folder structure in archive to store all documentation.

Step 3: Review of the metadata, microdata and documents, application of SDC by data curator. If quality checks are not passed, then the data provider will be notified and requested to address any issues and re-deposit.

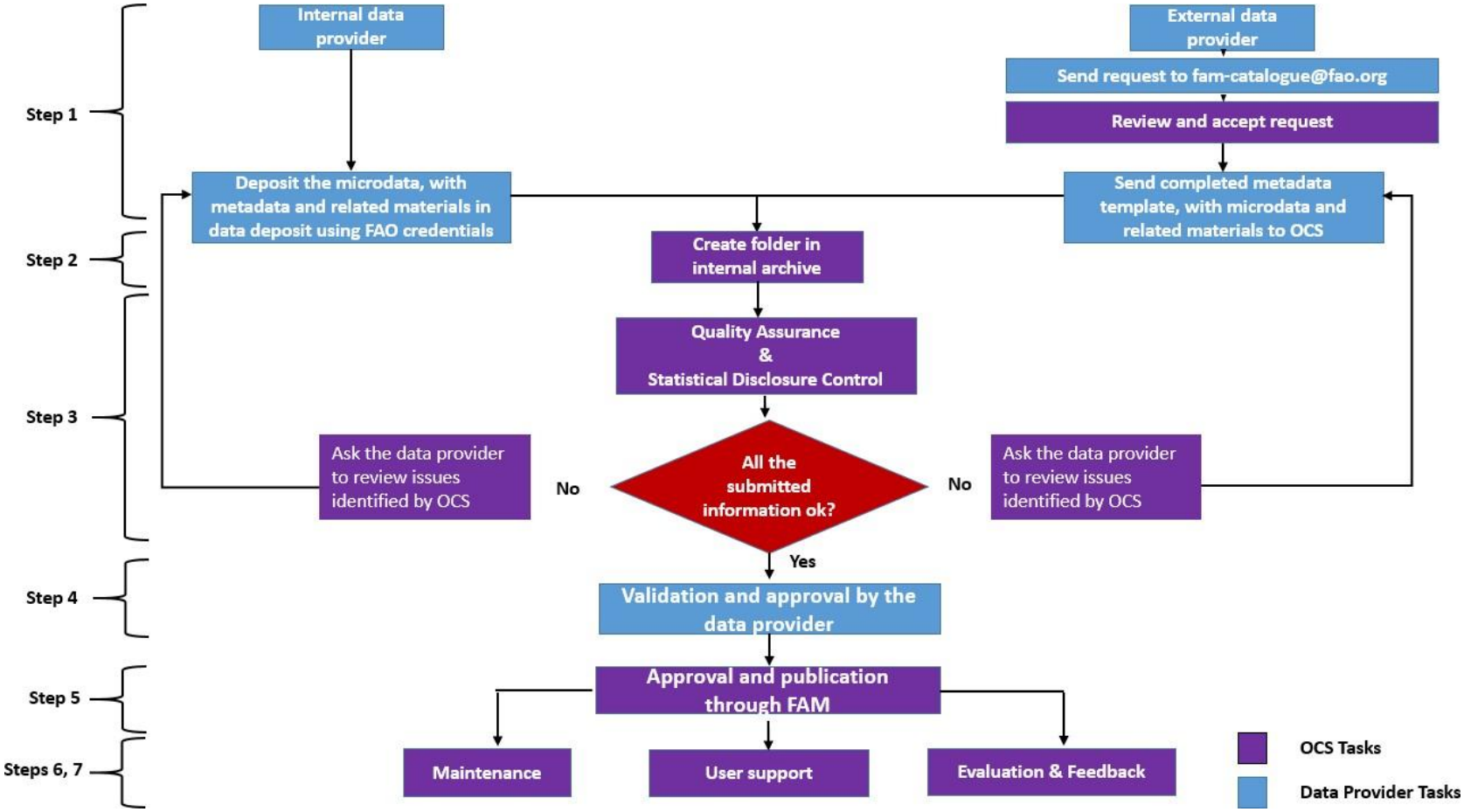
Step 4: Final validation and approval by the data provider.

Step 5: Final approval of Chief Statistician and uploading in the FAM catalogue.

Step 6: Review data access requests.

Step 7: Provide support to users, evaluate user satisfaction, and receive feedback on contents.

Figure 3: Workflow of the dissemination process for microdata received from the data deposit



5.2 Metadata life cycle for external contributing sources/catalogues

As stated in section 2, metadata could be harvested from other platforms that have food and agriculture related metadata and/or microdata disseminated. In all the approaches for harvesting metadata, which is already explained in Section 2.3.2, the metadata will be re-created or downloaded into Nesstar publisher to make sure it complies with OCS standards. All the related materials (technical reports, questionnaires etc.) will be downloaded from the "parent" source and integrated in the FAM catalog with the created/ harvested metadata.

The following steps will be followed to disseminate metadata from external contributing catalogs:

Step 1: Get permission from external contributor to share their metadata in the FAM.

Step 2: Harvest metadata by downloading the XML file from an external catalog or calling the XML file through API and download the external documents (technical reports, questionnaires, etc.). If XML file is not available for download, metadata is re-created from reports and materials.

Step 3: Create a folder for the harvested metadata and documents by using the naming convention for internal archiving.

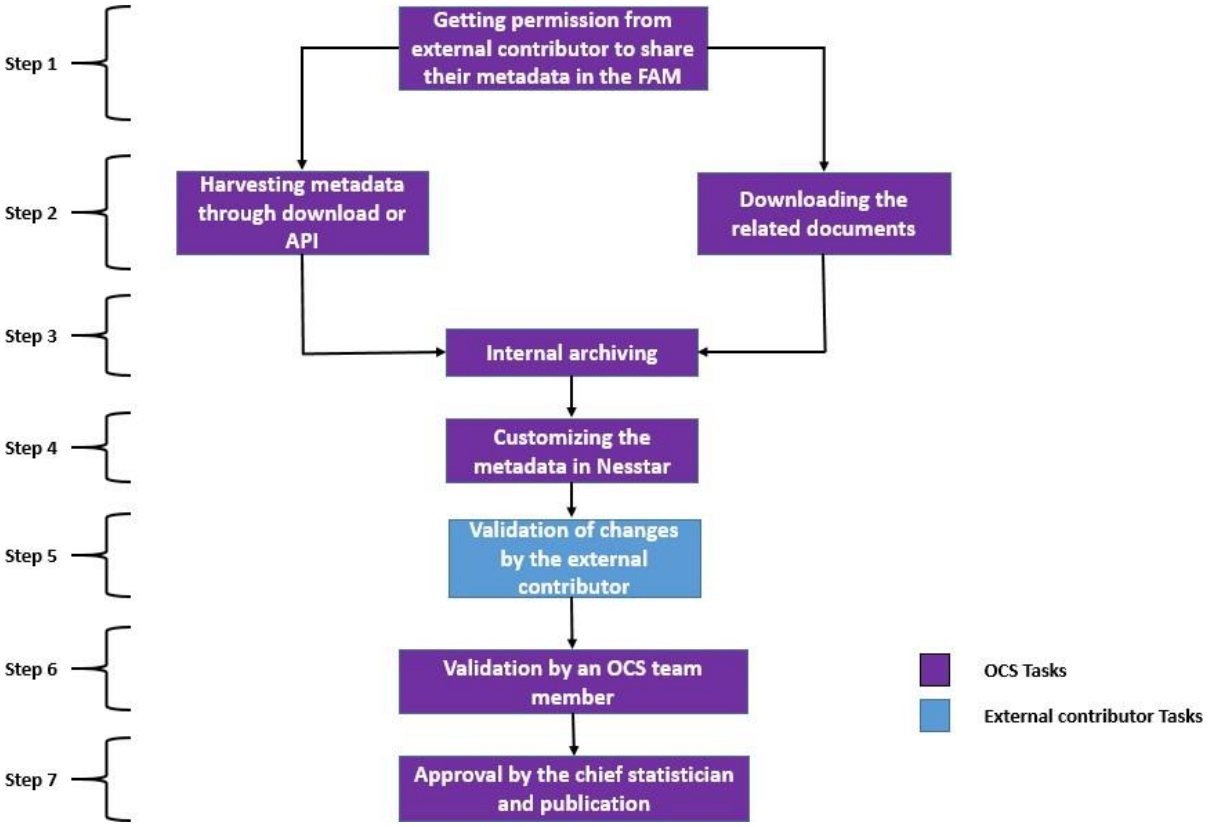
Step 4: Edit/ create the metadata in Nesstar and customize it to comply with OCS template and standards.

Step 5: In case major changes are made in the metadata, validate them with the external contributor.

Step 6: Validate the customized metadata by an OCS team member.

Step 7: Approval by the Chief Statistician and publication of the metadata and documents.

Figure 4: Workflow for metadata harvested from external catalogs



Annex 1: Naming convention for Food and Agriculture Microdata Library

Introduction

It is important to set a standard in terms of naming the file and folders in the archiving process. This will allow the organization of the files and folders in a proper way and facilitate reproducibility.

In fact all the metadata files must have a unique ID that will be populated in the catalog. Hence, having a clear and simple ID generation for files would help to avoid duplication.

The goal in naming the files and folders is to identify the country, the study/survey, the version of the files and all the other related documents and data.

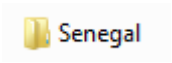
The folders' tree is organized as below:

In the following section, we are going to provide the naming convention for each level of the tree.

Folder naming

Level 1: country name

The country name should match with the countries in the M49 list. An example is



Level 2: study/survey

The study/survey folder is the second level of the tree. The elements of this level are:

ISO_Year(s)_Acronym of the study/survey_Frequency_Module

*The elements are separated by an "**underscore**"*

ISO (mandatory)

This is the 3 letter ISO alpha 3 code of the country as specified in the M49. In our example the ISO code for Senegal is **SEN**. It is a mandatory element of this level.

Year (s) (mandatory)

The year refers to the start of the data collection. In case the data collection is overlapping between years, you must consider the year the data collection started.

Example: SEN_2019

In case of panel studies or times series, the years (s) element is constituted by the first and last years of the study separated by a **hyphen**

Example a panel data (2016, 2017, 2018, 2019): SEN_2016-2019

Acronym

The acronym is the study abbreviation known for the survey. For example Agricultural Production Estimates Survey will have APES as acronym.

It is important to use the known survey acronym. The acronym should always remain in the original language of the study. In case the survey does not have a known acronym, create an acronym by combining the first letter of the words in the survey title.

Example: SEN_2019_APES

Frequency (optional)

Some studies/surveys are conducted many times during a year. In that case, the frequency element will be required. The frequency can be daily, weekly, monthly, every semester, every tree months etc. To indicate that dimension in the folder, the letter W (as wave) will be added. Then the numbering of the wave will follow.

In that case, it will be important to provide the time interval in the metadata to characterize whether it is daily, weekly, monthly etc.

The frequency is separated to the acronym by a **"hyphen"**

Example: First round of APES in Senegal conducted in semester 1 and second round conducted in semester two.

In this case, two folders at the second level will be created

SEN_2019_APES-W1

SEN_2019_APES-W2

In case the study is not conducted for many waves/rounds, this element is not required and shouldn't be in the name of the folder.

Module (optional)

Some studies/surveys might need to be archived considering each module separately. This could happen in case the data access to the module is different from the access to other modules.

In that case, the name of the second level folder should contain that information. Hence, an abbreviation needs to be create for the related module e.g. CC for Crop-cutting module, SP for social protection module etc.

In case no distinction is made between modules, this element of the naming will not be reflected.

However, it will be important to indicate in the metadata the survey module concerned.

The module is separated to the acronym or frequency by a **"hyphen"**

Example: Agricultural Practices (AP) module of the APES survey in Senegal in 2019

SEN_2019_APES-AP

Examples of different scenarios for the second level folder

Ex 1: Annual Agricultural Survey in Malawi in 2014

MWI_2014_AAS

Ex 2: Panel data on livestock survey (LIVE) in Malawi between 2012 and 2016, cattle module (CAT)

MWI_2012-2016_LIVE-CAT

Ex 3: Second round of the Annual agricultural survey of Malawi in 2016, plot module (PLOT)

MWI_2016_AAS-PLOT

Example of folders organization level 1 to level 2

Level 3: version

We need to distinguish two different categories in this level: the master folder and the adaptations' folders.

Master folder

The master folder is the one that contains the origin files as given or deposited by the data provider in the FAM Library without any change from OCS.

The name of the folder will be as follow:

ISO_Year(s)_Acronym of the study/survey_Frequency_Module_Masterversion_Language_M

Version (mandatory)

The version is most of the time v01. In case a second version of the study is released by the data provider, a new folder will be created and will be v02

Language (Mandatory)

This is the language of the study. The code of the language can be found here: http://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

M as Master (mandatory)

The letter M indicating Master is mandatory and should always be found in the name of the third level folder.

Example: SEN_2019_APES_v01_EN_M

Adaptations' folders

Adaptations are versions of the study that have been in some way changed/edited, maybe by the data curator or by the data provider in collaboration with the data curator. It can be removal of variables, reshaping of the datasets, anonymizing data etc.

The adaptations are placed in their own folders at the third level of the tree.

The naming of the adaptations' folders is a continuation of the master file folder even though they are at the same level (level 3 of the tree).

The naming is established as follow:

ISO_Year(s)_Acronym of the study/survey_Frequency_Module_Masterversion_Languague_M_AdaptationVersion_A_AdaptionInstitution

Adaptation version

The version of the adaptation start from v01 and should be in increment whenever a change is made in the study after submission of the master files. The version is mandatory if the folder is an adaptation

A (as Adaptation)

The letter A is mandatory if the folder is for an adaptation.

Adaptation Institution

It refers to the institution or department that made the changes in the Master files. It will be mandatory in the case of adaptations' folders.

The changes should be reflected in the metadata.

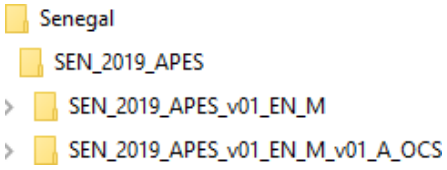
Example 1: APES Survey in Senegal in 2019 with the direct identifiers removed by OCS

SEN_2019_APES_v01_EN_M_**v01**_A_OCS

Example 2: after removal of the direct identifiers, anonymization was performed

SEN_2019_APES_v01_EN_M_**v02**_A_OCS

Example of folders organization level 1 to level 3



Level 4: Study folders

The fourth level has 4 sub-folders called study folders. It also contains the Nesstar file, the DDI and the RDF files.

Study folder 1: Data

It contains two folders: a folder with the original data and a folder for the data that should be distributed to users.

Study folder 2: Docs

It contains three sub-folders: the questionnaires, the reports and the technical documents.

Study folder 3: Programs

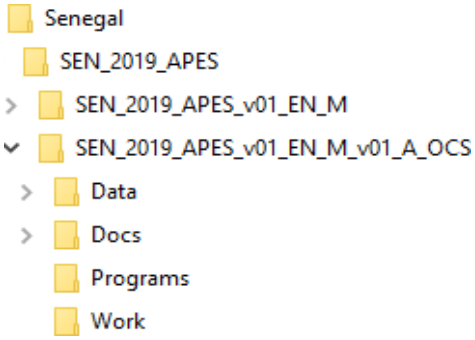
It contains all the statistical programs used (program for anonymization, program for cleaning etc.).

Study folder 4: Work

It is used to archive the temporary work files during the process of documenting the study.

The Nesstar file, the DDI and the RDF files are all present in the root of the level 4 folder.

Example of folders' organization: level 1 to level 4



Files, study ID and DDI document ID number naming

File names

The Nesstar, DDI and RDF files all take the same name as given to the version folder (level 3) in which they are stored

Study ID

The study ID has the same name as the version folder (level 3).

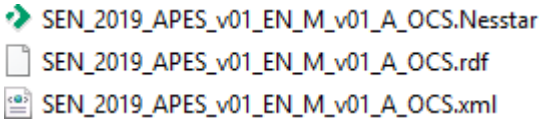
DDI document ID Number

The DDI document ID number is constituted as followed: DDI, the name of the level 3 folder and the Institution that prepare the DDI.

Example: DDI_SEN_2019_APES_v01_EN_M_FAO

DDI_SEN_2019_APES_v01_EN_M_v01_A_OCS_FAO

Example of File naming



Annex 2: Application for Access to a Licensed Dataset

First Name: [inserted automatically by system]

Last Name: [inserted automatically by system]

Organization: [inserted automatically by system]

Email: [inserted automatically by system]

Dataset requested: [inserted automatically by system]

This form must be filled and submitted by the Lead Researcher. Lead Researcher refers to the person who serves as the main point of contact for all communications involving this agreement. Access to licensed datasets will only be granted when the Lead Researcher is an employee of a legally registered non-commercial receiving agency (university, research center, national or international organization, etc.) on behalf of which access to the data is requested. The Lead Researcher assumes all responsibility for compliance with all terms of this Application for Access to a Licensed Dataset by employees of the receiving organization.
This requestwill be reviewed by a data release committee, who may decide to approve the request, to deny access to the data, or to request additional information from the Lead Researcher. A signed copy of this request form may also be requested.

This request is submitted on behalf of:

Receiving organization name: [manual entry]

Telephone (with country code): [manual entry]

Intended use:

Please provide a short description of your research and/or statistical project (i.e. research question, objectives, methods, expected outputs, collaborators/partners)
[User enters information]

List of expected outputs and expected dissemination outlet(s) and/or strategy:

[User enters information]

Expected completion date of research project

[User enters DD-MM-YYYY]

Research team members (other than Lead Researcher) which will have access to dataset. Provide name(s), title(s), and affiliation(s) of any other members of the Lead Researcher's team who will have access to the dataset:

[User enters information]

Identification of data files and variables needed

FAO provides detailed metadata on its website, including a description of data files and variables for each dataset. Researchers who do not need access to the whole dataset may indicate which subset of variables or cases they are interested in. As this specifies a subset of variables, rather than the whole dataset, providing this information may increase the probability that the data will be provided.

This request is submitted to access

Select one:

- ○ The whole dataset (all files, all cases)

- A subset of variables and/or cases as described below (note that variables such as the sample weighting coefficients and records identifiers will always be included in subsets):
  [User enters information]

Terms of Use of Data Access Agreement

- The Lead Researcher's organization and other researchers who will be involved in using the data at that organization must be identified. The Lead Researcher certifies that he/she is authorized to sign on behalf of the Lead Researcher's organization. If not, a suitable representative must be identified. Any violation of the Terms of Use of this agreement will be considered to have occurred on behalf of the Lead Researcher's organization, and FAO will take appropriate measures to sanction such misconduct, which may include denying any further and future access by the Lead Researcher organization or any other researchers involved in using the data ;
- The micro dataset will only be used for the stated statistical and/or research purpose. It shall not in any way be used for other purposes, including any administrative, commercial or law enforcement purposes.
- Any results derived from the micro dataset will be used solely for reporting aggregated information, and not for any specific individual entities or data subjects;
- The Lead Researcher nor anyone else authorized in this agreement shall take any action with the purpose of identifying any individual entity (i.e. person, household, enterprise, etc.) in the micro dataset(s). If such a disclosure is made inadvertently, no use will be made of the information, and it will be reported immediately to FAO;
- The micro dataset and other materials provided by the Food and Agriculture Microdata Catalogue will not be re-disseminated, or sold or otherwise shared with anyone other than the Lead Researcher and anyone else authorized in this agreement without the written agreement of FAO.
- No attempt will be made to produce links or matching among datasets provided by FAO, and any other datasets that could identify individuals or organizations.
- Any books, articles, conference papers, theses, dissertations, reports, or other publications that employ data obtained from FAO will cite the source of data in accordance with the Citation Requirement provided with each dataset.
- An electronic copy of all reports and publications based on the requested data will be sent to FAO.
- The Lead Researcher shall implement security measures to prevent unauthorized access to this micro dataset. The micro dataset must be destroyed upon the completion of the research, unless FAO obtains satisfactory guarantee that the micro dataset(s) can be secured and provides written authorization to the Lead Researcher in this respect.
- FAO may monitor, at its discretion, use of datasets obtained from the Food and Agricultural Microdata Catalogue, and decide whether an abuse has taken place under these Terms of Use. In such case, FAO may sanction users for violations. Penalties may include restrictions or denial of further access to the Food and Agriculture Microdata Catalogue.
- The designations employed in the dataset do not imply the expression of any opinion whatsoever on the part of FAO concerning the legal or development status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Information contained in the dataset is provided on an "as is" and "as available" basis. No guarantee is given that the

information is correct, complete or up-to-date. FAO does not represent or endorse the accuracy, completeness, authenticity or reliability of any information contained in the dataset. FAO shall not be held liable for any loss or damage arising from, or directly or indirectly connected to, the use of, reference to, or reliance on any dataset, including, but not limited to, any liability arising from any interpretation or inferences based upon the data, nor from any intentional or negligent misuse, errors, disclosure, undue transfer, loss or destruction of data that may occur

- The Lead Researcher will ensure that the Terms of Use are shared with any individual authorized to download datasets.
- FAO reserves the right to amend these Terms of Use at its own discretion. Any amendment affecting the conditions agreed upon under these Terms of Use will be notified to the Lead Researcher;
- Nothing contained in or related to these Terms of Use shall constitute or be interpreted as a waiver, express or implied, of the privileges and immunities of FAO.

○ I have read and agreed with the conditions

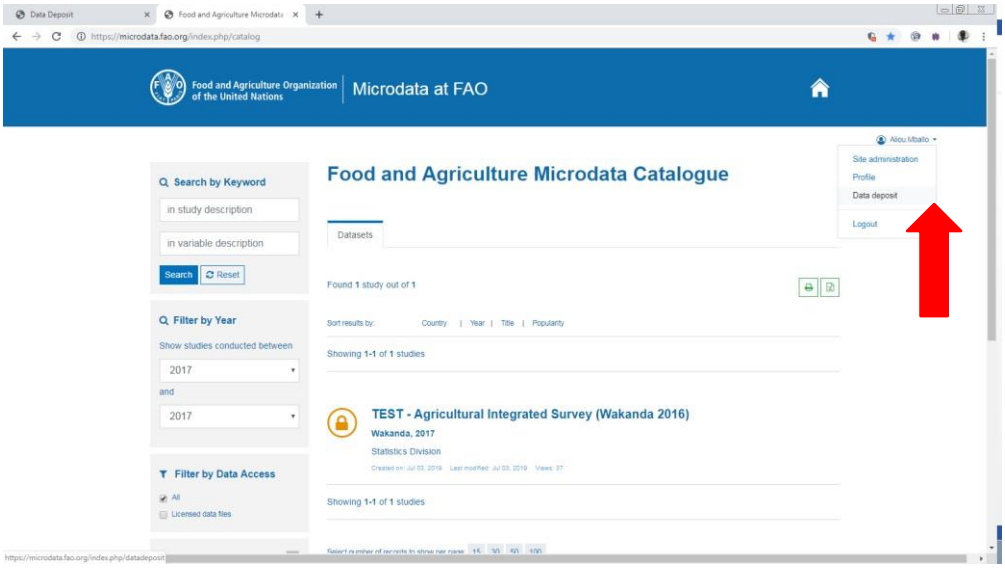
Annex 3: Data deposit

The data deposit is a system put into place which allows data providers to request dissemination of their metadata and/or microdata in the FAM catalog. Users of the data deposit should have, before starting the process, the information about the metadata, the micro datasets and all the relevant document (i.e. questionnaire, enumerator manuals, technical reports etc.).

Here are the steps to deposit metadata and/or microdata in the FAM catalog.

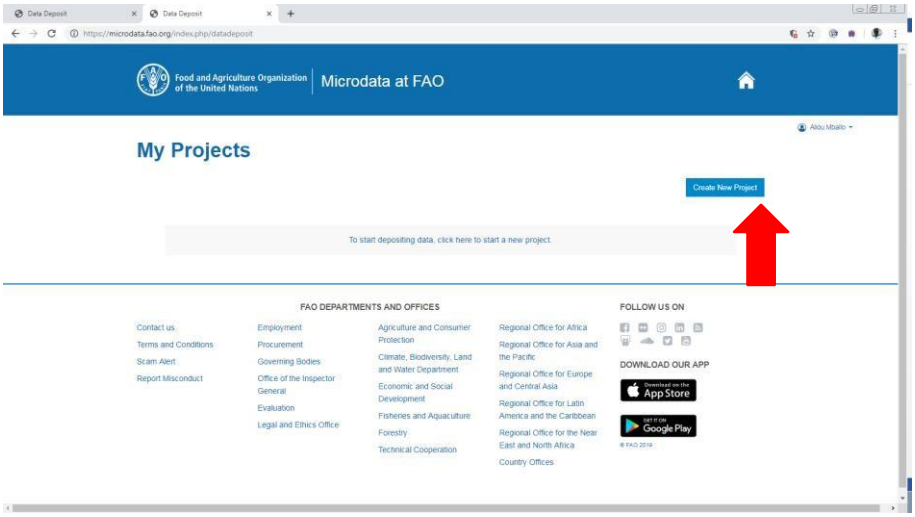
Step 1: Where to find the data deposit?

After logging into https://microdata.fao.org, click on your login name on the upper right of the page, and select on **"Data Deposit"** as seen in the figure below.

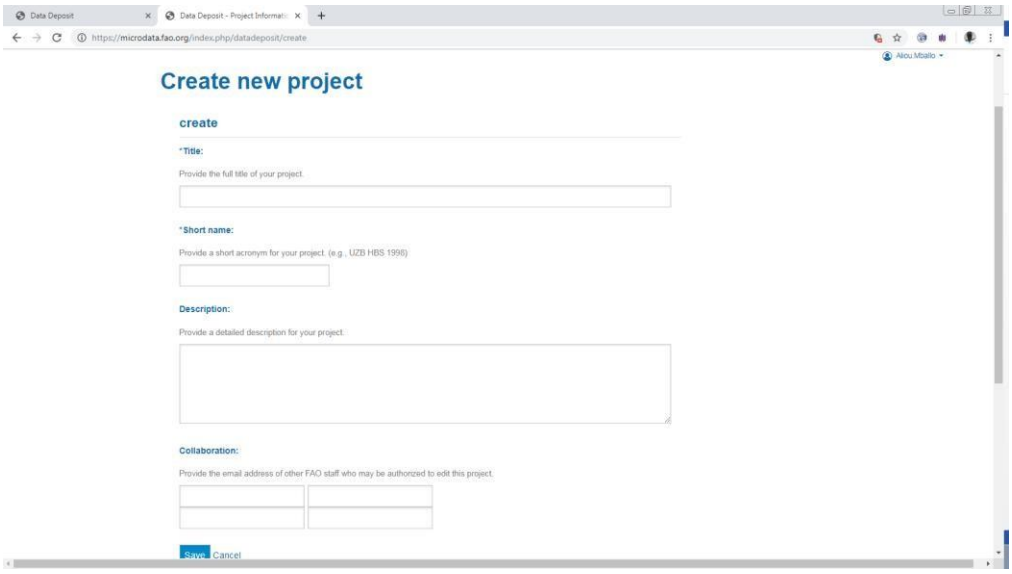


Step 2: Create/edit a project

If it is the first time you are depositing data, you will see the following page. Otherwise, you will also see the former projects you have created.



To start a new project, click on **"Create new project"**. Then the following page will appear:

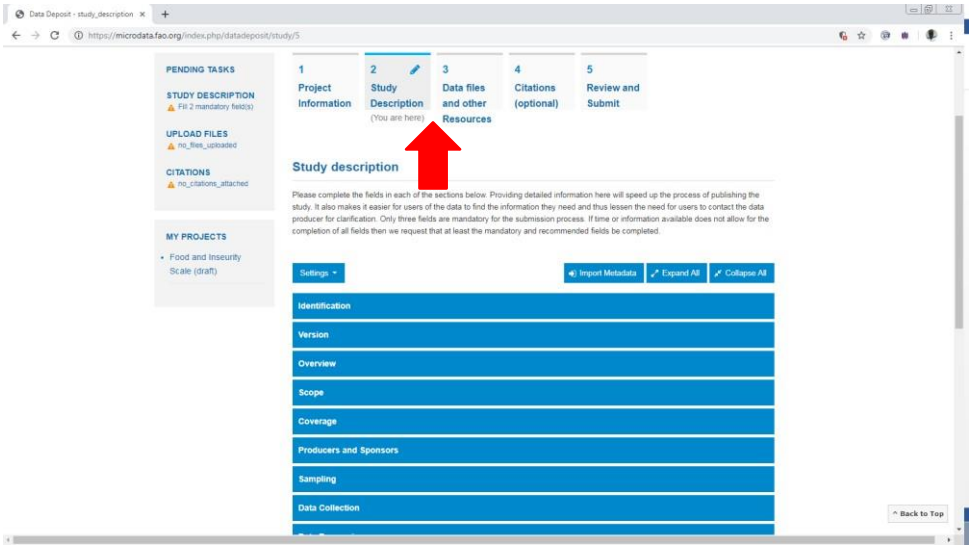


In this page, first you must provide the **Title of the study** (i.e. usually the name of the survey) and its short name or acronym. Then, complete the fields providing a short description of the project and give the email addresses of other FAO staff who may be authorized to edit the project.

Click **"Save"** after filling in all the information and a message should display indicating that the changes have been saved.

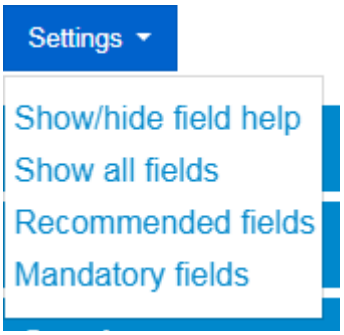
Step 3: Provide metadata

Click on the **"Study description"** tab to enter metadata about the study.

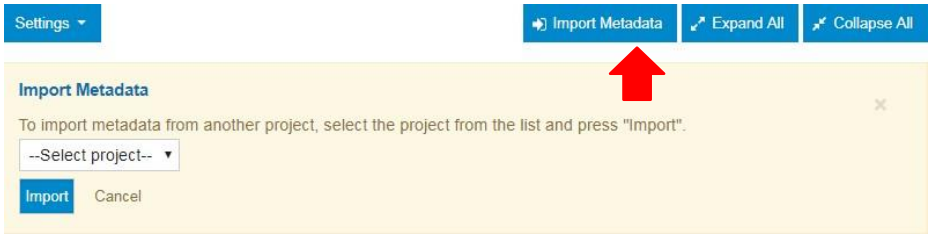


In this page:

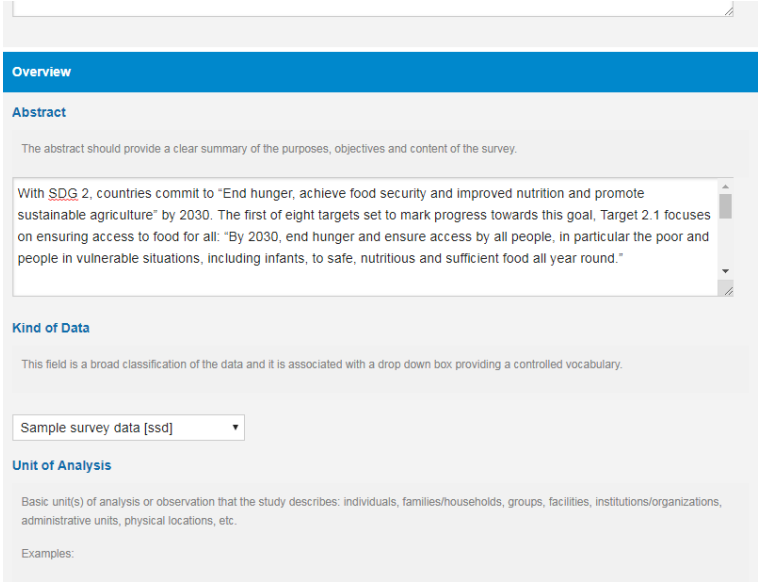
- There is the "Settings" tab that allows you to show or hide the help, show all fields, to see the recommended and mandatory fields. It is recommended to choose "Show/hide field help" for the first few studies that you enter. When help is displayed, more detailed instructions will be displayed for each metadata field.



- If you have a study with similar metadata, for example a previous wave of a longitudinal study, you can import metadata from another project by clicking "Import Metadata" and selecting the corresponding project.



- To see all the fields, you can click on "Expand All." Then, all the metadata elements will appear. Notably, some fields are mandatory, and others are not.

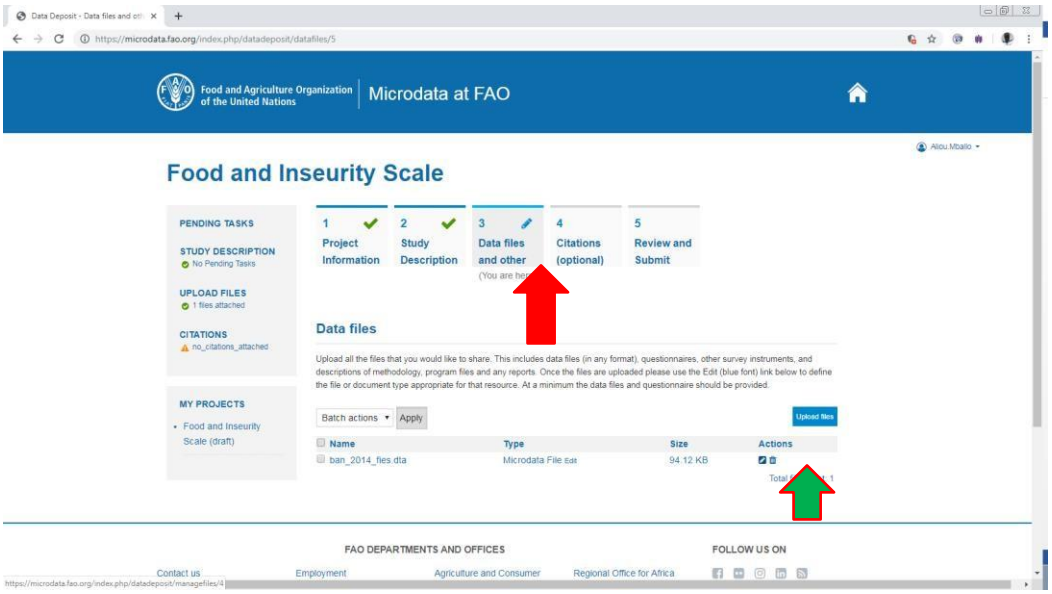


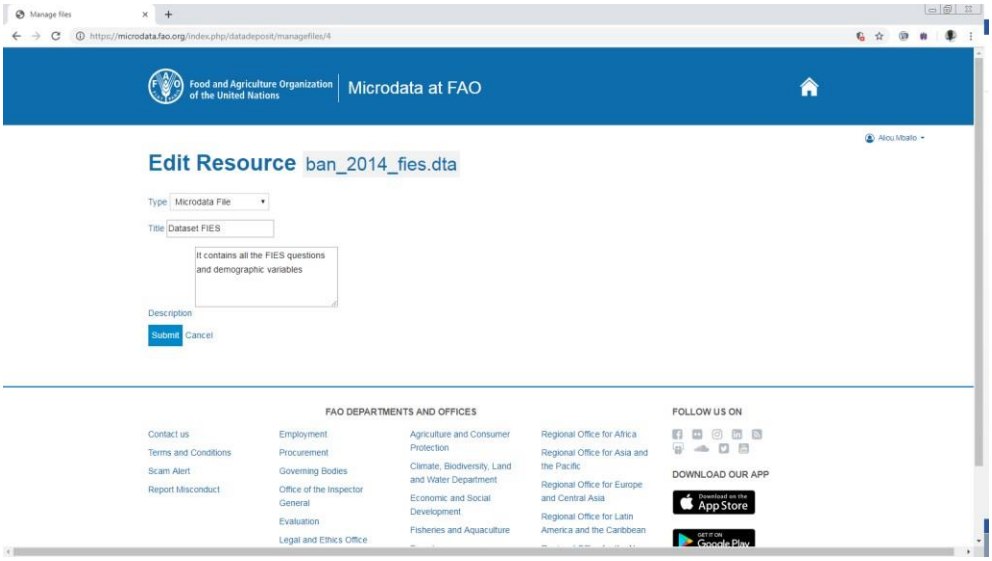
After having filled all the mandatory fields, and as many of the non-mandatory fields as possible, click **"save"** at the bottom of the page.

Step 4: Upload the microdata files and other resources

Next click on the "Data files and other Resources" tab to upload the datasets and all related technical documents such as questionnaires, technical reports etc.

After uploading each a file, you will need to edit it by choosing the “Type” of file, the title and a description of the content.



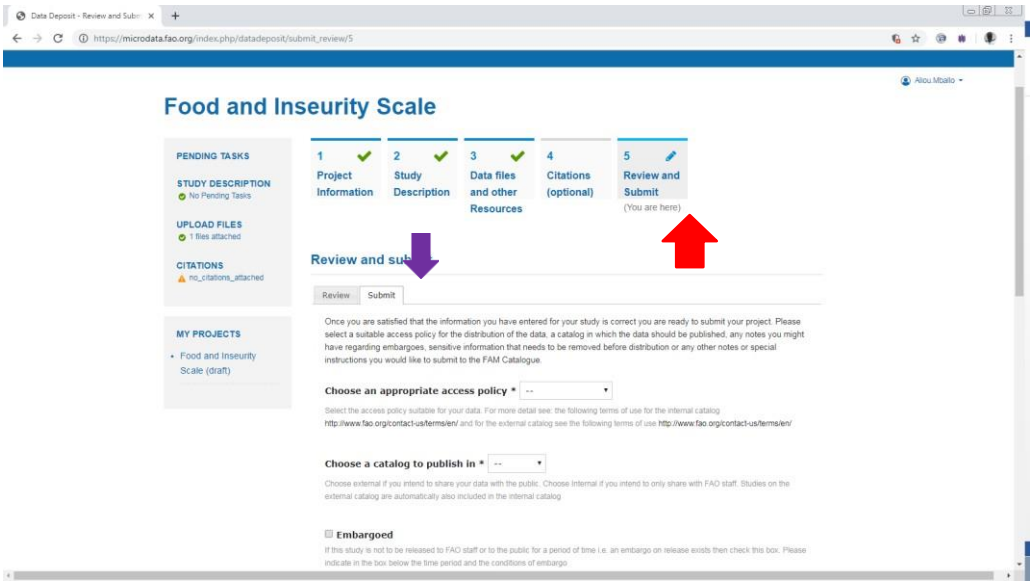


Step 5: Provide citations/references if applicable

In case you have published work that use the datasets, please provide the references by clicking on the “Citations (optional)” tab.

Step 6: Review and submit the project

To submit click on the **“Submit”** tab. Then you will have to choose the appreciate access policy, the catalog in which to publish and the additional information required.



References

Benschop T. et al. 2018. *Statistical Disclosure Control: A practical Guide.* Accessed at https://buildmedia.readthedocs.org/media/pdf/sdcpractice/latest/sdcpractice.pdf

CCSA. Microdata dissemination best practices https://unstats.un.org/unsd/accsub-

public/microdata.pdf

Creighton et al. DDI: a metadata standard for the social sciences. https://www.ddialliance.org/sites/default/files/ddiposterORMtg07.pdf

Dupriez O. and Boyko E. 2010. *Dissemination of microdata files: principles, procedures and practices.* IHSN Working Paper. n° 005. pp 68

IHSN. *Technical note in metadata standards*. Accessed at: http://www.ihsn.org/sites/default/files/resources/DDI_SDMX_IHSN_DRAFT.pdf

Organisation for Economic Cooperation and Development (OECD). 2007. "OECD Principles and Guidelines for Access to Research Data from Public Funding". www.oecd.org/dataoecd/9/61/38500813.pdf

Templ M. et al. 2015. *Statistical disclosure control for micro-data using R package sdcMicro*. Journal of Statistical Software. Volume 57. Issue 4.

Templ M. et al. 2014. *Introduction to Statistical Disclosure Control (SDC)*. IHSN Working Paper. n°007

United Nations Economic Commission for Europe (UNECE). 2007. *Managing Statistical Confidentiality and Microdata Access: principles and guidelines of good practice*. Accessed at: https://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.a nd.microdata.access.pdf

United Nations System: Personal Data Protection and Privacy Principles. Accessed at https://www.unsystem.org/personal-data-protection-and-privacy-principles