Evaluation of Microdata and Metadata quality

The *Pacific Data Hub Microdata Library (PML)* receives datasets from many different sources. As a general rule, we do not modify the datasets unless we work directly with the producer, except to apply statistical disclosure control and to format the data files for the convenience of users. Data files are always preserved in their original format, as well as in a Stata-consistent format. For dissemination the Library provides for data files to be converted into other commonly used formats such as SPSS, SAS, STATA or ASCII. Documents are stored in their original format but are usually disseminated as PDF files.

SPC works with various data producers to promote better practices of data management including variable and file naming rules and the use of labels. Because generally speaking the PML has no control over the data collection or management procedures used, there is no guarantee that these practices will have been used in any specific survey from its inception.

While an evaluation of microdata and metadata can be undertaken. Data are often provided "as is". We make all possible efforts to ensure that the metadata are as comprehensive as possible. This documentation includes, whenever possible, identification of problems and weaknesses in the datasets. It is, however, the responsibility of the researcher to make his/her own assessment of the reliability and suitability of the data for his/her specific purpose, based on all the information provided.

SPC collects microdata through data collections such as census and survey for a variety of purposes. Unlike input and output data that is typically aggregated, microdata is unique in that it is collected—and can be reported—at the individual, household, enterprise, and/or community level. It also tends to have two distinguishing characteristics: it is personally identifiable and can be sensitive.

Personally Identifiable Information is any information that can be used, on its own or in conjunction with other information that is linked or linkable to a specific individual, to determine the identity of an individual or otherwise locate or contact the individual. It includes:

- *Direct Identifiers*: such as the individual's full name, date of birth, mailing or home address, email address, telephone number, GPS coordinates, national identification number, physical/biological identifiers (physical appearance, through photo or video data collection, fingerprints, DNA, etc.); and
- *Indirect Identifiers*: These include unique, observable or other characteristics that may enable re-identification even when direct identifiers are removed. Risk of re-identification is closely linked to the population the sample is drawn from and understanding how likely an outlier in the data is an outlier in the population 2 .

Sensitive data is information that may pose a risk to the individual or firm if it is collected or released in a way that is linkable to the individual or firm responding to a survey. This type of data may include income, assets, or health status, the public release of which could harm survey respondents.

Data Quality Checking:

All datasets deposited with PML undergo quality checks to confirm the accuracy and usability of the data. Anomalies in data files and documents are corrected in consultation with data depositors. Missing values, errors and corrections are recorded as Data Quality Notes in the metadata provided with each dataset.

Data structure, completeness and correctness are checked- for example the structure, size and type, completeness, and correctness of the dataset agrees with description of the dataset content and with level of data curation within the PML repository. For example, it is important to make sure that, in all data files, the identification variables(s) provide a unique identifier. Use the duplicate function in SPSS or *isid* command in Stata to verify this. Verify the completeness of your data files by comparing the content of these files with the survey questionnaire. Make sure that data from all sections of the questionnaire are included in the dataset. Verify that the number of records in each file corresponds to what is expected.

Metadata Completeness is also checked for the microdata file for example whether a citation exists, including authorship, year, comprehensive title, persistent identifier (e.g. DOI). Any data files with data quality changes will receive a new version number. File naming and versioning is according to the Data Documentation Initiative (DDI) standard. Verify that all variables are labelled (variable labels) and that the codes for all categorical variables are labelled (value labels).

Evaluating a dataset is a crucial process following it's acquisition (please refer to data acquisition). This piece of work ensures the dataset is ready for preservation and documentation and is made of the following steps:

1) Tabulating aggregates from a report,
2) Integrity of the dataset,
3) Checking whether IDs are unique,
4) De-identification of the dataset,
5) Identification of sensitive variables,
6) Labelling all variables and all values.

1) **Tabulating aggregates from a report**: The very first thing to do when the dataset is acquired is to "test" its quality by comparing it with a few tables from a report, if a report there is. Most of the time, it is better to try to reproduce some of the most basic tables such as "Population by island and by sex" or "Households by island and by Type" or "Expenditure by type" or "Income by type"…etc. This way, it enables the curator to make sure they are working with the final dataset consistent with the external resources they are going to link in their documentation.

2) **Integrity of the dataset**: The integrity of a dataset can be evaluated by checking if the variables of the dataset are consistent with the questionnaire. It is important that all variables should be in the same order as the questions.

3) **Checking whether IDs are unique**: This is probably one of the most important and complex tasks when evaluating a dataset. It consists of evaluating the uniqueness of IDs in your dataset by concatenating all of the IDs and using a command (on Excel or Stata) to confirm if all households / individuals of the dataset are unique. ID variables can be for instance: “Island code”, “Enumeration Area”, “Household Number” or “Person Number”. If the dataset’s IDs are unique, it will then be possible to merge some variables from a dataset to another (e.g.: import the “weight” variable from the “Household” record to the “Person” one).

4) **De-identification of the dataset**: De-identification refers to the process of removing all the direct identifiers in a way that it is no longer possible to directly identify individuals or households. Direct identifiers can be “First Names”, “Last Names”, “Dates of Birth”, “GPS locations” or any other variable that is directly identifying.

5) **Identification of sensitive variables**: This process has to be done jointly between the curator and the producer of the data in order to identify all potential sensitive variables. Sensitive variables are those whose information should not be discovered for any respondent in the dataset due to ethical, political or legal reasons: e.g.: “Sexual Behaviour”, “Criminal Records”…etc. Sensitive variables can also be variables that when combined together, can contain personally identifying information – which is greatly relevant for small Pacific Island Countries. These sensitive variables can be: “Age” + “Gender” + “Religion” + “Level of Education”. Individually, these variables are not identifying someone specifically, but when combined all together - in some cases - they can.

6) **Labelling all variables and all values**: Lastly, it is important to make sure that all variables and all values are labelled as it will provide the researcher with a solid and well-documented dataset where all variables and values have a meaning. The labelling of the dataset can be done using any software (e.g.: Stata) but not Excel as it does not allow the application of labels whatsoever.

Evaluation of Metadata Quality

Once your documentation is fully completed using NESSTAR / Metadata Editor, it is important to make sure it is of best quality and that nothing of high importance has been omitted. Fortunately,

there are several ways to help us monitoring the quality of our newly-produced metadata[1]! Here is a list of some of them:

1) Validation tools provided in the NESSTAR / Metadata Editor software,
2) Checking if all "Recommended" fields have been entered,
3) DDI Reviewers' Feedback Form[2],
4) Checking at spelling mistakes and the formatting of your documentation,
5) Double-checking your documentation on NADA (data description, links to docs).

1) **Validation tools provided in the NESSTAR / Metadata Editor software**: NESSTAR / Metadata Editor are equipped with powerful tools that help you validating the completeness of your documentation. All these tools can be found under the "Tools" menu. The first tool is called "Validate Metadata", it enables the users to make sure they haven't forgotten any "Mandatory[3]" field in their documentation. The second tool is called "Validate External Resources" which verifies all "Mandatory" fields in the External Resources are filled in. Finally the last tool is called "Validate Dataset Relations", it is a nice way to check the structural integrity of the identifier variables and to make sure there is no duplicate in your dataset.

2) **Checking if all "Recommended" fields have been entered**: This is going to be a manual (but simple) task, it is also important to make sure all "Recommended[4]" fields are filled in as they provide fairly important information (such as: "Geographic Coverage" of the study, "Type of Version" of the dataset, "Dates of Collection").

3) **DDI Reviewers' Feedback Form**: An independent review of the data and metadata is highly recommended prior to publishing the final documentation on NADA. This form provided by the International Household Survey Network (IHSN) is a checklist that records how to enter each of the DDI sections. This document can be sent to an external reviewer in order to get some feedback on one's work. You can access it here.

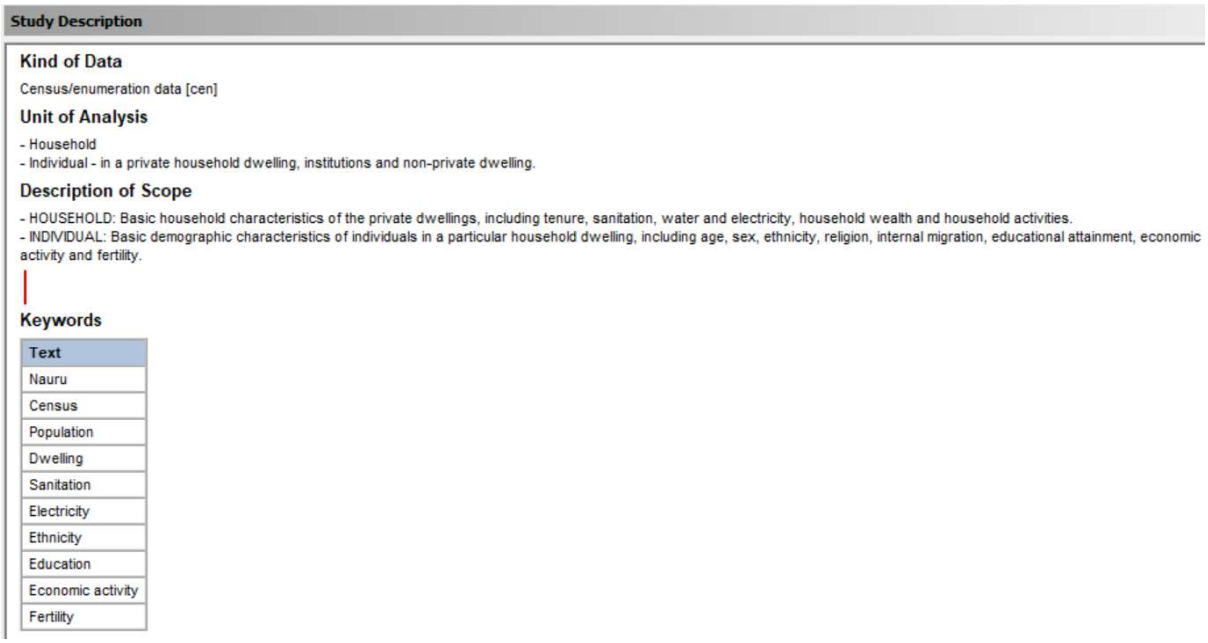
4) **Checking at the formatting of your documentation**: This may seem like a detail, but having a well-structured documentation that is pleasant to look at is more attractive for researchers and other users. Make sure there is no empty spaces after the end of a section like for instance at the end of "Description of Scope".

[1] First of all, you need to make sure you are using the IHSN templates in your documentation software. These templates can be found here: http://www.ihsn.org/software/ddi-metadata-editor under the "IHSN Templates" section.

[2] That can be downloaded here: https://data-archivists-guide.readthedocs.io/en/latest/_downloads/DDI-reviewers-feedback-form.pdf

[3] "Mandatory" fields are the most important pieces of information that cannot be omitted or skipped (e.g.: Title of your documentation, country where it took place, Primary Investigators…).

[4] "Recommended" fields are not mandatory but really important too and should not be skipped either if possible.



Study Description

Kind of Data

Census/enumeration data [cen]

Unit of Analysis

- Household
- Individual - in a private household dwelling, institutions and non-private dwelling.

Description of Scope

- HOUSEHOLD: Basic household characteristics of the private dwellings, including tenure, sanitation, water and electricity, household wealth and household activities.
- INDIVIDUAL: Basic demographic characteristics of individuals in a particular household dwelling, including age, sex, ethnicity, religion, internal migration, educational attainment, economic activity and fertility.

Keywords

Text
Nauru
Census
Population
Dwelling
Sanitation
Electricity
Ethnicity
Education
Economic activity
Fertility

In order to make the documentation readable for everyone, you should also make sure that all names of organizations, departments…etc are not abbreviated when firstly cited in each of the sections.

5) **Double-checking your documentation on PML**: Finally, this is where your final output will be published so it has to be perfect in terms of quality and aspect. Just a quick double-check of all the sections (Study Description and Data Description) is expected here. Should you also make sure the “External Resources” documents are all well-linked and hence, downloadable in the Documentation section.