Quick Reference Guide for Microdata Archivists

Quick Reference Guide for Data Archivists — World Bank Group Chief Statistician's Office (DECDG)

Acknowledgements

Version 2026-08 is a revision of Version 2019-04[^1], originally authored by Olivier Dupriez (World Bank) and subsequently updated by Diana Marcella Sanchez Castro and Matthew Welch (The World Bank). Their contributions to the guide's original structure, content, and conceptual framework are gratefully acknowledged.

This edition was revised and updated by Cathrine Machingauta and Mehmood Asghar. Key enhancements include:

- Updated references to World Bank tools that have replaced tools mentioned in earlier versions.

- Adoption of DDI-Codebook, now an ISO standard, replacing references to DDI 1.6 used in the original guide.

- Revisions and enhancements to the following sections:

- Introduction

- Folder organization and file-naming conventions

- Preparing data

- Documenting the Study Description

- External Resource Documentation

- Data Discoverability (new section)

- Generating Output

- Quality Review (including an updated quality review checklist)

- Administrative Metadata Documentation (new section)

- Publishing to NADA (new section)

These updates reflect current tools, standards, and practices for data curation, documentation, preservation, and dissemination.

[^1]: The Quick Reference Guide for Data Archivists. Publisher: The International Household Survey Network (IHSN). The World Bank, Development Data Group (DECDG) is a member of the IHSN Secretariat.

> **Disclaimer**
>
> Portions of this document were generated with the assistance of artificial intelligence (AI). The content has been reviewed, edited, and validated by the authors; however, users should independently verify information and exercise professional judgment when applying it.

Introduction

Microdata are detailed, unit-level data collected from surveys, censuses, and administrative systems. Each record represents a specific unit of observation, such as a person, household, business, or institution, and contains information about its characteristics, activities, or conditions. Because collecting, processing, and maintaining microdata requires a significant investment of time, expertise, and resources, these data are valuable institutional assets.

To ensure that microdata can be understood, discovered, accessed, and reused over time, they must be accompanied by comprehensive metadata. Metadata provide the context needed to find/identify a dataset, interpret the dataset, describe what the data measure, how they were collected, processed, and organized, who produced them, and under what conditions they can be accessed and used. The **Data Documentation Initiative Codebook (DDI-C)** and the **Dublin Core Metadata Initiative (DCMI)** are internationally recognized ISO standards - useful for providing a structured framework for documenting microdata and related resources in a consistent, machine-readable[^1], and interoperable[^2] manner.

In organizations that generate and process microdata, microdata archivists and data stewards play a critical role in applying these standards to preserve, document, organize, manage, and disseminate microdata while protecting respondent privacy and confidentiality. Through effective stewardship and standardized metadata, microdata remain understandable, discoverable, secure, accessible, and reusable, maximizing their value for research, policy development, and evidence-based decision-making.

This *Quick Reference Guide for Microdata Archivists* provides data archivists with guidelines for documenting micro-datasets in compliance with The Data Documentation Initiative Codebook (DDI Codebook/DDI-C) and The Dublin Core Metadata Initiative (DCMI) metadata standards[^3].

This Guide summarizes the process in the following chronological steps:

1. Organizing the folder structure
2. Gathering and preparing the data set
3. Gathering and preparing the documentation
4. Documenting the study, data file(s) and variables
5. Documenting external resources
6. Metadata Quality assessment
7. Generating the output for publication
8. Cataloging data, metadata and resources

Microdata archiving is part of the wider Data Management Lifecycle. It supports verification and processing; protection, documentation and packaging; storage and preservation; dissemination; and later evaluation and use.

This guide introduces the key concepts and practices for microdata archiving, including preparing data and organizing resources for cataloging and dissemination. It covers the creation of structured, machine-readable metadata, as well as the cataloging of microdata and related resources. Where relevant, the guide provides references to World Bank-supported open-source tools, including **sdcMicro** for data anonymization, the **Metadata Editor** for documenting datasets and resources and generating machine-readable metadata, and **NADA** for cataloging and disseminating data collections.

This guide is not intended to serve as a user manual for these tools. Comprehensive instructions for their installation, configuration, and use are available in their respective documentation. Additional information about these tools can be found here:

- sdcMicro - Statistical Disclosure Control

- Metadata Editor Documentation

- NADA Cataloging Software

To get started, the next section outlines the key principles and best practices for organizing datasets and related resources within a repository. Establishing a well-structured repository is an essential first step that facilitates efficient data curation, metadata creation, cataloging, and dissemination.

[^1]: Machine-readable metadata is structured metadata that software can process automatically.

[^2]: Interoperability refers to the ability of different systems, tools, and platforms to exchange, interpret, and use data and metadata consistently, enabling seamless integration and collaboration across various data management and analysis environments.

[^3]: DDI-Codebook (DDI-C) and DCMI (Dublin Core Metadata Initiative) are international XML metadata specifications. For more information on these standards, please visit https://ddialliance.org/ddi-codebook and https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

1. Before You Start: Organize Your Files

Documenting a dataset is significantly easier and more efficient when data files, metadata, documentation, scripts, and supporting resources are organized in a clear and consistent manner from the outset. A well-structured file organization system helps ensure that:

- **Files are easy to identify, locate, and understand**, reducing the time spent searching for information and helping collaborators navigate project resources efficiently.

- **Important files are protected from accidental overwriting, duplication, or loss**, preserving the integrity of the dataset and its associated documentation throughout the project lifecycle.

- **Sensitive or restricted content is kept separate from materials intended for sharing or public dissemination**, supporting compliance with data protection requirements and reducing the risk of unintended disclosure.

- **Documentation activities can be completed more quickly and accurately**, as all relevant materials are readily available and clearly categorized.

- **Collaboration is improved**, enabling team members to work with a common understanding of where files are stored and which versions should be used.

- **Data preservation and long-term reuse are enhanced**, making it easier to archive, locate archived versions, transfer, and maintain datasets over time.

Recommended Folder Structure

The following folder structure separates original data, working files, dissemination packages, documentation, programs, analytical outputs, and archived materials. This organization helps ensure that files are easy to locate, prevents accidental overwriting, and clearly distinguishes sensitive working files from materials intended for dissemination. We recommend that, before anything else, you create the necessary directories as follows:

| Folder Structure | Purpose |

|---|---|

| 🗂️ **UGA*2026*DHS** | **Root Directory** – Create a directory for the dataset and all associated resources (the archival package). We recommend using a clear and consistent naming convention that includes the country aor geographic scope (where applicable), the reference year or period, and a short identifier for the data collection. For example, **UGA*2026*DHS** for a "Demographic and Health Survey" of Uganda collected in 2026 or **KEN*2025*EMIS** for a Kenya Education Management Information System dataset for 2025. This directory serves as the root folder for all materials related to the microdata collectio, including data files, documentation,metadata, code, and supporting resources. Avoid spaces or special characters in folder names. Use underscores (_) or hyphens (-) as separators to ensure compatibility across operating systems, software applications, and repository platforms. A consistent naming convention improves organization, discoverability, version control, and long term preservation.|

| ├── 📁 **UGA*2026*DHS*v01*M** | Create sub-directories to hold different versions of the data. The original version of the data will always be designated as the Main file. In this example, the first version (v01) of the **M**ain data would be stored in the folder named **UGA*2026*DHS*v01*M.** As stated above, avoid spaces or special characters in the folder name. |

| ├──├── 📁 **Data** | The Data sub-directory contains datasets throughout their lifecycle, from acquisition to dissemination (if applicable). Data files are organized in marked sub-folders to minimize risk of overwriting and sharing of restricted information. |

| ├──├──├── 📁 **Original**
↳ Raw, Backup | Store source data exactly as received from the producer. These files should be preserved and remain unchanged. |

| ├──├──└── 📁 **Edited** | Store edited data separate from the raw data. |

| ├──├──├── 📁 **Working**
↳ Data Entry, Cleaning, Harmonization, Quality Assurance, Anonymization | This folder can hold intermediate files used during processing, validation, transformation, and anonymization. |

| ├──├──└── 📁 **Dissemination**
↳ v01*A*PUF, v02*A*PUF, v01*A*LUF | If data is being disseminated, create this dissemination sub-folder to separately store final approved datasets organized by release version and access type. In this example, the files prepared for dissemination are considered adaptations of the main data, so the folder names include the adaptation version information and end with *A*PUF, *A*SUF etc. |

| ├──├── 📁 **Documentation** | Create documentation sub-directories to hold materials that describe, support, or govern the study and its data. As needed, materials can be further organized into separate sub-directories as shown below. |

| ├──├──├── 📁 **Administrative**
↳ Acquisition, Agreements, Correspondence, Permissions | Store administrative records related to dataset acquisition, licensing, and governance. |

| ├──├──├── 📁 **Project Materials**
↳ Questionnaires, Manuals | Store data collection instruments and fieldwork guidance. |

| ├──├──├── 📁 **Reports**
↳ Project Reports, Methodology, Sampling Reports, Evaluations | Store reports describing the data production process, methodology, and findings. |

| ├──├──├── 📁 **Technical**
↳ Codebooks, Data Dictionaries, Sample Design, Geospatial material, data processing and quality assurance documents etc. | Store documentation required to understand and use the data. |

| ├──├──└── 📁 **Photos**
↳ Fieldwork, Instruments, Maps | Store photographs and visual resources associated with the study. |

| ├──├── 📁 **Programs** | Create sub-directories for scripts and code used to process, validate, anonymize, tabulate, and analyze data as needed. |

| ├──├──├── 📁 **Data Management**
↳ Data Entry, Cleaning, Anonymization, QA | Store programs used to prepare dissemination-ready datasets. |

| ├──├──└── 📁 **Analysis**
↳ Tabulation, Indicators, Research | Store statistical and analytical programs used to produce outputs. |

| ├──├── 📁 **Outputs**
↳ Tables, Figures, Maps, Publications | Store products generated from the data, including statistical outputs and publications. |

| └──└── 📁 **Archive**
↳ Previous Releases, Superseded Files, Deposits | Store historical versions and archived materials retained for preservation and audit purposes. |

Example: Dissemination Folder Structure

The dissemination package is a curated subset of the archival package containing only the resources approved for distribution. While the archival package preserves all materials associated with the data lifecycle, the dissemination package includes only those resources that can be shared and are required to support data discovery, access, interpretation, and reuse.

The dissemination folder contains the dissemination package, which includes the data and resources approved for sharing, publication, or distribution to authorized users. Depending on the dissemination model and access restrictions, this package may contain data files, metadata, documentation, codebooks, analytical code, replication programs, and other materials that help users understand, interpret, analyze, and reuse the data.

Not all resources contained in the main project or archive folder should be included in the dissemination package. Certain documents may contain sensitive, confidential, proprietary, personally identifiable, or otherwise restricted information and should therefore be retained only in the archival collection. Examples include internal correspondence, administrative records, quality control materials, disclosure assessments, security-related documentation, and other materials that are not intended for public release.

Before dissemination, all files should be reviewed to ensure they comply with applicable confidentiality, privacy, licensing, legal, and organizational requirements. The dissemination package should contain only those resources necessary to support the appropriate use, interpretation, and reproducibility of the disseminated data.

``

📁 Dissemination

│

├── 📁 UGA*2026*DHS*v01*M*v01*A_PUF

│ ├── 📁 Data

│ ├── 📁 Documentation

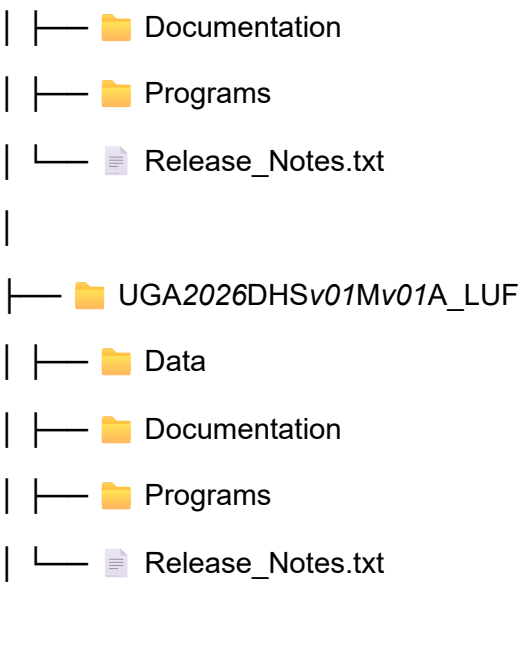
│ ├── 📁 Programs

│ └── 📄 Release_Notes.txt

│

├── 📁 UGA*2026*DHS*v01*M*v02*A_PUF

│ ├── 📁 Data



`

Dissemination Version Naming Convention

| Folder Name | Description |

|---|---|

| v01APUF | First release of the Public Use File package |

| v02APUF | Second release of the Public Use File package |

| v01ALUF | First release of the Licensed Use File package |

| v01ACUF | First release of the Confidential Use File package |

| Release_Notes.txt | Documents changes introduced in the release |

::: tip Best Practice

Store each dissemination release as a complete, self-contained package containing the distributed data, documentation, programs, and release notes. This ensures that previous releases remain reproducible and that users can clearly identify which version of the data was used.

When creating structured metadata for a dataset, use the dataset folder name as the Primary ID whenever possible. This ensures a clear and traceable link between the metadata record and the corresponding data files.

For example, when documenting the main dataset, the Primary ID should be **UGA*2026*DHS*v01*M**. If you are documenting Version 2 of the Public Use File, the Primary ID should be **UGA*2026*DHS*v01*M*v02*A_PUF**.

To accommodate diverse user needs and support long-term preservation and accessibility, data should be archived and disseminated in multiple widely used formats, such as Stata, SPSS, CSV, and ASCII.

Using a consistent naming convention helps maintain traceability, supports version control, and makes it easier to manage and identify related metadata and data assets over time.

:::

Once the folders and files have been properly organized, the next step is to assess the data to ensure it is adequately prepared for curation, archiving, and, where appropriate, dissemination. This review helps confirm that the data meet quality, documentation, and preservation requirements before the curation process begins. The key considerations for data preparation are covered in the next section.

---

2. Gathering and Preparing the Data Set

Gathering and preparing data is a process that requires great care. Prior to documenting a dataset, it is important to ensure that you are working with the most appropriate version of all the concerned data files. If the dataset is meant for public release, one should work with the final, edited, anonymous version of the dataset. If the dataset is being documented for archiving and internal use only, one may include the raw data as well as the final, fully edited files. If you are working with administrative data, more information on documentation of administrative data is provided in the annex. The Metadata Editor provides you with the possibility of documenting the specificity of each version of the dataset.

Much of the quality of the output generated by the Editor will depend upon the prior preparatory work. Although you can make changes to the data from the Metadata Editor, it is highly recommended that the necessary checks and changes be made in advance using a statistical package and a script. This ensures accurate and replicable results.

This section describes the various checks and balances involved in the data preparation process, such as: making a diagnostic on the structure of your data, cleaning it and identifying the various variables at the outset. Listed below are some common data problems that users encounter:

- Absence of variables that uniquely identify each record of the dataset

- Duplicate observations

- Errors from merging multiple datasets

- Encountering incomplete data when comparing the content of the data files with the the source documentation, data collection instruments, data specifications, or system design documents

- Unlabelled data

- Variables with missing values

- Unnecessary or temporary variables in the data files

- Data with sensitive information or direct identifiers

Some practical examples using a statistical package are provided in Section A: Data Validations in Stata.

> **Note:** If you are working in a data archive, be careful not to overwrite your original variables. Since managing databases involves several data-checking procedures, archive a new version in addition to the original. Work on this new version, leaving the original data files untouched.

The following procedures are recommended for preparing your dataset(s):

2.1. Data Files Should Be Organized in a Hierarchical Format

Look at your data and visualize it to understand its structure. It is preferable to organize your files in a hierarchical format instead of a flat format. In a hierarchical format, columns contain specific information about all possible units of analysis and rows form the individual observations (households, establishments, products, communities/countries, or any combination of those). Hierarchical files are easier to analyse, as they contain fewer columns that store the same information and are more compact. A flat format contains multiple columns with information on only one specific unit of analysis, so the information becomes redundant. For example, the information provided in one column is about the household head, and the row provides information on the child in the household.

Table 1. Flat Format

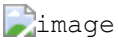
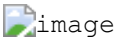


Table 2. Hierarchical Format



Tables 1 and 2 illustrate the two data structures. They contain the same information about six people on age and their relationship with the head of the household. The flat dataset (Table 1) stores the information on each family member in a new column. Note that for every additional member or characteristic, the dataset gets flatter and wider. The hierarchical dataset (see Table 2) has one observation and one row per person. Each variable contains a value that measures the same attribute across people and each record contains all values measured on the same person across variables. For every additional member characteristic, the dataset maintains the same number of columns, gains additional rows and gets longer and less flat compared to *Table 1*. The first two columns of this dataset have a hierarchical structure, where the ID member column is nested inside the ID household column.

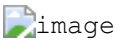
Hierarchical files are easier to manage. Suppose in this example that there were many characteristics measured for everyone, the hierarchical structure would be a more convenient format because for each new characteristic, the dataset creates only one additional column, whereas, in the flat structure, it would create as many columns as there are people in the data with such characteristics.

Datasets With Multiple Units of Analysis Should Be Stored in Different Data Files

It is recommended that you store your data in different files when you have multiple observational units. For example, *Table 3* shows a dataset that has both household-level data (columns on the type of dwelling and walls material) and individual-level data (columns on the marital status, work status, and worker category). Note that storing both levels of information in one dataset will result in a repetition of household characteristics for each household member. In Table 3, the information about the columns 'type of dwelling' and 'wall material' is repeated for everyone. Sometimes, this

duplication is inefficient, and it is easier to have the dataset broken down by observational unit, into multiple files. In this example, it would be simpler to create two files: one for the household characteristics and another for the individual characteristics. The two files can be connected through a unique identifier, which in this case will be the household ID and member ID. We discuss the need for this unique identifier further on in this text as well.

Table 3. Single data set with more than one observational unit

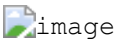


Columns in a Dataset Should Represent Variables, Not Values

It is recommended that columns represent variables (e.g., sex, age, marital status) and rows represent observations (e.g., individuals, households, firms, products and so forth). In some datasets, columns instead of describing variables or attributes, describe values, which means that one variable is broken into segments and each one is stored in different columns. While this dataset structure can be useful for some analysis, the standard data structure where columns are variables and not values is the norm.

For example, *Table 4* (options 1 and 2) gives information at the individual-level on marital status, relationship with the head of the household and age. The difference between both tables is how the variable 'age' is reported. Option 1 had broken the variable 'age' into segments. This practice makes your data: (i) messier, it has values of the variable as headings, and (ii) inefficient, it increases the size of the dataset. Option 2 is recommended since there is only one heading and store of the information occupies less space, allowing the user to identify the structure of the data in a clear manner.

Table 4. Data Structures: Hypothetical datasets



2.2. Check File Structure and Coverage

The dataset should contain all variables, fields, or attributes defined in the source documentation, data collection instruments, data specifications, or system design documents, except those intentionally excluded for reasons such as confidentiality, privacy protection, legal restrictions, or data minimization requirements.

Verify the completeness of the data files by comparing their contents with the relevant source documentation, including questionnaires, forms, administrative records, data collection instruments, data dictionaries, system specifications, codebooks, or metadata documentation. This review helps ensure that all expected variables have been captured and appropriately documented.

Variables should be organized in a logical and consistent manner that reflects the structure of the data source, business process, or data production workflow. For example, variables may be grouped by subject area, module, record type, processing stage, or functional domain. A well-

organized dataset makes it easier for users to understand relationships among variables, navigate the data efficiently, and connect the data to the supporting documentation.

Data management tools can be used to generate an inventory of variables, including their names, labels, data types, formats, and other characteristics. Reviewing this inventory helps verify that no variables have been omitted, that metadata are complete and accurate, and that variables are organized consistently across files. This process also provides an opportunity to identify duplicate, undocumented, or improperly labeled variables before the data are archived or disseminated.

Example: The Stata command *-describe-* displays the names, variable labels and other characteristics, which helps us verify that no variables have been omitted in the database. It simultaneously confirms that all variables are correctly ordered. Refer to Example 7 for further details.

2.3. Verify That the Number of Records in Each File Corresponds to Expectations

The technical and methodological documentation should provide enough information to establish reasonable expectations about the size and structure of the dataset. Verify that the number of records in each file is consistent with what is described in the documentation, data production process, or source systems.

For example, if the documentation states that a dataset contains records for 50,321 households, establishments, businesses, facilities, beneficiaries, or other units of observation, the corresponding data file should contain a similar number of records. If discrepancies exist, they should be explained and documented.

Even when expected record counts are not explicitly provided, it is often possible to perform reasonableness checks by examining the relationships between files. For example:

A file containing individual-level records would typically contain multiple records for each household-level record.

A transaction-level file would generally contain more records than the file containing the entities associated with those transactions (for example, customers, firms, facilities, or beneficiaries).

An event-level file would typically contain multiple records for each subject, case, or reporting unit represented in a master file.

A file containing service encounters, claims, purchases, assessments, visits, or observations would generally contain more records than the file containing the persons, households, organizations, or locations associated with those records.

When datasets contain hierarchical relationships, the number of records in lower-level files should be broadly consistent with the documented structure of the data and the expected number of related records per unit. Significant deviations may indicate missing data, duplicate records, processing errors, or incomplete extracts and should be investigated and documented.

These checks help confirm that all expected records have been captured and that the relationships between files are consistent with the documented design and intended use of the data.

2.4. Each Observation in Every File Must Have a Unique Identifier

Before you check for uniqueness of the identifiers in your files, you need to figure out the unit of analysis. Even if you are not the data producer, it is often easy to identify it. You can always review the documentation to see if the information has been provided. Below, some examples of units of analysis:

Table 5. Unit of Analysis by Study Type

| Study Type | Unit of Analysis |

|---|---|

| Income/Expenditure/Household Surveys | • Households
• Individuals
• Consumption items |

| Enterprise Surveys/Census | • Firms
• Establishments/Plants |

| Agricultural Surveys/Census | • Households
• Crop area |

| Research Data | • Schools
• Financial transactions
• Exported products
• Municipalities/precincts |

Once you recognize the unit of analysis, the next step is to identify the column that uniquely identifies each record. If a dataset contains multiple related files, each record in every file must have a unique identifier. The data producer can also choose multiple variables to define a unique identifier. In that case, more than one column in a dataset is used to guarantee uniqueness. These identifiers are also called **key variables**[^1] or **ID variables**. The variable(s) should not contain missing values or have any duplicates. They are used by statistical packages such as SPSS, Stata, R or Python when data files need to be merged for analysis

The absence of a unique identifier is a data quality issue, so one needs to ensure that the unique IDs remain fixed/present during the data cleaning process. If this correction is not possible, the archivist should note the anomalies in the documentation process.

::: tip Best Practices

- It is recommended that ID variables be defined as a numeric since sorting and filtering records is much more efficient when variables are numeric.

- ID variables should not contain spaces, special characters or accents, since they may suffer modifications when the dataset is converted in different formats.

- For the convenience of users of the data, avoid identifiers consisting of too many variables. For example, in a household-level microdata file, the household identifier should ideally be a single variable (which you may create by concatenating a group of variables), and the individual identifier should be the combination of only two variables (the household ID, and the sequential number of each member).

- It is recommended that you generate an ID based on a sequential number, however, keep in mind that it should not be too long because statistical packages and spreadsheet programs store a

number of digits of precision, so opening a data set that contains ID variables with many characters, might result in truncated fields. For instance, the limit of the number of characters in Microsoft Excel is 15, so it changes any digits past the fifteenth place to zeroes.

- If you prepare your data files for public dissemination, it may be preferable to generate a unique household identification that would **not** be a compilation of geographic codes (because geographic codes are highly identifying). This recommendation is to ensure anonymity and will be explained in further detail later on in this text. The following example shows how to construct a unique identifier without using detailed information provided by

the geographic codes.

:::

Example: Creating a Unique Identifier

Suppose the unique identification of a household is a combination of variables PROV (Province), DIST (District), EA (Enumeration Area), HHNUM (Household Number). Options 2 and 3 are recommended. Note that if option 3 is chosen, it is crucial to preserve (but not distribute) a file that would provide the mapping between the original codes and the new HHID.

| PROV | DIST | EA | HHNUM | HHID (concatenated) | HHID (sequential) |

| --- | --- | --- | --- | --- | --- |

| 12 | 01 | 014 | 004 | 1201014004 | 1 |

| 12 | 01 | 015 | 001 | 1201015001 | 2 |

| 13 | 07 | 008 | 112 | 1307008112 | 3 |

| Etc | Etc | Etc | Etc | Etc | Etc |

Option 1 uses a combination of four variables (PROV, DIST, EA, HHNUM). *Option 2* generates a concatenated ID. *Option 3* generates a sequential number.

Once you recognize the unit of analysis and the variable that uniquely identifies it, the following checks are suggested:

- Even if a dataset contains a variable labeled as a "unique identifier," it is important to verify that the variable truly identifies each record uniquely. To confirm the uniqueness of an identifier, or to determine which variable or combination of variables uniquely identifies observations, users can employ the *Duplicates* procedure in SPSS, the *isid* command in Stata, the *isid()* function in R, or the *is_unique* property of a pandas index in Python (see Table 6) to verify that each observation is uniquely identified. For more details, refer to Example 1 and Example 2.

Table 6. Check for unique identifiers: STATA/R/PYTHON/SPSS Commands

STATA

```stata
```

use “household.dta”

isid key1 key2

`

R

`r

my_data <-

load("household.rda")

id <-c("key1" , " key2")

library(eeptools)

isid(my_data, id, verbose=FALSE)

`

Python

`py

import pandas as pd

df = pd.read_stata("household.dta")

Below syntax returns True IDs in index are unique

---

df.set*index(["key1", "key2"]).index.is*unique

Below assertion fails if IDS are not unique

---

assert df.set*index(["key1", "key2"]).index.is*unique

`

SPSS

` sps

Load dataset and choose from the menu:

- Data > Identify Duplicate Cases

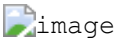
- Select Key Variables

`

- Finally, check that the ID variable for the unit of observation doesn't have missing or assigned zero/null values. Ensure that the datasets are sorted and arranged by their unique identifiers.

Table 7 below gives a hypothetical example. In this dataset, the highlighted columns (hh1, hh2, hh3) are the key variables, which means that they are supposed to make up the unique identifier. However, looking at those variables, we can identify some problems: the key variables do not uniquely identify each observation as they have the same values in rows 4 and 5, they also have some missing values (represented by asterisks), assigned zero values and some null values (those that say NA, don't know). All these issues suggest that those variables are not the key variables, and one needs to go back and double-check the data documentation. Alternatively, the archivist could check with the data producer and ask them how to fix these variables, in case those are indeed the key variables.

Table 7. Check for unique identifiers: Hypothetical data set



Example 3 in Section A: Data Validations in Stata provides further details and describes the steps involved in performing a validation when the identifier is made of multiple variables.

2.5. Identifying Duplicate Observations

One way to rule out problems with the unique identifier is to check if there are duplicate observations (records with identical values for all variables, not just the unique identifiers). Duplicate observations can generate erroneous analysis and cause data management problems. Some possible reasons for duplicate data are, for example, the same record being entered twice during data collection. They could also arise from an incorrect reading of the questionnaires during the scanning process if paper-based methods are being used.

Identifying duplicate observations is a crucial step. Correcting this issue may involve eliminating the duplicates from the dataset or giving them some other appropriate treatment.

Statistical packages have several commands that help identify duplicates. *Table 8* shows examples of these commands in STATA, R, Python and SPSS. The STATA command *-duplicates report-* generates a table that summarizes the number of copies for each record (across all variables). The command *-duplicates tag-* allows us to distinguish between duplicates and unique observations. For more details, refer to Example 4.

Table 8. Check for duplicates observations: STATA/R/PYTHON/SPSS commands

STATA

```stata
use “household.dta”
duplicates report
duplicates tag,
generate(newvar)
```

R

```r
my_data <-
load("household.rda")
household[duplicated(household),]
```

Python

```py
import pandas as pd
df=pd.read_stata("household.dta")
```

list all duplicated records in a dataframe

df[df.duplicated()]

`

SPSS

`sps

Load dataset and choose from the menu:

- Data > Identify Duplicate Cases

execute.

`

2.6. Ensure That Each Individual Dataset Can Be Combined into a Single Database

For organizational purposes, microdata is often stored in different datasets. Therefore, checking the relationship between the data files is an essential step to keep in mind throughout the data validation process. The role of the data producer is to store the information as efficiently as possible, which implies storing data in different files. The role of the data user is to analyse the data as holistically as possible, which could sometimes mean that they might have to join all the different data files into a single file to facilitate analysis. It is essential to ensure that each of the separate files can be combined (merged or appended depending on the case) into a single file, should the data user want to undertake this step.

Use statistical software to validate that all files can be combined into one. For a household survey, for example, verify that all records in the individual-level files have a corresponding household in the household-level master file. Also, verify that all households have at least one corresponding record in the household-roster file that lists all individuals. Below, some considerations to keep in mind before merging data files:

- The variable name of the identifier should be the same across all datasets.

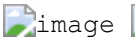
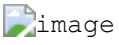
- The ID variables need to be the same type (either both numeric or both string) across all databases.

- Except for ID variables, it is highly recommended that the databases don't share the same variable names or labels.

Example: Joining Household and Child Datasets

A household survey is disseminated in two datasets; one contains information about household characteristics and the other contains information on the children (administered only to mothers or caretakers). To build a dataset containing all the information about the household characteristics, including where the children live, one needs to combine these files. Users are thus assured that all observations in the child-level file have corresponding household information.

Joining data files: Hypothetical data set

Statistical packages have some commands that allows us to combine datasets using one or multiple unique identifiers. *Table 9* shows examples of these commands/functions in STATA, R and SPSS. For more details, refer to Example 5.

Table 9. Joining data files: STATA/R/PYTHON/SPSS commands

STATA

```stata
use “household.dta”
merge 1:m hh1 hh2 hh3
using
"individuals.dta"
```

R

```r
household <- load("household.rda")
individuals <- load("individuals.rda")
md <- merge(household, individuals,
by = c("hh1", "hh2", “hh3”),
all =TRUE)
```

```

**Python**

```py

import pandas as pd

household = pd.read_stata("household.dta")

individuals = pd.read_stata("individuals.dta")

md = pd.merge(

household,

individuals,

on=["hh1", "hh2", "hh3"],

how="outer"

)

```

SPSS

```sps

Load dataset and choose from the menu:

- Data > Merge Files > Add Variables

- Select the data file to merge

- Select Key Variables

```

Panel datasets should be stored in different files as well. Having one file per data collection period is a good practice. To combine the different periods of a panel dataset, the data user could merge them (Adding variables to the existing observations for the same period) or append them (Adding observations for a different period to the existing variables). To make sure that panels can be properly appended, the following checks are suggested:

- Check for the column(s) that identifies the period of the data (Year, Wave, Series, etc.).
```

- The variable names and variable types should be the same across all datasets.

- Ensure that the variables use the same label and the same coding across all datasets.

To combine datasets vertically, use the following

- SPSS: *"Append new records"*

- STATA: *append*

- Python: *pd.concat(household, individuals)*

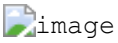
- R: rbind(household, individuals) or bind_rows(household,individuals)

2.7. Check That the Data Types Are Correct

Do not include string variables if they can be converted into numeric variables. Look at your data and check the variables' types, particularly for those that you expect to be numeric (age, years, number of persons/employees/hours, income, purchases/expenditures, weights, and so forth). If there are numeric variables stored as string variables, your data needs cleaning.

For example, *Table 10* contains a data set at the individual-level with some variables that should be numeric. The columns B (Age) and E (Working Weeks) are stored as numeric variables, which is fine. However, the variables 'Number of working of hours per week' (Column G), 'Number of persons working at the business' (Column H) and 'Monthly Income' (Column I) are loaded as strings because there are non-numeric values (don't know, skip, refused to answer) and some missing values present. Those variables need to be cleaned and converted from string variables to numeric variables.

Table 10. Checking Data Types: Hypothetical data set



Statistical packages have some commands that allows us to make such conversions. *Table 11* shows examples of these commands/functions in STATA, R, PYTHON and SPSS.

Table 11. Convert string variables to numeric: STATA/R/PYTHON/SPSS commands

STATA

```stata
use "individual.dta"
destring (varname),generate
```

R

```r

individual <-

load("individual.rda")

replace}

as.numeric(individual$varname)

```

Python

```py

import pandas as pd

df = pd.read_stata('individual.dta')

df["varname*num"] = pd.to*numeric(df["varname"], errors="coerce")

```

SPSS

```sps

Load dataset and choose from the menu:

- Data > Transform > Recode into Same individual$varname <-

- Select the variable

- Select "Old and New Values" and Recode it

- Select "Convert numeric strings to numbers ('5'->5)

```

2.8. Check for Variables With Missing Values

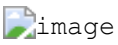
Getting data ready for documentation also involves checking for variables that do not provide complete information because they are full of missing values. This step is important because

missing values can have unexpected effects on the data analysis process. Typically, missing values are defined as a character (.a, .b, single period or asterisks), special numeric (-1, -2) or blanks. Variables entirely comprised of missing values should ideally not be included in the dataset. However, before excluding them, it is useful to check whether the missing values are expected according to the questionnaire, and the skip patterns.

For example, a hypothetical household survey at the individual-level (Table 10) provides information about the respondent's employment status. The survey identifies if the respondent is employed in Column D, and then provides information about the worker category in Column E, but only for those who reported being employed in Column D. This means that those who answered 'unemployed' in column D should have a valid missing value in column E. In other words, this is a pattern in the missing values that should be observed and duly noted.

On the other hand, Columns F and G are used to determine if the people who are not employed are looking for a job and are actively seeking it. These questions are not asked to the employed people (those who answered "yes" in Column D), which mean that again, the missing values in those columns correspond with what is expected. However, Column H contains information for all employed individuals, so missing values in this column suggest that there is a problem in the data and should be addressed. Therefore, one should not blindly delete missing values at the outset without checking for these patterns.

Table 12. Checking for Missing Values: Hypothetical data set



In SPSS, use the function *"Missing Value Analysis"* and in R, do as shown in *Table 12*. You can also use the STATA command *-misstable summarize-* that produces a report that counts all the missing values. You can also use the *-rowmiss()-* command with *-egen-* to generate the number of missing values among the specified variables. For more details, refer to Example 6.

Table 13. Counting Missing Values: STATA/R/PYTHON/SPSS commands

STATA

```stata
use “individual.dta”
misstable summarize
```

R

```r
```

my_data <-

load("individual.rda")

colSums(is.na(individual))

colMeans(is.na(individual))

`

Python

`py

import pandas as pd

df = pd.read_stata('individual.dta')

df.isna().sum()

to see number and percentage of missing values, use:

---

missing = pd.DataFrame({

"Missing": df.isna().sum(),

"Percent": (df.isna().sum() / len(df) * 100).round(2)

})

print(missing)

`

SPSS

`sps

Load dataset and choose from the menu:

- Data > Analyze > Missing Value Analysis

- Select “Use All Variables”

`

::: tip Best Practices

Since there are different reasons for missing values, data producer

should code them with negative integers or letters to distinguish

the missing values and valid data. For instance, (− 1) might be the

code for "Don't Know", (-2) the code for "Refused to Answer" and

(-9) code for "Not Applicable".

:::

2.9. Check Improper Value Ranges

It is helpful to generate descriptive statistics for all variables (frequencies for discrete variables; min/max/mean for continuous variables) and verify that these statistics look reasonable. Just as there are variables that must take on only specific values, such as "F" and "M" for gender, there are also some variables that can take on several values (such as age or height). However, those values must fit a particular range. For example, we don't expect negative values, or typically see values over 115 years for age.

Values for categorical variables should be guided by the questionnaire (or separate documentation for constructed variables). If we have an education variable that has 9 response options in the questionnaire, the corresponding 'education' variable in the dataset should have 9 categories. We should not observe more than 9 unique values for this variable. Similarly, for any questions for which the options are only "yes", "no" and "other", we should not observe more than these 3 unique values. When out of range values exist, this might signal data cleaning issues.

Table 14 shows examples of some commands/functions in STATA, R, PYTHON and SPSS.

Table 14. Generate descriptive statistics: STATA/R/PYTHON/SPSS Commands

STATA

`stata

use “individual.dta”

summarize

```

**R**

```r

individual <-

load("individual.rda")

summary(individual)

```

Python

```py

import pandas as pd

df = pd.read_stata("individual.dta")

df.describe()

```

SPSS

```sps

Load dataset and choose

from the menu:

- Data > Analyze > Descriptive Statistics > Frequencies

- Select “Statistics"

`

## 2.10. Verify Weights, Design Variables, and Adjustment Factors (Where Applicable)

Some microdata collections include weights, adjustment factors, or design variables that are required for producing valid estimates and analyses. Where such variables exist, verify that they are included in the dataset, clearly labeled, properly documented, and consistent with the accompanying methodology documentation.

For sample-based datasets, weighting variables are often provided to enable users to produce estimates that are representative of a larger target population. In these cases, the documentation should clearly describe how the weights were constructed, how they should be applied, and any limitations associated with their use. Basic validation checks, such as reviewing minimum and maximum values, identifying missing values, and confirming alignment with the documented methodology, can help detect potential issues.

In some datasets, adjustment factors may be included to account for non-response, calibration, post-stratification, benchmarking, or other statistical corrections. Where applicable, these factors should be clearly identifiable and adequately documented.

Not all microdata require weights. For example, administrative records, transaction data, operational systems, registries, and complete censuses may not include weighting variables because they are intended to represent the full population of interest or are not derived from a sample. However, even in these cases, any transformations, adjustment factors, or derived analytical variables used to support analysis should be documented and preserved.

Where the data are based on a sample design (surveys), verify that the variables identifying stratification levels, clusters, primary sampling units, or other design elements are included and clearly documented. These variables are often necessary for estimating sampling errors and producing statistically valid analyses.

More generally, ensure that any variables required to correctly interpret, aggregate, weight, or analyze the data are present, clearly identifiable, and accompanied by sufficient documentation to support their appropriate use by data users.

## 2.11. Variables and Codes for Categorical Variables Must Be Labelled

**Variable Labels**

Variable labels should be concise, precise, and informative. They provide a clear description of the information contained in a variable and help users understand how the data relate to the corresponding literal questions. Without meaningful variable labels, it can be difficult to interpret the contents of a dataset or link variables back to the questionnaire. Therefore, all variables should be clearly labelled.

Even when variables are labelled, the following good practices should be followed:

- Variable labels should be informative, accurate, and as concise as possible. While software packages may allow relatively long labels (for example, up to 80 characters in Stata and 255 characters in SPSS), shorter labels are generally easier to read and manage.

- Avoid using the full literal question as a variable label. Literal questions are often lengthy and may exceed recommended label lengths. Instead, provide a brief description that captures the essence of the question.

- Each variable should have a unique label. The same label should not be used for different variables, as this can create confusion and make analysis more difficult.

- Labels should clearly distinguish between related variables and use consistent terminology throughout the dataset.

- Variable labels should complement, not replace, detailed variable descriptions. While labels provide a short summary, the **variable description**[^2] should capture the full wording of the question, interviewer instructions, concepts being measured, derivation methods, or any other contextual information needed to interpret the data correctly.

- Well-documented labels and descriptions improve data quality by making datasets easier to understand, review, validate, and reuse. They also support metadata extraction, search, and discovery, and help automated tools accurately interpret variables and generate reliable outputs.

**Value Labels**

Label values are used for categorical variables. To ensure the correct encoding of data, it is important to check that the stored values in those variables correspond to what is expected according to the questionnaire. In the case of continuous variables, we also suggest the checking of ranges. For instance, if the question is about the number of working hours, the variable should not have negative values.

You can compare variable labels in the dataset to those in the questionnaire using the *--codebook-* Stata command or *--labelbook-*. Refer to Example 8 for further details.

## 2.12. Assess Variable Relevance

**Temporary, calculated or derived variables should not be disseminated**. Remove all unnecessary or temporary variables from the data files. These variables are not collected in the field and present no interest for users.

The data producer could generate variables that are only needed during the quality control process but are not relevant to the final data user. For example, the variable "merge" in Stata is generated automatically after performing the check described in the Numeral *1.6*, when the data producer wants to see if the datasets match properly. Variables that group categories of a question, dummy variables that identify a question's category are all variables produced during the coding process that are not relevant once the analysis is completed.

There are cases in which calculated variables may be useful to the users, so they must be documented in the metadata. For example, most Labor Force Surveys (LFS) contain derived

dummy variables to identify the sections of the population that are employed or unemployed. These variables are generated using multiple questions from the dataset and are essential elements of any LFS. Most data users prefer to make use of them instead of computing them on their own, to reduce the risk of error. This is a strong argument to make a case for keeping these variables in the dataset, despite them being a by-product of other original variables.

To be useful, those variables that remain in the dataset must be well documented, else they, they may be useless to or misunderstood by users.

## 2.13. Compress the Variables to Reduce the File Size

Compress the variables consist of reducing the size of the data file without loss of precision or modifying the information that it provides. Listed below are some reasons why compressing a data set may be a useful practice for at least three reasons: First, it makes faster the process of creating backups, uploading and downloading data files from your data repository or any microdata catalog. Second, it reduces the time that data users will need to spend working with the data. Additionally, it will make the data more accessible to the different type of users; sometimes the data size will impose restrictions on those users who lack high computational power. Third, it will help to free up disk space in the server where you store your data

### Example: Compressing Variables to Reduce File Size

*Table 15* shows two versions of one dataset that provides individual-level information about the year of the first union, age, school attendance, and health insurance. There is no difference in the appearance of both datasets. However, version 1 was saving uncompressed and version 2 compressed. In the uncompressed version, the variables "ID" and "Year" are stored as double, which means that they can store number with high decimal precision, but they are designed to only record information of integer numbers between -32,767 and 32,740. So, the compressed version changed the storage type of these variables to int and saves 6 bytes per observation. Similarly, other variables like "age" and "school attendance" are stored as a byte in the compressed version, which saves 7 bytes per observation when are compared to the uncompressed version. Let's suppose that one has a data set with 500 variables like these, the total savings would be 3,500 bytes per observation; if this data set has 50,000 observations, it means that the savings in memory space would be around 175 megabytes.

**Table 15. Compressing the Variables: Hypothetical data set**

Use the *compress* command in Stata, or the *compress* option when you save a SPSS data file.

## 2.14. Protect respondent privacy

Keep in mind that microdata are granular data with records describing individual units such as persons, households, businesses or institutions. Because these data contain detailed information about respondents, they may pose a risk of identification or divulging sensitive information if they are not properly protected. Steps need to be taken to ensure that the privacy of respondents is

protected. This is important to maintain public trust, meet ethical and legal obligations, and enable data to be shared and used responsibly for research and policy analysis.

Therefore all datasets intended to be used with automated tools, prepared for analysis, or released for dissemination must not contain direct identifiers or personally identifiable information (PII).

Before using or sharing a dataset, verify that all files have been reviewed to ensure that direct identifiers and other sensitive information that could directly or indirectly reveal the identity of respondents have been removed or treated. Examples include names, addresses, telephone numbers, email addresses, GPS coordinates, national identification numbers, and similar identifying information. Any variables containing direct identifiers should be excluded from datasets shared with others and from any datasets uploaded to online automated tools or external platforms.

If the dataset is intended for public release, it must first be transformed into an anonymous version suitable for dissemination. Removing direct identifiers is an essential first step in protecting respondent confidentiality and privacy. However, data anonymization should always begin with a careful review of the data to identify any variables that may pose a disclosure risk.

### Resources to Check for PII and Apply Statistical Disclosure Control[^3] Measures

- How to search datasets for PII

- How to deidentify datasets

- An anonymization practice guide

- Introduction to the theory of anonymization for microdata

- A guide to using the graphic user interface to sdcMicro

- Introduction to Statistical Disclosure Control (SDC)

::: tip Suggestion

If you are in the process of establishing a data archive and plan

to document a collection of microdata, undertake a full inventory of

all existing data and metadata before you start the documentation.

Use the IHSN Inventory Guidelines and Forms to facilitate this

inventory (available at www.ihsn.org).

:::

The next section focuses on organizing and preparing external resources for long-term preservation and, where appropriate, dissemination. These resources include all materials produced throughout the data lifecycle, not just the datasets themselves.

Examples include technical documentation, such as questionnaires, code lists, manuals, and methodological reports that are essential for data users; administrative and operational reports that may inform the design and implementation of future data collection projects; and supporting

materials, such as stakeholder feedback, workshop proceedings, and records of decisions made during questionnaire development. Preserving these resources alongside the data helps ensure transparency, reproducibility, and the long-term value of the microdata collection.

[^1]: See section 3 -- *Importing data and establishing relationships* for more information on key variables.

^2]: See [Variable Description section under Creating Structured Metadata.

[^3]: Statistical Disclosure Control (SDC) is the application of statistical and data modification techniques to datasets and statistical outputs to prevent the identification of individuals or organizations and the disclosure of confidential information, while maintaining the usefulness of the data for analysis.

---

# 3. Gathering and Preparing the Documentation

Documentation should make the data and the processes used to produce them verifiable. This includes recording methods, decisions, transformations, and known quality issues clearly enough for others to review.

All information related to the study throughout the data production lifecycle may be useful and should be archived (even if not all will be disseminated to the public). This includes not only technical documents such as the questionnaires or list of codes (obviously needed by data users), but also administrative reports (potentially useful for implementation of future microdata collection projects), and other documents such as a compilation of the comments provided by stakeholders at the time the questionnaire was designed, etc. All archived materials should follow a standardized folder structure and file naming convention to facilitate discovery, preservation, and future reuse. Resources to be included if available include:

**Administrative Documents and Governance Records**

- Project charters, budgets, and planning documents

- Key correspondence related to data collection, processing, or management

- Meeting minutes and decision records

- Steering committee or advisory group decisions

- Approval, ethical clearance, and authorization documents

- Data access, sharing, and licensing agreements

- Data governance policies and procedures

**Methodology and Data Production Documentation**

- Data collection or data acquisition methodology

- Data source descriptions and data lineage documentation (i.e. records of where data came from and the main steps through which they were collected, combined, transformed, or revised.)

- Sampling design documentation (where applicable)

- Weighting methodologies and calculation procedures (where applicable)

- Data integration, linkage, or matching procedures

- Data processing workflows and business rules

- Non-response adjustment procedures (where applicable)

- Fieldwork, operational, or system implementation reports

- Pilot, pre-test, or system testing reports

- Project timelines and production schedules

- Version history and change logs for data and metadata

- Records linking significant edits or transformations to the responsible script, decision, or approval.

**Geospatial Resources**

- GIS shapefiles and boundary files

- Geographic reference maps

- Spatial sampling or coverage maps

- Geographic code lists and crosswalks

- Geospatial metadata and documentation

**Data Collection and Source Materials**

- Questionnaires, forms, or data collection instruments

- Administrative forms and reporting templates

- Computer-assisted data collection instruments (CAPI/CATI/CAWI)

- Data submission templates and specifications

- Code lists, classifications, and controlled vocabularies

- Consent forms and participant information sheets (where applicable)

- Reference materials used during data collection

- User guides and operational instructions

**Data Processing and Quality Assurance Documentation**

- Data validation specifications and business rules

- Data entry and capture procedures

- Data transformation, integration, and processing workflows

- Data cleaning and editing logs

- Quality assurance and quality control reports

- Data consistency and validation checks

- Imputation methodologies

- Derived variable specifications and algorithms

- Version histories and change logs

- System testing and verification reports

**Access, Confidentiality, and Dissemination Documentation**

- Data access policies

- Terms of use

- Privacy and confidentiality procedures

- Disclosure risk assessments

- Anonymization and de-identification reports

- Data licensing information

- Access restrictions and distribution conditions

- Data sharing agreements

- Metadata publication and dissemination records

**Outputs, Analysis, and Communication Materials**

- Statistical reports and analytical publications

- Research papers, policy briefs, and working papers

- Dashboards and data visualizations

- Presentations and training materials

- Communication and outreach materials

- Technical notes and methodological papers

- Tables, indicators, and summary statistics

- Photos, videos, and other supporting media

- Documentation of data use cases and applications

**Computer programs**: (scripts used for data entry, editing, anonymization, tabulation and analysis)

::: tip Preserving Documents

Documents available in electronic format (MS-Word, Excel, and others) must be preserved in their original format and in PDF format.

All documents available only on hard copy/paper must be scanned. Use low resolution graphics, and black & white option (unless it is crucial to preserve colours e.g. where color conveys meaning or interpretation) to avoid large file sizes. A minimum scanning resolution of 300 dpi is recommended. Save the scanned documents in searchable PDF format where possible.

Maintain checksums or other file integrity verification mechanisms for digital preservation and periodically verify that files remain accessible and uncorrupted. Scan all resources with an updated virus detection application. Document the procedures used to maintain file integrity and accessibility over time.

For preservation copies, prefer formats with open, published specifications, wide adoption, good metadata support, and a reasonable history of backward compatibility. Retain the original file as

well. Maintain more than one copy of critical materials, keep at least one copy separate from the main storage location, and periodically test that files and restoration procedures still work.

:::

Organizing data and resources is a critical part of the documentation and archiving process. However, simply organizing files and documentation does not make them interoperable, machine-readable, or ready for ingestion into a data catalog. To ensure consistency, discoverability, and long-term usability, these resources must also be described using structured metadata that complies with established standards.

The next section introduces the metadata standards commonly used for documenting microdata and related resources, and explains how standards-compliant, machine-readable metadata can be created to support cataloging, discovery, preservation, and dissemination.

---

# 4. Completing the Metadata

Once all data and documentation materials have been assembled, organized, and verified, they should be documented in accordance with the relevant metadata standards: DDI Codebook (DDI-C) for datasets and Dublin Core Metadata Initiative (DCMI) standards for external resources. This process produces structured, machine-readable metadata that can be stored in XML (Extensible Markup Language) and JSON (JavaScript Object Notation) formats. Structured metadata supports the long-term preservation of data and documentation in repositories, enables metadata exchange and interoperability across catalogs, and enhances the discovery, access, and dissemination of data and related resources through searchable online platforms.

A thorough completion of the DDI-C and DCMI elements will significantly raise the value of the archiving work by providing users with the necessary information to put the study into its proper context and to understand its purpose.

## Completing the Study Documentation

The DDI-C metadata standard provides structured metadata for a dataset, capturing information on the identification, authorship, ownership, purpose, background methodologies, source information, provenance, quality control, access, physical file structures, variables/variable groupings, and related materials of a single dataset. Generating DDI-C compliant metadata requires the completion of this information organized into five key sections: Document Description, Study Description, File Description, Variable Description, Variable Groups, and External Resources.

The Metadata Editor is a specialized tool designed to create structured, machine-readable metadata that complies with internationally recognized metadata standards. This section introduces the key metadata elements required for creating comprehensive, structured metadata for microdata. It explains why such metadata is important and links to the Metadata Editor User Guide for detailed, step-by-step instructions for completing each section in the Metadata Editor.

### 4.1. Good Practices for Completing the Document Description

The Document Description section contains metadata about the XML or JSON metadata record itself rather than the study being documented. In other words, it describes the metadata document, its authorship, and its version history.

As a best practice, the Document Description should capture information such as:

- The metadata producer(s)

- Producer affiliation(s)

- Roles and responsibilities of the metadata contributors

- Date of metadata production

- Metadata version number

- DDI document identifier (DDI ID)

- Metadata creation and revision history

Maintaining this information supports version control, accountability, and traceability by clearly identifying who created the metadata, when it was produced, and which version of the metadata record is being used. This is particularly important when metadata are updated over time, maintained by multiple contributors, or exchanged between organizations. By documenting the metadata itself, users can assess its provenance, determine whether they are working with the most current version, and better understand how the metadata record has evolved over time.

The table below shows how this information might be filled in in the Metadata Editor.

| Field | Description | Example |

|---|---|---|

| **Metadata Producer** | Name of the person(s), team(s), or organization(s) responsible for creating, updating, or maintaining the metadata documentation. Use the **Role** attribute to distinguish different types of contributions to the metadata production process, such as documentation, quality review, editing, or metadata management. | **Example** <br>*Name:* Uganda Bureau of Statistics (UBOS) <br>*Role:* Metadata documentation <br><br>*Name:* IHSN <br>*Role:* Metadata review |

| **Date of Production** | The date (ISO format YYYY-MM-DD) on which the metadata document was created or last updated. This refers to the production of the metadata document itself, not the date on which the data were collected, published, disseminated, or archived. | |

| **Metadata Document Version** | Metadata development is often an iterative process. As errors are corrected, documentation is enhanced, or new information becomes available, updates to the metadata may be required. This element is used to identify and describe the current version of the metadata document. It is good practice to include a version number, date, and a brief description of any significant changes from previous versions. | **Example:** <br>*Version 02 (July 2026). This version is identical to Version 01, except that the Data Quality Assessment section has been updated and additional external resources have been documented.* |

| **Metadata Document ID Number (DDI ID)** | A metadata document should be assigned a unique identifier that distinguishes it from all other metadata records. Organizations should establish and consistently apply an identification scheme aligned with local repository, catalog, or archival practices. Where possible, the identifier should be linked to the collection's primary identifier. A recommended structure is: <br>
<br> PRODUCER*DDI*COUNTRY*YEAR*COLLECTION*VERSION* *<br><br>where: <br>- country is the 3-letter ISO country code (where applicable) <br>- producer is the abbreviation of the organization responsible for the metadata <br>- collection is a short identifier for the dataset, census, administrative system, registry, transaction system, or other microdata collection <br>- year is the reference year, reporting period, or collection start year <br>- version is the metadata document version number* | ***Example:*** *<br>The metadata document for a*

*Demographic and Health Survey prepared by the Uganda Bureau of Statistics in 2026 could be assigned the identifier: <br>* *UBOS*DDI*UGA*2026*DHS*v01 <br><br>*The metadata document for an Education Management Information System dataset could be assigned:* <br> MOE*DDI*KEN*2025*EMIS_v01 |

## 4.2. Good Practices for Completing the Study Description

A thorough completion of the DDI-C and DCMI elements will significantly raise the value of the archiving work by providing users with the necessary information to put the study into its proper context and to understand its purpose.

The DDI requires completion of the following sections: Document Description, Study Description, Datasets, Variables Groups, and External Resources. Recommendations for DDI-C fields included in the World Bank Group microdata template are provided below.

### Study Description

#### Title Statement

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Title | titl | Full authoritative title for the work.<br><br>**Guidance:** Do not include survey acronym in the title. Use title case and include reference years where appropriate. | Synthetic Data for an Imaginary Country, Sample, 2023 |

| Identification No | IDNo | Unique string or number.<br><br>**Guidance:** Use a consistent ID scheme.| **WLD*2023*SYNTH-SVY-EN*v01*M** |

|Other Identifiers| identifiers |This metadata element is provided to store information on the other identifiers of the study (for example a Digital Object Identifier, or the study identifier in another data catalog). The identifiers (key) and is composed of two sub-elements ( type and identifier ). This is a repeateble field that allows capturing multiple identifiers.|

| Sub-title | subTitl | Secondary title.<br><br>**Guidance:** Optional and rarely used. | A synthetic hierarchical dataset for simulation and training purposes |

| Alternate Title / Acronym | altTitl | Commonly used title or acronym.<br><br>**Guidance:** Use official abbreviation. | **SYNTH-SVY-EN 2023** for the synthetic file or **DHS 2015** for a 2015 DHS survey |

| Parallel / Translated Title | parTitl | Title translated into another language. | Données synthétiques pour un pays imaginaire, échantillon, 2023 |

#### Authoring & Contributors

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Authoring Entity / Primary Investigator | AuthEnty | Person or agency responsible for substantive content.<br><br>**Guidance:** Include institution and affiliation. | National Statistics Office |

| Other Identifications / Acknowledgments | othId | Collaborators and contributors not otherwise recorded. | |

#### Production Statement

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Producer | producer | Person or organization responsible for production. | Technical assistance in questionnaire design |

| Copyright | copyright | Usage rights for the resource. Provide copyright or usage rights when relevant.| © 2017, Popstan Central Statistics Agency |

| Production Date | prodDate | Date when the marked-up document/marked-up document source/data collection/other material(s) were produced (not distributed or archived).<br><br>Track production dates for all versions. Provide at least month and year in ISO format YYYY-MM or YYYY-MM-DD.| 2023-05-01 |

| Production Place | prodPlac | Production location. | World Bank |

| Funding Agency / Sponsor | fundAg |The source(s) of funds for production of the work including abbreviation and affiliation.<br><br>List organizations that contributed cash or in-kind financing. Include government funding sources where applicable. | UNHCR-World Bank Joint Data Center on Forced Displacement |

| Grant Number | grantNo | Grant or contract number. | KP-P174174-TF0B5124 |

#### Distribution Statement

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Contact Persons | contact | Responsible contact persons. | dataproc@cso.org |

| Depositor | depositr | Institution depositing the collection. | World Bank |

| Date of Deposit | depDate | The date that the work was deposited with the archive that originally received it. Record the date the data collection was deposited with the receiving archive using ISO date format YYYY-MM-DD.| 2023-02-02 |

#### Series Statement

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Series Name | serName | The name of the series to which the work belongs. | |

| Series Information | serInfo | Contains a history of the series and a summary of those features that apply to the series as a whole. | This dataset is part of a collection of fully synthetic data generated, for training and simulation purposes, for an imaginary middle-income country. The dataset is available in English and French. A full population dataset (~10 million individuals) is also available in English and French as a "synthetic census dataset". |

#### Version

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Version | version | Also known as release or edition. Use a standard convention for dataset releases or editions. | v02 |

| Version Description | verStmt |Version statement for the work at the appropriate level: marked-up document; marked-up document source; study; study description, other material; other material for study.<br><br>Use a version

number followed by a version label. Distinguish raw, edited internal-use, and edited dissemination versions where applicable. | v00: Basic raw data, obtained from data entry; v01: Edited data for internal use only; v02: Edited, anonymous dataset for public distribution|

| Version Responsibility | verResp | The organization or person responsible for the version of the work.| World Bank, Development Data Group |

#### Authorization

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Authorizing Agency | authorizingAgency | Name of the agent or agency that authorized the study. | Office of Health Statistics |

| Authorization Statement | authorizationStatement | The text of the authorization. | Authorization received on 2010-11-04 |

---

## Study Information

#### Overview & Coverage

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Keywords | keyword | Words or phrases that describe salient aspects of a data collection's content. Use keywords to summarize content or subject matter. Prefer standard thesauri where available. | |

| Topic Classifications | topcClas | The classification field indicates the broad substantive topic(s) that the data cover. Select topics from a standard thesaurus where possible. Topic classifications support search and catalog discovery. | |

| Subjects / Scope | subject | Subject information describing the data collection's intellectual content. Describe the themes covered by the study. The scope should summarize modules or themes, not geographic coverage. | HOUSEHOLD: household characteristics, education, water and sanitation; WOMEN: maternal health, marriage, contraception; CHILDREN: birth registration, immunization, anthropometry. |

| Abstract | abstract | Provide a clear summary of the purpose, objectives, and content. Ideally written by a researcher or statistician familiar with the study. | The dataset is a relational dataset of 8,000 households households, representing a sample of the population of an imaginary middle-income country. The dataset contains two data files: one with variables at the household level, the other one with variables at the individual level. It includes variables that are typically collected in population censuses (demography, education, occupation, dwelling characteristics, fertility, mortality, and migration) and in household surveys (household expenditure, anthropometric data for children, assets ownership). The data only includes ordinary households (no community households). The dataset was created using REaLTabFormer, a model that leverages deep learning methods. The dataset was created for the purpose of training and simulation and is not intended to be representative of any specific country. |

| Time Period | timePrd | The time period to which the data refer. This differs from data collection dates. | |

| Dates of Data Collection | collDate | Contains the date(s) when the data were collected. Enter start and end dates in ISO format YYYY-MM-DD. If collection

occurred in waves, enter each wave separately and identify the cycle. | 2006-09-02 to 2006-10-17 |

| Country | nation | Indicates the country or countries (or “economies”, or “territories”) covered in the study (but not the sub-national geographic areas). If the study covers more than one country, they will be entered separately. Consistent country names and spelling should be used across an organization (for example, if "Democratic Republic of Congo" is the recommended name for the organization, the information should should not be entered as "Congo, Democratic Republic", "Congo, Dem. Rep.", "DR Congo", or other option). In user templates, a controlled vocabulary can be set at the country name level, at the country code level, or at the combined name/code level. | Uganda (UGA) |

| Geographic Coverage | geogCover | Geographic representativeness. Describe the geographic level at which the data are representative, not simply where data were collected. | National coverage |

| Geographic Unit | geogUnit | Lowest level of geographic aggregation covered by the data. Enter the lowest level of geographic aggregation covered by the data. | District |

| Geographic Bounding Box | geoBndBox | The fundamental geometric description for any dataset that models geography. Provide the minimum bounding box, using west/east longitudes and south/north latitudes. Required if a geographic bounding polygon is included. | |

| Geographic Bounding Polygon | boundPoly | This field allows the creation of multiple polygons to describe in a more detailed manner the geographic area covered by the dataset. Use polygons only to define outer boundaries of the covered area. This supports coordinate-based discovery and is not intended for detailed mapping.| Separate polygons may be used for non-contiguous areas such as islands or exclaves.|

| Unit of Analysis | anlyUnit | Basic unit of analysis or observation that the file describes: individuals, families/households, groups, institutions/organizations, administrative units, etc. | Enter standard units of analysis where applicable, such as household, person, enterprise, commodity, or plot of land. |

| Universe | universe | The universe is the group of persons (or other units of observations, like dwellings, facilities, or other) that are the object of the

study and to which any analytic results refer. The universe will rarely cover the entire population of the country. Sample household surveys, for example, may not cover homeless, nomads, diplomats, community households. Population censuses do not cover diplomats. Facility surveys may be limited to facilities of a certain type (e.g., public schools). Try to provide the most detailed information possible on the population covered by the survey/census, focusing on excluded categories of the population.<br><br>For household surveys, age, nationality, and residence commonly help to delineate a given universe, but any of a number of factors may be involved, such as sex, race, income, veteran status, criminal convictions, etc. In general, it should be possible to tell from the description of the universe whether a given individual or element (hypothetical or real) is a member of the population under study. Describe the population of interest in detail, including covered and excluded groups. This is the study-level universe, not variable-level skip patterns. | All de jure household members |

| Kind of Data | dataKind | The type of data included in the file: survey data, census/enumeration data, aggregate data, clinical data, event/transaction data, program source code, machine-readable text, administrative records data, experimental data, psychological test, textual data, coded textual, coded documents, time budget diaries, observation data/ratings, process-produced data, etc. Select the broad classification of the data from the controlled vocabulary. | Sample survey data |

---

## Study Development

#### Development Activities

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Development Activity | developmentActivity | Information on the development activity including a "description" of the activity, name of each "participant", each "resource" used, and each "outcome" of the activity. | |

| Study Development | studyDevelopment | Describe the process of study development as a series of development activities. | |

| Study Budget | studyBudget | Describe the project budget in as much detail as needed. Include budget line ID, label, amount, currency, and source of funding where available. | |

---

## Methodology

#### Methods

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Time Method | timeMeth | The time method or time dimension of the data collection. | |

| Data Collectors | dataCollector | The entity (individual, agency, or institution) responsible for administering the questionnaire or interview or compiling the data. Record agencies or persons responsible for data collection. Include name, abbreviation, affiliation, and role. Usually record the agency rather than individual interviewers. | Name: CSO; Affiliation: Ministry of Planning; Role: Planner |

| Collector Training | collectorTraining | Describe interviewer training, process testing, and standards compliance. Use type to distinguish training aspects when needed. | Two-week interviewer training |

| Frequency of Data Collection | frequenc | For data collected at more than one point in time, the frequency with which the data were collected. Describe the frequency of repeated data collection; use controlled vocabulary where available. | annual |

| Sampling Procedure | sampProc |The type of sample and sample design used to select the survey respondents to represent the population. Include sample size, selection process, stratification, stages of selection, design omissions, level of representation, non-response strategy, sample frame, and variables identifying strata and PSU where relevant. Reference detailed sampling documents provided as external resources. | 8000 households selected |

| Sample Frame | sampleFrame |Sample frame describes the sampling frame used for identifying the population from which the sample was taken. Describe the sample frame, including the listing exercise, validity period, custodian, update procedure, frame unit, reference period, and use statement where available. | |

| Sample Frame Name | sampleFrameName | Provide the name of the sample frame used to identify the population from which the sample was drawn. | WLD*2023*SYNTH-CENS-EN*v01*M |

| Target Sample Size | targetSampleSize | Provides both the target size of the sample (this is the number in the original sample, not the number of respondents) as well as the formula used for determining the sample size. Include the formula used to determine the sample size where available. | |

| Sample Size | sampleSize | Provide planned and actual sample sizes with the unit and number. | 8000 households |

| Deviation from Sample Design | deviat | Sometimes the reality of the field requires a deviation from the sampling design (for example due to difficulty to access to zones due to weather problems, political instability, etc). If for any reason, the sample design has deviated, this can be reported here. This element will provide information indicating the correspondence as well as the possible discrepancies between the sampled units (obtained) and available statistics for the population (age, sex-ratio, marital status, etc.) as a whole.<br><br>Report deviations from sample design caused by field constraints or other implementation realities. | |

| Mode of Data Collection | collMode | The method used to collect the data; instrumentation characteristics. Select the manner in which information was gathered from a controlled vocabulary when available. | Face-to-face interview |

| Research Instrument / Questionnaires | resInstru | The type of data collection instrument used. Describe questionnaires or instruments used. List each instrument, language, design process, stakeholder review, pretest, and reference external resources. | Household, women, and children questionnaires were adapted from MICS3 model questionnaires and translated into local languages. |

| Instrument Development | instrumentDevelopment | Describe any development work on the data collection instrument. Describe development work, review processes, pilots, testing, and agencies or persons consulted. | |

| Notes on Data Collection | collSitu | Description of noteworthy aspects of the data collection situation. Document noteworthy events, pilot/pre-test details, interview duration, field arrangements, languages used, and corrective actions during collection. | The pre-test took place from August 15-25, 2006 and included 14 interviewers who later became supervisors. |

| Actions to Minimize Losses | actMin | Document actions taken to minimize data loss. | |

| Weighting | weight | This field only applies to sample surveys. The use of sampling procedures may make it necessary to apply weights to produce accurate statistical results. Describe here the criteria for using weights in analysis of a collection, and provide a list of variables used as weighting coefficient. If more than one variable is a weighting variable, describe how these variables differ from each other and what the purpose of each one of them is. List weighting variables and explain how they differ and how each should be used. Describe calculation, non-response adjustment, and normalization where available. | HHWEIGHT |

| Data Processing | dataProcessing | Describe data entry, editing, verification, recoding, tabulation, software, double entry, structural checks, and analysis processing. Reference processing guidelines or programs as external resources. | Describes various data processing procedures not captured elsewhere in the documentation, such as topcoding, recoding, suppression, tabulation, etc. Document data editing included office editing and coding, data entry, structure checking, verification entry, secondary editing, export to SPSS, recoding, adding weights, and data quality tabulations. |

| Coding Instructions | codingInstructions | Describe specific coding instructions used in data processing, cleaning, acquisition, or tabulation. Provide human-readable coding instructions and command code where relevant; identify formal language for commands. | RECODE V1 TO V100 (10 THROUGH HIGH = 0).|

---

## Quality Assessment

#### Quality Statement

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Quality Statement | qualityStatement |This structure consists of two parts, "standardsCompliance" and "otherQualityStatements". List standards followed during study execution. Include standard name, producer, and how the study complied. Use Other Quality Statement for additional quality notes. | Study complied with specified survey methodology and data documentation standards. |

| Standards Compliance | qualityStatement.standards | List the standard name, producer, and how the study complied with it. | Standard: DHS Program Methodological Standards; Producer: DHS Program|

| Other Quality Statement | qualityStatement.otherQualityStatement | Use for quality explanations that do not fit standards compliance. Data collection followed documented supervision, editing, and validation procedures. | |

#### Data Appraisal

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Other Forms of Data Appraisal | dataAppr |Document other issues pertaining to data appraisal. Describe other actions taken to assess reliability or quality, including post-enumeration surveys, comparisons with other sources, and data quality tables. Document both the procedures that were planned and any important deviations in implementation. Explain known limitations and their implications so users can judge whether the data are suitable for their intended analysis.| Data quality tables reviewed age distributions, missing values, sex ratios at birth, population pyramids, and anthropometry scatter plots. |

| Response Rate | respRate | The percentage of sample members who provided information. <br><br>Report household or unit response rates based on the original sample. Provide rates by stratum when possible and ensure consistency with sample size and records in the data. | Household response rate: 96.3%; women response rate: 96.0%; children response rate: 97.7%. This is a synthetic dataset; the "response rate" is 100%. |

| Estimate of Sampling Error | EstSmpErr | Measure of how precisely one can estimate a population value from a given sample. For sample surveys, describe sampling error calculations, indicators, software, methods, and reports or programs provided as external resources. | Sampling errors were calculated using the SPSS Complex Samples module and Taylor linearization method. |

#### Post-Evaluation Procedures

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Evaluator | ex*post*evaluation.evaluator | The "Evaluator" element identifies the person(s) and/or organization(s) involved in the ex-post evaluation. Include name, affiliation, abbreviation, and role. | Affiliation: United Nations; Abbr.: UNSD; Role: Consultant |

| Evaluation Process | ex*post*evaluation.evaluation_process | A description of the ex-post evaluation process. This may include information on the period when the evaluation was conducted, cost/budget, relevance, institutional or legal arrangements, etc. Describe the post-evaluation process, especially when the study is repeated or ongoing. | Independent review after completion of field operations. |

| Evaluation Outcomes | ex*post*evaluation.outcomes | A description of the outcomes of the ex-post evaluation. It may include a reference to an evaluation report. Summarize evaluation findings and recommendations. | Recommendations were adopted for the next survey round. |

#### Data Quality Controls

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Cleaning / Control Operations | cleanOps | Methods used to "clean" the data collection, e.g., consistency checking, wild code checking, etc. Describe data cleaning methods such as consistency checks, wild-code checks, supervision, field controls, and corrective actions. | Field editors reviewed questionnaires daily for missed questions, skip errors, and inconsistencies. |

| Processing Checks | ProcStat | Processing status of the data collection. Document file-specific checks such as consistency checking and wild-code

checking. Reference external resources with check specifications where available. | Range checks and consistency checks applied to file-level variables. |

| Missing Data | dataMsng | This element can be used to give general information about missing data, e.g., that missing data have been standardized across the collection, missing data are present because of merging, etc. Document missing data conventions. Missing values should be coded and distinguished from not applicable and zero values. | Codes 98 = Do not know; 99 = Missing. |

| Imputation | imputation | Imputation or replacement techniques used to correct inconsistent or unreasonable data. Summarize imputation methods and reference external resources where available. | Random imputation was used for missing dates within calculated constraints. |

#### Fieldwork Quality Assurance

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Supervision | control_operations | This element contains information on the oversight of the data collection, i.e. on methods implemented to facilitate data control usually performed by the primary investigator. Describe team structure, supervisor roles, field editor responsibilities, central office visits, and corrective actions. | Interviewing teams included supervisors and field editors; central staff conducted periodic field visits. |

| Compliance with Data Collection Standards | qualityStatement.standards | Compliance with relevant data collection standards. Describe whether the study complied with international survey recommendations or institutional fieldwork standards. | Survey procedures complied with documented international recommendations. |

---

## Data Access

This section describes access conditions and terms of use for the data collection.

#### Access & Use

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Location of Data Collection | accsPlac | Storage location. Provide the location where the data are stored and a URL or address if available. | |

| Availability Status | avlStatus | Availability statement. Indicate availability status such as unavailable, embargoed, superseded, or new edition imminent where relevant. | |

| Confidentiality Declaration | confDec | This element is used to determine if signing of a confidentiality declaration is needed to access a resource. <br> <br> **Guidance:** Indicate confidentiality declarations required for access when relevant. Leave blank if no confidentiality issue applies. | Users agree not to attempt to identify any person, establishment, or sampling unit. |

| Citation Requirement | citReq | Text of requirement that a data collection should be cited properly in articles or other publications that are based on analysis of the data. <br><br> **Guidance:** Provide the citation users must use. Include primary investigator, dataset name and abbreviation, reference year, version, and official access URL where available. | World Bank. (2023). *Synthetic Data for an Imaginary Country, Sample*, 2023 [Dataset]. World Bank, Development Data Group. https://doi.org/10.48529/MC1F-QH23. |

| Deposit Requirement | deposReq | User reporting requirements. | |

| Access Conditions | conditions | Conditions of use. <br><br>**Guidance:** Indicates any additional information that will assist the user in understanding the access and use conditions of the data collection. Note that if this information is subject to change, it should not be captured in the study metadata but provided in a data catalog page. | Public Use Dataset |

| License | license | License information. A legal document giving official permission to something with the resource. | |

| Embargo | embargo | Embargo information. Provides information on variables/nCubes which are not currently available because of policies established by the principal investigators and/or data producers. | |

| Disclaimer | disclaimer | Disclaimer statement.<br><br> **Guidance:** A disclaimer limits the liability that the data producer or data custodian has regarding the use of the data. A standard legal statement should be used for all datasets from a same agency. | The following formulation could be used:<br><br> The user of the data acknowledges that the original collector of the data, the authorized distributor of the data, and the relevant funding agency bear no responsibility for use of the data or for interpretations or inferences based upon such uses. |

| Metadata Access | metadataAccs | Metadata access conditions.<br><br>**Guidance:** This section describes access conditions and terms of use for the metadata. | |

---

For a full list of DDI-C fields, see the DDI XML Schema and the Metadata Editor Documentation. Field level examples are also provided in the Metadata Editor.

Much of this information can be obtained from reports, technical documentation, methodological notes, project proposals, data collection instruments, user and operational manuals, system specifications, data dictionaries, and other supporting materials created throughout the lifecycle of the microdata collection. These resources often contain the information needed to accurately document the data, its production process, and its intended use.

**Why is this important?**

**The Study Description serves as the foundation of the metadata record.**

It helps users determine whether a dataset is relevant for their research, understand its strengths and limitations, and assess whether it is appropriate for a particular analytical purpose.

A well-documented Study Description captures above-mentioned metadata elements as well as information on data quality, comparability, and limitations. As a result, the comprehensive metadata supports discovery and evaluation of datasets in data catalogs and preserves institutional knowledge for future users.

**Supports Data Quality and Transparency**

The Study Description is also a critical component of data quality documentation. By describing how the microdata were collected, acquired, processed and managed, it enables users to assess the reliability, coverage, and suitability of the data for their intended purposes.

For example, the Study Description may document the following elements that would shed light on quality:

- Data sources, collection methods, or data acquisition processes

- Coverage, scope, target population, or units of observation

- Sampling methods, sample size, or selection procedures (where applicable)

- Response rates, completeness measures, and known sources of error or bias

- Data collection, extraction, integration, or processing procedures

- Quality assurance, validation, and data verification processes

- Data transformations, harmonization, imputation, or adjustment methodologies

- Changes from previous versions, collection cycles, or releases

- Limitations, constraints, and considerations for data interpretation and use

Without adequate documentation, users may incorrectly interpret findings, make invalid comparisons, or draw conclusions that are inconsistent with the data's scope, context, and methodological limitations.

**Supports Automated Discovery and Analysis**

As data catalogs increasingly incorporate automated search and analytical tools, detailed study-level metadata becomes even more important. Automated systems rely on information in the Study Description to understand the subject matter, population, methodology, and coverage of a dataset. This enables them to make relevant and meaningful recommendations, improve search results and dataset discovery.

**Example**

Consider a dataset titled *Household Survey 2025.* From the title alone, it is impossible to know what the survey was about, who was interviewed, where it was conducted, or when the data were collected. The Study Description provides this information by documenting the survey objectives, target population, geographic coverage, data collection methods, and reference period. This context enables users to correctly interpret the data and assess whether it is suitable for their research needs.

For detailed guidance on how to complete the Study Description, see the section on *Documenting microdata in the Metadata Editor*.

### 4.3. Good Practices for Completing the File Description

The File Description section captures metadata about the physical data file(s) that make up a dataset. Recommendations for completing the file description metadata are provided in the table below.

## File Description

| Field | DDI Element | Description | Example |

|---|---|---|---|

| File Description | fileDscr | This section stores metadata about the physical data files that make up a research collection.| |

| File Name | fileName | The official name of the data file.| hh_data.dta |

| File Contents | fileCont |Details what the data file contains, including its nature and any special characteristics. | Synthetic data, household-level variables, sample of 8,000 households |

| Missing Data | dataMsng | Information on missing data conventions. | |

| Data Check | dataChck | A global element used at the file level to document the extent of processing checks and data cleaning operations performed on a data file| Consistency checks were performed by Data Producer.|

| File Type | fileType | The technical format of the file (e.g., SAS, SPSS, Stata, CSV). | |

| Version statement | verStmt | Records the specific version/edition of the physical data file being described. Version statement captures information onversion number version , version date versionDate , and notes .| |

|Dimensions| dimensns |Physical dimensions of the file, such as the total number of cases/rows caseQnty and columns varQnty .| |

| Processing Status | procStat |Information regarding the processing status of the data file (e.g., whether it has been cleaned or weighted).||

| Notes | notes | Any notes required to understand and use the file correctly including how to merge with other datasets. | |

---

Some information, such as file names, record counts, variable counts, and basic file characteristics, may be extracted automatically from the data files when using the Metadata Editor. However, other details, including quality assurance procedures, data processing and transformation activities, validation checks, and the treatment of missing or inconsistent data, may need to be obtained from technical documentation, methodological reports, project records, or consultations with data producers and subject matter experts.

**Why is this important?**

File-level metadata provides essential context about how a dataset was created, processed, and structured. It helps users understand the contents and quality of the data files and supports the correct interpretation and use of the data.

Well-documented file descriptions:

- **Help users identify the correct data file and version** for their analysis.

- Provide **transparency** about data processing, validation, quality assurance procedures, known limitations, exclusions, or caveats.

- Document **how missing values, inconsistencies, and errors were handled**.

- **Support reproducibility** by recording details that may not be apparent from the data alone.

- Facilitate **long-term preservation and reuse** of the dataset.

- Improve **discoverability and usability** in data catalogs and repositories.

- **Enable automated search and analysis tools** to better understand the structure and characteristics of the data.

The File Description section is also an important component of data quality documentation. Information about missing data, validation rules, consistency checks, editing procedures, and quality assessments **helps users evaluate the reliability and limitations** of the dataset.

Without this information, users may incorrectly assume that the data are complete or error-free, potentially leading to inaccurate analyses and conclusions.

**Example**

A dataset contains 10,000 records and 250 variables. The file description may indicate that:

- Records with duplicate identifiers were removed during processing.

- Age values outside the valid range were reviewed and corrected.

- Missing income values were coded as -99.

- Consistency checks were performed between household and individual records.

- The file represents the second corrected release of the dataset.

This information provides important context that cannot be determined simply by looking at the data file itself and helps ensure that users understand how the final dataset was produced and what limitations should be considered during analysis.

## 4.4. Good Practices for Completing the Variable Description

The DDI-Codebook Variable Description section captures detailed metadata about each variable in a dataset. The table below provides some recommendations for completing variable metadata. when using the Metadata Editor, variable information like name, type, intrvl, label, value labels, width and summary statistics are automatically extracted from the data.

#### Variable Description

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Variable | var | Identifies and documents a single variable in a data file. The variable element contains information about the variable's name, label, format, location, question text, valid values, categories, summary statistics, and other variable-level metadata used to describe and interpret the data. See the data preparation section for guidance on variable naming. | name : hhid varFormat type : character <br> intrvl : discrete<br><br> name : hhinc <br> varFormat type : numeric <br> intrvl : continuous|

| Variable Label | labl | All variables should have a label that provides a short but clear indication of what the variable contains. Ideally, all variables in a data file will have a unique label. File formats like Stata or SPSS often contain variable labels. Variable labels should identify the questionnaire item, clearly describe the content, and indicate whether variables are constructed. Avoid ALL CAPS and duplicate labels. | Annual per capita real expenditure in local currency |

| Variable Format / Data Type | varFormat | Data type. | Numeric |

| Pre-question Text | preQTxt | The pre-question texts are the instructions provided to the interviewers and printed in the questionnaire before the literal question. This does not apply to all variables. Do not confuse this with instructions provided in the interviewer’s manual. Copy the text printed in the questionnaire before the literal question, not interviewer manual instructions.| Check age |

| Literal Question | qstnLit | Copy the exact question wording as asked during the interview. | Does (name) attend any organized learning or early childhood education programme? |

| Post-question Text | postQTxt | Text describing what occurs after the literal question has been asked.| Go to next module |

| Interviewer Instruction | ivuInstr | Specific instructions to the individual conducting an interview. Copy interviewer manual instructions. Repeat instructions across variables when the same guidance applies to multiple variables. | |

| Categories | catgry | A description of the categories (codes with labels) that apply to the variable, for categorical variables. The element also includes summary statistics at the category level.| 1=Male;2=Female |

| Category Value | catValu | The explicit response option for a category. | |

| Category Statistics | catStat | May include frequencies, percentages, or crosstabulation results. | |

| Category Level | catLevel |Used to describe the levels of the category hierarchy.| |

| Recoding and Derivation | derivation | Used only in the case of a derived variable, this element provides both a description of how the derivation was performed and the command used to generate the derived variable, as well as a specification of the other variables in the study used to generate the derivation. For derived variables, document source variables, transformations, assumptions, and reference programs or reports when calculations are complex. | AGE_GRP was obtained by recoding S1Q3 age in years into age groups 0-4, 5-9, etc. |

| Coding Instructions | codInstr | Any special instructions to those who converted information from one form to another for a particular variable. Record special coding instructions associated with a variable. | Due to a system error, value 27 for NBWFBPC should be recoded as invalid value 99. |

| Summary Statistics | sumStat | One or more statistical measures that describe the responses to a particular variable and may include one or more standard summaries, e.g., minimum and maximum values, median, mode, etc. Select appropriate statistics based on measure. Avoid frequencies for ID variables and avoid means/standard deviations for nominal codes. | Minimum, maximum, mean, standard deviation, frequency|

Comprehensive variable documentation is essential because

- It **provides the context required to accurately interpret, analyze, and reuse data**. Variable descriptions preserve the meaning, origin, and construction of each variable, ensuring that datasets remain understandable long after data collection has been completed.

- They also **improve discoverability in data catalogs and support reproducible research** by documenting how variables were collected, coded, and, where applicable, derived.

- Moreover, **modern automated catalogs, search engines, and data assistants rely heavily on metadata** to find relevant datasets and variables. Without this information, automated techniques may apply inappropriate analytical methods or draw incorrect conclusions.

Beyond supporting interpretation and reuse, **documenting variables contributes to data quality**. The process of reviewing and describing variables often helps identify inconsistencies, coding errors, undocumented transformations, unclear labels, and discrepancies between the questionnaire and the dataset. As a result, variable documentation serves as an important quality assurance step, improving the accuracy, reliability, and usability of the data.

Variable descriptions help explain:

- what the variable measures

- how the information was collected

- the population to which the question applies

- any instructions provided to interviewers or respondents

- how response categories and codes should be interpreted

- how derived or calculated variables were constructed.

**Example**

A variable named *AGE* may appear straightforward, but the accompanying description may clarify whether it refers to age at the time of the interview, age at last birthday, age at first marriage, or age at a specific reference date. Without this additional context, users may incorrectly interpret the data and produce misleading results.

> **Note:** When using a tool such as the Metadata Editor, metadata elements including variable names, variable labels, category labels, data types, field lengths, and field widths can be automatically extracted from a dataset during import.<br>Additional metadata, such as field definitions, descriptions, permissible values, coverage statements, data collection or processing notes, usage instructions, and information about derived or calculated variables, should be completed using available documentation, including questionnaires, data dictionaries, technical reports, methodological documentation, field manuals, processing specifications, or other materials that describe how the data were collected, managed, transformed, and produced.<br>To improve the quality and usability of the documentation, ensure that variables and response categories are assigned clear, meaningful labels and that variable names follow established naming conventions. Variable names should be concise, descriptive, and free of special or unsupported characters that may not be accepted by commonly used statistical and data analysis software. Well-structured variable names and labels make datasets easier to understand, analyze, and maintain over time.<br>While variable metadata can be updated in

> the Metadata Editor, we recommend applying variable naming and labeling standards before importing the dataset whenever possible. See the section on Gathering and Preparing the Dataset for additional guidance.

### 4.5. Creating Variable Groups

The Variable Groups section is optional and is used to organize variables into logical, thematic, or hierarchical groupings without altering the underlying data files. These groups are virtual, they exist only within the metadata and do not affect the structure or content of the dataset itself.

Variable groups help organize large and complex datasets into meaningful sections, making them easier for both humans and systems to navigate, understand, and discover. By grouping related variables together, users can quickly identify relevant content without having to review hundreds or thousands of individual variables.

**Why is this important?**

Variable groups improve the usability and discoverability of a dataset by providing a structured view of its contents. They help users understand how variables relate to one another and make it easier to locate information on specific topics.

Well-designed variable groups can:

- Improve navigation of large datasets
- Help users quickly identify variables relevant to their research
- Support thematic browsing and searching in data catalogs
- Provide additional context about the organization of the questionnaire or study
- Enhance machine-readable metadata for automated search and discovery tools

**Example**

Variable groups may mirror the structure of the questionnaire or be organized around key analytical themes. For example, variables could be grouped into categories such as:

- Demographics
- Education and Literacy
- Employment and Income
- Health and Nutrition
- Information and Communication Technology (ICT)
- Internet Use
- Housing Characteristics
- Agriculture
- Household Assets

Variable groups make metadata easier to explore. Datasets become more accessible to users, and data catalogs can provide a richer and more intuitive discovery experience.

**In a Nutshell**

Together, the Document Description, Study Description, File Description, Variable Description, and Variable Groups sections provide a comprehensive and structured description of a dataset and its supporting documentation. Following metadata standards when capturing information about the study, data files, individual variables, and the relationships between them, promotes greater interoperability, comparability, and consistency across datasets, systems, and repositories.

These metadata elements improve data quality by encouraging a systematic review of the data and its documentation, while also supporting long-term preservation, discoverability, and reuse. Rich metadata ensures that datasets remain understandable and usable long after the original project team is no longer available, preserving valuable institutional knowledge and context.

Comprehensive metadata also enables data catalogs, repositories, and automated tools to more effectively locate, interpret, connect, and analyze data resources. The Metadata Editor streamlines this process by providing an efficient way to create standards-compliant, machine-readable metadata that can be shared, published, exchanged, and preserved over time.

::: tip Documenting microdata in the Metadata Editor

To generate structured metadata for datasets using the Metadata Editor, you will need to:

- ✓ create a Metadata Editor Project

- ✓ complete the Document Description metadata

- ✓ complete the Study Description metadata

- ✓ import the prepared data file(s) to your project

- ✓ complete the File Description for each imported file

- ✓ complete the Variable Description metadata and (optional) create Variable Groups

- ✓ complete the External Resource metadata for each resource

- ✓ review and export the machine-readable metadata (ready for sharing/publication in online catalogs).

Detailed instructions on how to create a project in the Metadata Editor are provided here.

Note that if you are documenting a population census and have very large data files, it is recommended to split the files by geographic area. Typically, you will have a file at individual level, one at the household level, and possibly one at the community level, for each State or Province. In such case, import all files for one State or Province only. You will import the other data files after you complete the documentation of the files. This will considerably reduce the time needed to save your files. The Metadata Editor will allow you to replicate the metadata from the documented files to all other data files that you will import later.

For detailed guidance on how to document a study in the Metadata Editor, see the Documenting Microdata section of the Metadata Editor User Guide.

Detailed information and examples on how to import data are provided in the Metadata Editor guide - see Documenting Data → Microdata → Import and Document the Dataset → Data Files.

:::

It is equally important to document the supporting documents and resources that enable data users to better understand the data, replicate analyses, and correctly interpret results. These supporting materials, often referred to as external resources, are documented using the Dublin Core Metadata Initiative (DCMI) standard. Documenting these resources alongside the data helps organizations to improve transparency, reproducibility, and the long-term usability of the micro-dataset collection.

The next section explains how to document external resources using the DCMI standard and describes the metadata elements used to capture and manage these materials.

---

# Creating Variable Groups

Variable groups are optional, but will help organize the data for the user into specific subject of use categories. This will be particularly useful to the user in the case of data files that contain many variables and are not organized by topic (some flat files contain hundreds or even thousands of variables).

The Metadata Editor allows you to group variables found in various separate data files. For example, education data may be found in various locations and the disparate variables grouped together. Also, a same variable can belong to more than one group.

Variable groups are "virtual". The variables themselves are not moved or grouped. They remain untouched in the data files.

The variable groups will appear under a menu item "Data dictionary". The only reason for grouping variables is to allow users to easily locate variables related to their topic of interest. If your dataset contains very few variables, there is no justification for grouping them.

If you decide to create variable groups (and sub-groups if needed), make sure that ALL variables in the dataset belong to at least one group.

Variable groups also have their own DDI elements which include Type, Label, Text, Definition, Universe, and Notes. These elements are optional and will in most cases be left empty.

| Type | This is a controlled vocabulary<br>field. It best identifies the<br>manner the variables are grouped<br>together. This field is optional. |

| --- | --- |

| Label | The label used to identify the<br>group should be clear and relate<br>to the type chosen. If these are<br>grouped by subject, then the<br>subject should be clearly stated<br>etc. |

| Text | Include additional text to<br>clarify the reason or purpose for<br>grouping the variables. This<br>field is optional. |

| Definition | This optional field is used to<br>define the variable group. |

| Universe | This optional field defines the<br>universe relevant to the selected<br>grouped variables. The variables<br>for example can be grouped as<br>“Fertility Data” and the universe<br>restricted to women between the<br>ages of 15-49. |

| Notes | Additional space for further<br>optional explanatory notes. |

For more information on creating variable groups in the Metadata Editor, see the Documenting Data → Microdata → Import and document the dataset → Variable groupings section of the documentation.

# 5. Documenting External Resources

## Good Practices for Completing External Resource Metadata

The External Resources section is used to document and describe materials that are related to a dataset, study, project, or data collection but exist outside the primary data files. These resources may include questionnaires, data collection instruments, interviewer or field manuals, reports, publications, methodological documents, technical papers, tabulations, maps, training materials, codebooks, data processing guidelines, and other supporting documentation that provides additional context for understanding and using the data.

External resources are typically documented using the Dublin Core Metadata Initiative (DCMI) standard, which provides a simple and widely adopted framework for describing digital resources. DCMI is a compact metadata standard used here to describe related resources such as questionnaires, manuals, reports, scripts, photographs, and maps.

Key metadata elements include:

- Resource title

- Creator

- Publisher

- Date

- Description

- Subject

- Language

- Format

- Access location or URL

### External Resources

#### Resource Description

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Label | label | Hyperlink label. | Household Questionnaire |

| Resource | resource | Link to resource. | documentation/questionnaire_household.pdf |

| Type | type | Resource type. | Document Questionnaire |

| Title | title | Resource title. | Household Questionnaire |

| Subtitle | subtitle | Resource subtitle. | Wave 1 Education Module|

| Date Created | date | Creation date. | 2024-06 |

| Format | format | Digital format. | PDF |

| Description | description | Resource description. | This is the education module of the household questionnaire for the survey... |

| Abstract | abstract | Resource abstract. | |

| Table of Contents | tableOfContents | Table of contents. | CHAPTER 1 INTRODUCTION |

### Creators and Contributors

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Author(s) | creator | Resource authors. | National Statistics Office |

| Contributor(s) | contributor | Resource contributors. | World Bank |

| Publisher(s) | publisher | Resource publishers. | National Statistics Office |

### Coverage and Language

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Country | country | Covered country. | Uganda |

| Language | language | Language used. | English |

### Identifiers and Rights

| Field | DDI Element | Description | Example |

|---|---|---|---|

| ID Number | idno | Document identifier |A DOI if one exists |

| Rights | rights | Usage rights. | © 2017, Popstan Central Statistics Agency |

### Subjects

| Field | DDI Element | Description | Example |

|---|---|---|---|

| Subjects | subjects` | Resource subjects. | Sampling; Questionnaire design; Fieldwork; Data processing |

### Why is This Important?

- Documenting external resources helps **preserve valuable contextual information that may not be captured within the data files themselves**.

- It enables users to **locate**, **access**, and **understand** supporting materials that provide important information about the data, its production, methodology, processing, quality, and intended use.

- Well-documented access and rights information provides users with a clear understanding of who can access the resource, under what conditions, and any restrictions governing its use, redistribution, or reuse.

- Comprehensive documentation of external resources improves the **discoverability** of data and related materials in catalogs and repositories.

- It supports **long-term preservation** by ensuring that critical documentation and supporting resources remain linked to the data over time.

- Well-documented external resources also enable **automated tools, search systems, and metadata platforms to connect datasets with the documentation required for accurate interpretation, analysis, reproducibility, and reuse**.

## Importing External Resources

When documenting external resources, you will need to organize your resources in your local folder by resource type following the folder structure described in *Organizing Your Files* section.

Some resources might be composed of more than one file (for example, the CSPro data entry application includes multiple files that should not be separated). In such cases, zip them into one single file, and import it as a single resource.

For documents available in multiple formats (for example, a questionnaire available in Excel and in PDF), you may create two separate resources, or zip the files into one single file. In such case, list the different formats available in the "Content/ Description" field.

::: tip Best Practices - **Naming Convention for External Resources**

- Use file names short, but self-explanatory about the content of

the document.

- Preferably, use lower cases.

- Avoid spaces to delimit words.

- Be consistent with your method of naming across all files. For

instance, if you use underscores to delimit words, keep it that

way in all files.

- Use only alphanumeric characters, underscores or dashes. Avoid

using special characters (!@#$%^&*()~) or any accented characters.

- If you intend to have an archive useable and downloadable across

multiple countries, use English names for your files.

:::

The Metadata Editor documentation provides guidelines under Documenting Data → Microdata → Import and Document the Dataset → External resources on how to import and document your resources in the Editor.

Once the data and external resources have been documented, it is essential to conduct a quality assessment before the metadata are stored, published, or shared. Metadata quality review helps ensure that the documentation is complete, consistent, accurate, and compliant with applicable metadata standards, thereby improving discoverability, interoperability, and long-term usability.

The next section explores metadata quality assessment and presents some recommended approaches, tools, and techniques that can be used to evaluate and improve metadata quality.

---

title: Microdata Documentation Quality Review Checklist

description: A practical checklist for reviewing the completeness, consistency, quality, and usability of microdata metadata and supporting documentation.

---

# 6. Metadata Quality Assessment

An **independent review** of the metadata is highly recommended prior to publishing the final output. There are two recommended methods to conduct a quality assessment. The review should consider relevant dimensions such as accuracy, reliability, coherence, comparability, timeliness, accessibility, usability, interpretability, and discoverability.

## I. Metadata Editor Assessment Tool

The Metadata Editor is equipped with a built-in AI assisted metadata assessment tool to assess and improve project metadata when review is initiated. More information on the metadata review can be found in the documentation - see Metadata reviewer.

## II. Microdata Documentation Quality Review Checklist and Feedback Form

This checklist is designed to support the review of microdata documentation before publication, dissemination, or long-term preservation. It can be used to assess whether metadata prepared using the DDI-Codebook (DDI-C) and Dublin Core Metadata Initiative (DCMI) standards is complete, consistent, clear, and usable by data users, data catalogs, repositories, and AI-enabled discovery tools.

The checklist is organized around the main metadata sections commonly used to document microdata: **Document Description**, **Study Description**, **Data Files**, **Variables**, **Variable Groups**, and **External Resources**. A final section is included for reviewing the dataset landing page or repository record before publication. A downloadable, editable version is available <a href="/media/microdata-documentation-quality-review-checklist-checkboxes.docx" download="quality-review-checklist.docx">here</a>.

## How to Use This Checklist

Use this checklist during metadata review to identify missing information, inconsistencies, unclear descriptions, formatting issues, and areas where additional documentation is needed. For each item, reviewers may indicate whether the required information is provided, partially provided, not provided, or not applicable, and may record comments and recommended actions.

Suggested review values:

- **Provided**: The information is complete and clear.

- **Partially provided**: Some information is available, but it is incomplete or unclear.

- **Not provided**: The information is missing.

- **Not applicable**: The item does not apply to the dataset or study being reviewed.

Suggested action values:

- **None**: No action is required.

- **Add**: Missing information should be added.

- **Fix**: Existing information should be corrected.

- **Check**: The information should be verified before publication.

---

#### Review Information

| Field | Details |

|---|---|

| Country | |

| Language | |

| Study title | |

| Dataset ID | |

| Dataset version | |

| Submitted by | |

| Date submitted | |

| Metadata format provided | ☐ DDI XML ☐ JSON/XML export ☐ Metadata Editor project package ☐ Other: |

| Data files provided | ☐ Yes ☐ No |

| Supporting documentation provided | ☐ Yes ☐ No |

| External resources provided | ☐ Yes ☐ No |

| Reviewed by | |

| Review date | |

| Metadata standard(s) used | ☐ DDI-C ☐ DCMI ☐ Other: |

| Standard template used | ☐ Yes ☐ No ☐ Not applicable |

| New or revised metadata produced by reviewer | ☐ Yes ☐ No |

| Name of revised metadata file, if applicable | |

---

#### 1. Document Description

The **Document Description** section describes the metadata record itself. It should identify who produced the metadata, when it was created, the metadata version, and the unique identifier assigned to the metadata document.

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Study title | Title is complete, correctly formatted, and consistent with the Study Description. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Metadata producer | Name and affiliation of the person or organization that produced the metadata are provided. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Date of production | Date is provided in a consistent format, preferably ISO format. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| DDI document version | Metadata version is provided and follows the agreed versioning convention. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| DDI document ID | Identifier is provided and is consistent with the Study Description ID, where applicable. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Revision history | Major changes to the metadata record are documented, where relevant. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

---

#### 2. Study Description

The **Study Description** section captures information about the study as a whole, including its purpose, scope, methodology, coverage, producers, sponsors, data collection, data processing, data quality, access conditions, and citation requirements.

##### 2.1 Identification

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Title | Full study title is provided, including reference year(s), and is written consistently. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Subtitle | Subtitle is provided only where needed and adds meaningful context. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Abbreviation | Abbreviation is clear, standardized, and includes the reference year where appropriate. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Study type | Study type is selected from a standard or controlled vocabulary, where available. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Series information | Series description explains objectives, ownership, scope, coverage, periodicity, and related rounds where applicable. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Translated title | Translated title is included where relevant and special characters display correctly. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| ID number | Study ID is clear, unique, and follows the agreed naming convention. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 2.2 Version

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Version description | Dataset version is clearly described and follows a standard naming convention. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Production date | Production date is provided in a consistent format, preferably ISO format. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Version notes | Notes explain what distinguishes this version from previous or future versions. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 2.3 Overview and Scope

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Abstract | Abstract clearly summarizes study objectives, scope, coverage, methodology, and key context. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Kind of data | Kind of data is selected from a standard or controlled vocabulary, where available. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Unit of analysis | Unit(s) of analysis are clearly stated, such as household, person, enterprise, school, facility, or plot. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Description of scope | Main topics, questionnaire modules, and analytical coverage are clearly described. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Keywords | Keywords are relevant, consistently formatted, and preferably based on a controlled vocabulary. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Topics classification | Topic classifications are relevant and preferably based on a controlled vocabulary. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 2.4 Coverage

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Country | Country name is provided in full and is consistently formatted. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Geographic coverage | Geographic coverage is clear and states any exclusions or limitations. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Universe | Target population is clearly described and does not rely on overly broad terms unless accurate. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Temporal coverage | Reference period or time period covered by the data is documented where relevant. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 2.5 Producers and Sponsors

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Primary investigator | Main agency or organization responsible for the study is identified. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Other producers | Co-producers or implementing partners are identified where applicable. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Funding | Funding agencies, donors, and in-kind contributors are documented where applicable. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Acknowledgments | Technical contributors, advisory groups, or other important contributors are acknowledged where appropriate. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 2.6 Sampling

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Sampling procedure | Sampling design, sample size, stratification, sample frame, replacement policy, and key sample variables are described. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Deviation from sample design | Differences between planned and actual sample design are documented. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Response rates | Response rates are documented clearly and disaggregated where relevant. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Weighting | Weight variables and weighting methodology are described, or self-weighting is explicitly stated. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 2.7 Data Collection

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Dates of data collection | Data collection dates are provided in a consistent format, preferably ISO format. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Time periods | Time periods covered by the data are documented where relevant. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Mode of data collection | Data collection mode is clearly identified, such as face-to-face, telephone, web, CAPI, CATI, CAWI, or mixed mode. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Notes on data collection | Fieldwork organization, training, supervision, and notable implementation issues are documented. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Questionnaires | Questionnaire titles, versions, languages, and content are documented. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Data collectors | Organizations responsible for data collection are identified where relevant. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Supervision | Field supervision structure and quality control mechanisms are described. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 2.8 Data Processing and Appraisal

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Data editing | Data editing methods, software, checks, and related documentation are described. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Other processing | Data entry, cleaning, anonymization, tabulation, analysis, and processing workflows are documented where relevant. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Sampling errors | Sampling error estimates, variance estimation methods, and related documentation are described where applicable. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Other data appraisal | Other data quality assessments, validation checks, limitations, and known issues are documented. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 2.9 Data Access, Rights, and Contacts

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Access authority | Organization or role responsible for granting data access is identified. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Confidentiality | Confidentiality statement is provided and is appropriate for the access level and sensitivity of the data. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Access conditions | Terms of access are clearly stated, such as public use, licensed use, restricted access, or confidential. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Citation requirements | Recommended citation includes study title, producer, country, reference year, version, and access location where applicable. | ☐ Provided ☐ Partially provided ☐ Not provided | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Disclaimer | Disclaimer is provided where appropriate and is consistent with organizational requirements. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Copyright | Copyright or rights statement is provided where appropriate. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Contact persons | Contact role, unit, or organization is provided for questions about the study or access process. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

---

#### 3. Data Files

The **Data Files** section reviews the physical data files included in the dataset. It checks whether file-level metadata is complete, whether files are listed logically, and whether relationships between files are documented and validated.

##### 3.1 File Checks

| Item | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Order of appearance | Data files are listed in a logical order, such as household before individual files or questionnaire section order. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Fix ☐ Check |

| File relationships | Key variables and relationships between files are documented and validated using statistical software, data processing tools, or metadata validation tools. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Completeness | Data files are available for all relevant questionnaire sections and derived datasets. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Fix ☐ Check |

| File structure | Record counts, variable counts, file format, and structure are documented where relevant. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| File naming | File names are clear, stable, and consistent with agreed naming conventions. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Fix ☐ Check |

##### 3.2 File Description

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Name | File name matches the data file and should not be changed unless necessary. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Fix ☐ Check |

| Content | Short description explains what the file contains and links it to questionnaire sections where applicable. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Producer | File producer is identified, typically the same as the study producer unless otherwise specified. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Version | File version is documented where file-level versioning is used. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Processing checks | File-level editing, validation, consistency checks, and quality review procedures are documented where relevant. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Missing data | File-level missing data conventions are documented where relevant. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Notes | File-specific notes, limitations, caveats, or special handling instructions are provided where relevant. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

#### 4. Variables

The **Variables** section reviews variable-level metadata, including variable names, labels, categories, question text, universe statements, missing values, derivation logic, and data quality indicators.

##### 4.1 Variable Checks

| Item | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Variable names | Variable names are stable, clear, and do not contain unsupported characters. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Fix ☐ Check |

| Variable labels | All variables have unique, clear, and descriptive labels. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Categories | Categorical variables have complete and clear value labels. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Statistics options | Summary statistics are appropriate for the variable type and do not produce unnecessary outputs for IDs or other non-analytical variables. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Weights | Appropriate weights are applied or documented where relevant. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Data type | Data types are correct and appropriate for the values stored in each variable. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Fix ☐ Check |

| Measure | Measurement level is appropriate, such as nominal, ordinal, interval, or ratio. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Time variable | Time variables are correctly identified where applicable. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Min and max values | Minimum and maximum values are documented and checked for out-of-range values where data are available. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Decimals | Decimal settings are appropriate and consistent with the data. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Missing data | Missing values are documented consistently and are not confused with valid response categories. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 4.2 Variable Description

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Definition | Definitions are provided for key concepts, derived variables, and variables that require clarification. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Universe | Universe is specified for variables where the population or skip pattern is restricted. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Source of information | Source is documented where relevant, such as respondent, household head, administrative record, or derived source. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Concepts | Concepts are documented where defined in questionnaires, manuals, or methodological documents. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 4.3 Questions and Instructions

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Pre-question text | Instructions appearing before the question are captured where applicable. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Literal question | Exact question text is attached to the corresponding variable where available. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Post-question text | Instructions appearing after the question, including skip instructions, are captured where applicable. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Interviewer instructions | Relevant interviewer instructions from questionnaires or manuals are documented. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 4.4 Imputation, Derivation, Security, and Notes

| DDI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Imputation | Imputation methods are documented for variables where values were imputed. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Recoding and derivation | Recoding and derivation logic is documented for calculated or transformed variables. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Security | Confidentiality or sensitivity indicators are documented where relevant. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Notes | Variable-specific notes provide useful clarification, caveats, or known limitations. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

---

#### 5. Variable Groups

The **Variable Groups** section is optional and is used to organize variables into logical, thematic, or hierarchical groups without changing the underlying data files.

| Item | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Variable groups | Variable groups are provided where useful for navigation, discovery, or thematic organization. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Coverage | Groups cover all relevant variables or clearly indicate why only selected variables are grouped. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Group labels | Group labels are clear, concise, and meaningful to data users. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Group structure | Hierarchical or thematic structures are logical and consistent. | ☐ Provided ☐ Partially provided ☐ Not provided ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

---

#### 6. External Resources

The **External Resources** section reviews supporting documentation and related materials described using DCMI metadata elements. External resources may include questionnaires, manuals, reports, scripts, maps, presentations, publications, methodological documents, anonymization reports, and other materials that help users understand and use the data.

##### 6.1 External Resource Checks

| Item | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Questionnaires | All questionnaire versions are provided in PDF and, where available, original editable formats. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Supporting documentation | Relevant technical, methodological, administrative, and analytical documentation is provided. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Programs and scripts | Data entry, editing, anonymization, tabulation, and analysis scripts are preserved where available. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Reports | Reports and key analytical outputs are provided in PDF and, where available, original editable formats. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Links | All links to external resources are valid and use stable or relative paths where appropriate. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Resource labels | Each external resource has a short, explicit, and user-friendly label. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Add ☐ Fix ☐ Check |

| File formats | Preservation and access formats are appropriate, readable, and documented. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 6.2 Identification

| DCMI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Type | Resource type is indicated using a standard or controlled vocabulary where available. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Title | Resource title matches the cover page or clearly describes the content. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Subtitle | Subtitle is included where applicable and matches the resource. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Author or creator | Person, agency, or organization responsible for creating the resource is identified. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Date | Date is provided at least to month and year where available. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Country | Country or geographic relevance is documented where applicable. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Language | Language is documented for documents and other language-dependent resources. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Format | File format is documented, such as PDF, DOCX, XLSX, CSV, DO, R, PY, JPG, PNG, or ZIP. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Identifier | Resource identifier is provided where available or needed for traceability. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 6.3 Contributors and Rights

| DCMI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Contributors | Contributors are documented where relevant. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Publisher | Publisher or issuing organization is documented where relevant. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Rights | Rights, license, use conditions, or access restrictions are documented where relevant. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

##### 6.4 Content

| DCMI element | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Description | Resource description clearly explains the content and purpose of the resource. | ☐ All ☐ Some ☐ None | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Abstract | Abstract is provided for documents where useful. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Table of contents | Table of contents is provided where useful and does not need to include page numbers. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Subjects | Subjects or keywords are relevant and preferably based on a controlled vocabulary. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

---

#### 7. Data Catalog / Repository Review

This section replaces legacy media-focused checks and supports review of the online dataset landing page, repository record, or catalog display before publication.

| Item | Expected review criteria | Status | Reviewer comments | Action |

|---|---|---|---|---|

| Dataset landing page | Landing page displays the study title, abstract, citation, access information, and key metadata clearly. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Branding and ownership | Responsible organization or publisher is clearly identified. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Fix ☐ Check |

| Metadata display | Metadata displays correctly and special characters render properly. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Fix ☐ Check |

| Downloads and access links | Data, documentation, and external resource links are functional and point to the correct files. | ☐ All ☐ Some ☐ None ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Searchability | Dataset can be discovered using title, country, year, topics, keywords, and other relevant metadata. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Persistent links | Stable URLs, persistent identifiers, or catalog links are provided where applicable. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Accessibility | Documents and catalog pages are readable, accessible, and usable by intended audiences. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Fix ☐ Check |

| AI readiness | Metadata is sufficiently descriptive and structured to support machine indexing, search, and AI-enabled discovery. | ☐ Yes ☐ No ☐ Not applicable | | ☐ None ☐ Add ☐ Fix ☐ Check |

---

#### 8. Overall Metadata Quality Review

Use this section to summarize cross-cutting quality issues that may affect the usability, interoperability, preservation, or dissemination of the dataset.

| Quality dimension | Review question | Assessment | Reviewer comments | Action |

|---|---|---|---|---|

| Completeness | Are the required metadata sections sufficiently completed? | ☐ Good ☐ Partial ☐ Needs work | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Consistency | Are titles, IDs, dates, versions, producers, and access statements consistent across sections? | ☐ Good ☐ Partial ☐ Needs work | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Clarity | Is the metadata clear, understandable, and useful to data users? | ☐ Good ☐ Partial ☐ Needs work | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Data quality documentation | Are processing, validation, missing data, and data appraisal information adequately documented? | ☐ Good ☐ Partial ☐ Needs work | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Reproducibility | Are derivations, transformations, scripts, and processing steps documented in enough detail to support reproducibility? | ☐ Good ☐ Partial ☐ Needs work | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Interoperability | Does the metadata follow relevant DDI-C and DCMI conventions consistently? | ☐ Good ☐ Partial ☐ Needs work | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Preservation | Are original files, preservation formats, and supporting documentation sufficiently retained? | ☐ Good ☐ Partial ☐ Needs work | | ☐ None ☐ Add ☐ Fix ☐ Check |

| Discoverability | Can users and systems find the dataset using meaningful topics, keywords, coverage, and descriptions? | ☐ Good ☐ Partial ☐ Needs work | | ☐ None ☐ Add ☐ Fix ☐ Check |

---

#### Reviewer Summary and Recommendations <br><br>

##### Overall Assessment

☐ Ready for publication

☐ Ready after minor corrections

☐ Requires substantial revision before publication

☐ Not ready for publication

##### Summary of Main Issues <br><br>

<!-- Add a concise summary of the main documentation issues identified during review. -->

##### Required Corrections <br><br>

<!-- List corrections that must be completed before publication. -->

##### Recommended Improvements <br><br>

<!-- List improvements that would strengthen documentation quality but are not blockers. -->

##### Reviewer Notes <br><br>

<!-- Add any additional notes, observations, or follow-up recommendations. -->

---

## Optional Reviewer Sign-off

| Field | Details |

|---|---|

| Reviewer name | |

| Reviewer role/unit | |

| Review completed on | |

| Final recommendation | |

| Follow-up required | ☐ Yes ☐ No |

| Follow-up owner | |

Once the quality assessment has been completed and you are satisfied that the metadata and associated resources meet the required standards, the next step is to generate the machine-readable metadata files for publication, cataloging, and sharing. These metadata files enable discovery, interoperability, and exchange across repositories and data catalogs.

The next section explains how to generate standards-compliant, machine-readable metadata and prepare it for publication in online catalogs and other dissemination platforms.

---

# 7. Generating the Output for Publication

When you have completed the necessary checks and addressed quality issues, the next step is to export the metadata for publication in NADA or other platforms, archiving and sharing as needed. The Metadata Editor is equipped with the functionality to export metadata to machine readable formats, generating the DDI Codebook in XML and JSON for microdata, and RDF for external resource metadata. These files are machine and human-readable, and can be ingested into online cataloging systems line NADA, DataVerse etc. Users can also export the metadata to a PDF report formatted for human readability and easy navigation across the metadata sections. Generating the PDF document is optional, and the file can also be distributed along with the data as an external resource. Note that whenever the metadata is updated, these files need to be re-generated to capture the revisions.

::: warning

The PDF report will include a list of all external resources

related to the study. This list should include this PDF report

itself. **Before** you generate it, make sure you create one entry

in the External Resources for documenting this report. Immediately

after you generate the PDF report, import it in the Editor.

:::

One thing to keep in mind is that in a dataset with a large number of variables may produce a document that is very long. If the report is in excess of 300 or 350 pages, you may want to split this report (e.g., produce one report with the study metadata, and one with the files and variables metadata), or change the content options (e.g., not including a frequency table for all variables).

If your agency has a website, you may upload this PDF directly to the web server. The IHSN recommends the use of a proper DDI-C compliant cataloguing system, such as the one provided by its National Data Archive (NADA) application. NADA is an open source package, available free of charge at ihsn.org.

The Metadata Editor also facilitates exporting all the data, resources and metadata to a project level .ZIP that can be pushed to a catalog, shared or re-imported into the Metadata Editor. The next section covers cataloging data and and metadata.

# 8. Cataloging Data, Metadata and Resources

Once the data and metadata have been reviewed, validated, and approved, they are ready to be published in an online data catalog. Publishing data in a catalog makes them easier to discover, access, understand, and reuse by both internal and external audiences.

## Why Catalog Data?

An online catalog provides several important benefits not limited to the following:

- **Facilitates data discovery** by making datasets searchable and easier to locate.

- **Improves data understanding** by allowing users to review metadata, documentation, and related resources before requesting access.

- **Supports informed data use** by providing information on methodology, coverage, quality, and terms of use.

- **Enables secure data access** by applying appropriate access controls and restrictions where necessary.

- **Promotes data reuse and transparency** by increasing the visibility and accessibility of research and statistical outputs.

## How It Works

Open-source cataloging platforms such as NADA (National Data Archive) can readily ingest metadata documented using the DDI Codebook (DDI-C) and Dublin Core (DCMI) standards. Once imported, the metadata can be published through a searchable web interface, alongside data files, documentation, questionnaires, reports, and other related resources.

Detailed guidance on adding metadata and data content to a NADA catalog is available in the NADA Documentation:-

- NADA Documentation

Additional Resources

- International Household Survey Network (IHSN) Data Archiving Resources

These resources provide practical guidance on metadata standards, catalog management, data dissemination, and best practices for preserving and sharing microdata collections.

# Running Validations and Diagnostics

The Metadata Editor includes a useful series of diagnostic and validation modules (see the drop down menu *Tools*): these range from very simple validations (such as the *Tools-Validate Metadata*) to complex visual displays that iterate through each variable and provides feedback to the archivist at the variable level.

- *Validate Metadata*: verifies that all mandatory fields are filled in.

- *Validate External Resources*: verifies all mandatory fields in the External Resources are filled in.

- *Health Check*: displays a popup window that provides some information and diagnostics regarding the R package. The Metadata Editor uses R to import and export the data. The health check option tells the which version of R is being used on the machine. In addition, the Health Check will provide information on the Environment Path and the result of the execution of the R script.

- *Validate Dataset Relations*: this option is used to validate hierarchically related datasets. The 'Base key variables' should be the variables that uniquely identify a case within that file.

- *Translation Manager*: provides the user with the ability to translate the interface of the Metadata Editor into any language. Selecting this option will display the Translation Manager interface. When using the option for the first time, the English template will be displayed with all the labels that require translation.

In addition to these validations, it is recommended that you generate the DDI document (in the Editor, use the Export DDI" function) and verify the size of the resulting \[.xml\] file. A fully documented survey with a large number of variables should not produce a file larger than 10Mb. Very large DDI files often indicate errors in the selection of summary statistics (for example, frequencies are produced for a variable like the household ID in a sample household file).

# Conclusion

Effective data curation, archiving, and dissemination extend beyond simply preserving datasets. They require a systematic approach to organizing data and supporting resources, preparing materials for long-term preservation, documenting them using recognized metadata standards, and ensuring that metadata are accurate, complete, and fit for purpose.

This guide has introduced the key concepts and processes involved in preparing microdata and related resources for cataloging and dissemination. From organizing repository structures and assessing data readiness, to documenting datasets with **DDI-C** compliant metadata, describing external resources using **DCMI** standards, conducting metadata quality assessments, and generating machine-readable metadata for publication, each step contributes to making data more discoverable, accessible, interoperable, and reusable.

The value of metadata standards extends beyond documentation. By transforming metadata into structured, machine-readable formats, datasets become visible and understandable to search engines, data catalogs, repositories, applications, and automated discovery systems. This enables datasets to be discovered, linked, and reused across platforms and communities that may never directly access the original repository, significantly expanding the reach and impact of the data.

By adopting the guidelines and practices described in this guide, and leveraging open-source tools such as **sdcMicro**, the **Metadata Editor**, and **NADA**, data producers and curators can create high-quality, standards-compliant metadata that supports long-term preservation, enhances machine-driven discovery, facilitates integration with broader data ecosystems, and maximizes the value of data assets.

Ultimately, well-documented data are more than a record of what was collected. They provide the context, transparency, and evidence needed for users, applications, and automated systems to find, understand, trust, reproduce, and reuse data effectively. In doing so, metadata serves as the bridge between data preservation and data utilization, ensuring that valuable data resources remain discoverable, accessible, and meaningful for years to come.