Collecting OIMS-compatible metadata for data assets

3rd draft documentation for Template version 1.5

Gideon Kruseman[1*], TBD

[1] Alliance Bioversity International & CIAT

* corresponding author: g.kruseman@cgiar.org

May 15, 2023

This working paper shares insights from work in progress by the CGIAR Foresight Initiative, and is shared for discussion. This paper has gone through a light review process at the initiative team leadership level and is shared with the approval of the relevant work package leader, but it has not been formally peer reviewed.

The CGIAR Initiative on Foresight combines state-of-the-art analytics, innovative use of data, and close engagement with national, regional and global partners to offer better insights into alternative transformation pathways that can inform choices and sharpen decision-making today, leading to more productive, sustainable and inclusive food, land and water systems in the future. More information can be found at https://www.cgiar.org/initiative/24-foresight-and-metrics-to-accelerate-inclusive-and-sustainable-agrifood-system-transformation/.

Acknowledgements

This report was prepared by work package 3 team and financially supported by the CGIAR Initiative on Foresight and Metrics to Accelerate Food, Land and Water Systems Transformation. CGIAR is a global research partnership for a food-secure future, dedicated to transforming food, land, and water systems in a climate crisis. We would like to thank all funders who supported this research through their contributions to the CGIAR Trust Fund: https://www.cgiar.org/funders/ .

Preface

By WP leader or Keith/Elisabetta, if appropriate

Table of Contents

Introduction

Metadata is a critical component of data management in today's data-driven world, as it provides context, description, and meaning to data, enabling it to be discoverable, understandable, and reusable. It serves as a bridge between the data and its users, making it possible to access and reuse data across different contexts and applications. As such, metadata plays an important role in the proper management, preservation, and sharing of data.

Recent years have seen a growing focus on the development of FAIR (Findable, Accessible, Interoperable, and Reusable) metadata standards, which aim to make data more discoverable and reusable by enabling automated discovery and integration of data (Drechsler et al, 2015). The adoption of FAIR metadata standards ensures that data is not only machine-readable but also human-readable, making it easier for users to locate, access, and understand data. This makes FAIR metadata principles essential for effective data management, sharing, and reuse (Wilkinson et al 2016; Recker et al 2019).

Research has shown the importance of metadata in ensuring effective data management and reuse (Choudhury & Shreeves 2013). Incomplete or inaccurate metadata can lead to data being misused, misinterpreted, or underutilized (Borgman 2012). Another study by Tenopir et al. (2011) found that the quality of metadata is critical to the reuse and citation of data, as well as the identification of data sources.

Additionally, metadata plays a vital role in ensuring data interoperability across different systems and applications (Duke et al 2014). According to Dappert et al. (2017a), the use of consistent metadata standards can help ensure that data is correctly interpreted by different systems and enable data integration across different domains.

Overall, metadata is a critical component of effective data management, preservation, and sharing. The adoption of FAIR metadata standards ensures that data is discoverable, accessible, interoperable, and reusable, making it easier for users to locate, access, and understand data. The use of consistent metadata standards is also essential for ensuring data interoperability and integration across different systems and domains.

The user guide aims to facilitate the provision of necessary metadata for archiving data sets, including those for open access and restricted use (National Science Foundation 2015). It specifically focuses on data sets related to agriculture, agronomy, breeding, and human subjects research. With the increasing complexity of data analysis tools and metrics, particularly in developing countries, the need for effective foresight analysis has become critical. The guide seeks to reduce barriers to access and use of the CGIAR's core foresight models, databases, and systems-level metrics.

By following the best practices outlined in the guide, data managers can ensure that metadata is of high quality, FAIR compliant, and can be effectively used and shared across various disciplines and research communities. Additionally, the guide references the CG-Core metadata standards, which provide a comprehensive framework for describing and managing research data. The guide aims to support data managers in creating metadata that enables the effective management, preservation, and sharing of data for present and future use.

The metadata template was originally developed by the International Maize and Wheat Improvement Center (CIMMYT). It is based on the CG-Core metadata standards but vastly expanded to incorporate

important missing metadata. The template was originally developed for tabular human-subjects and human-subjects related research survey data has been expanded to accommodate other data types from agronomy and breeding research. The idea was to have a template in a structured way that would allow the data to be exported to an OIMS-compatible metadata format (Kruseman 2022).

Metadata is essential for archiving data (United Nations 2019). Archiving data ensures that it is available for future use and is an essential component of open science. Metadata facilitates data discovery and reuse, making it a critical component of archiving.

In this user guide, we will provide an overview of the metadata template (version 1.5), including six sections: Descriptive Metadata, Technical Metadata, Structural Metadata, Controlled Vocabularies, information on the metadata contributor and information on any tools used to extract or geberate metadata. For each section, we will provide an explanation of the metadata fields included in the template, best practices for filling out each field, and references to additional resources.

We hope this user guide will be a valuable resource for data managers in creating effective metadata for archiving data.

Overview of the template

The metadata template includes three main types of metadata: descriptive metadata, technical metadata, and structural metadata.

Descriptive metadata is a critical component of data management as it provides contextual information about data that enables users to understand, access, and effectively use it. Descriptive metadata contains information about the origin, purpose, context, and contents of the data, which are essential for its discoverability and reuse.

Several studies have shown the importance of descriptive metadata in ensuring effective data management and reuse. Borgman (2012) found that incomplete or inaccurate metadata can lead to data being misused, misinterpreted, or underutilized. Similarly, the quality of metadata is critical to the reuse and citation of data, as well as the identification of data sources (Tenopir 2011). Furthermore, we should emphasize the role of descriptive metadata in ensuring the transparency, credibility, and trustworthiness of data (Koskela et al 2016).

The importance of descriptive metadata is also highlighted by the adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) metadata standards. The FAIR principles require that data is described with sufficient metadata to enable its discovery and reuse. Inadequate or missing descriptive metadata can lead to data being overlooked or difficult to use, undermining the FAIR principles (Wilkinson et al., 2016).

In summary, descriptive metadata is crucial for effective data management, preservation, and sharing. It facilitates data discovery and reuse, making it a critical component of archiving. The adoption of FAIR metadata standards emphasizes the importance of descriptive metadata in enabling the effective management and reuse of research data. Descriptive metadata can be provided both at the level of datasets as well as at the level of data files.

The chapter on dataset level metadata provides a good overview of the CG Core metadata schema and its importance for describing information products published by CGIAR centers and entities. The chapter explains that most of the CG Core metadata fields are included in the template, while some are provided at a later stage in the data management process, such as the persistent identifier. The chapter also covers the key CG Core metadata fields, including identifier, title, subtitle, description, type, language, license, rightsHolder, accessRights, issued, embargoDate, creator, contributor, temporal coverage, spatial or geographic coverage, subject, concept or keyword, and specific socio-economic metadata fields that are sometimes relevant to other domains.

The chapter provides clear definitions for each metadata field and highlights whether it is included in the template or not. It also provides explanations of some key concepts and compound variables such as subtitle, creator, and contributor. Moreover, the chapter covers specific metadata fields that are not part of the CG-core metadata schema but are relevant to certain types of data, such as unit of analysis, time method, data source, frequency, and domains.

The chapter concludes with sections on ethics approval related to the data, sensitive information, publications related to the dataset, and model output or metrics generated with a model, tool, or algorithm. The explanations are clear and concise, making it easy for users to understand the metadata fields and their importance in describing information products. Overall, the chapter provides a

comprehensive overview of the CG Core metadata schema and the template, making it a useful reference for anyone working with information products published by CGIAR centers and entities.

At data file level we provide specific descriptive metadata but more importantly the technical metadata.

Technical metadata is essential in describing the technical characteristics of data files, including format, structure, and storage location. This information is necessary for managing and preserving data files, as well as for discovering and reusing data. Technical metadata provides essential information about data files that facilitate data management, preservation, and sharing. For instance, data files stored in different locations or in different formats may require different preservation strategies. Technical metadata can also provide information about the quality and completeness of the data files.

Technical metadata plays a vital role in data preservation and sharing (Devarakonda et al. 2013). Technical metadata can help data curators to identify and manage different versions of data files, to monitor the quality of the data files, and to ensure the long-term preservation of the data files. In addition, technical metadata is essential for ensuring data interoperability across different systems and applications. The use of consistent metadata standards can help ensure that data is correctly interpreted by different systems and enable data integration across different domains (Dappert et al. 2017b).

The technical metadata section of the metadata template includes fields such as file name, description, storage location, data format, recording, file format, file type, origin, data file structure, metadata, metadata location, location of backups, and backup frequency. These fields provide a comprehensive description of the technical characteristics of the data files and are critical for managing and preserving data files.

In summary, technical metadata is essential for managing and preserving data files, discovering and reusing data, and ensuring data interoperability. By including technical metadata in data files, data managers can ensure that data files are preserved and shared effectively and that they can be used across different systems and applications.

Structural metadata is essential for understanding the internal structure of the data files and for interpreting the data correctly (Michener and Jones, 2012; Rugg and McInerney 2012). It includes information on the data type, unit of measurement, variable name, variable description, and other relevant information about the data structure. Structural metadata is important for data reuse, data interpretation, and data integration across different systems.

For example, if a researcher wants to use data from different sources for an analysis, the structural metadata of each data set needs to be consistent so that the data can be integrated properly. The absence or inconsistency of structural metadata can lead to errors and incorrect conclusions (Ghosh, 2014).

Furthermore, structural metadata is essential for the reproducibility of research. To reproduce the results of a study, researchers need to have access to the data and the associated metadata, including the structural metadata. Without proper structural metadata, it may be impossible to reproduce the original results or conduct new analyses on the data.

Together, these different types of metadata provide a comprehensive picture of the dataset, enabling users to discover, understand, and effectively use the data. By including all three types of metadata in the template and following best practices for filling out each field, data managers can ensure that the dataset is effectively managed, preserved, and shared for present and future use.

In the template we make use of two types of non-tabular metadata fields types. The first is compound variables or attributes. The second is variables or attributes that allow multiple values. The latter is self explanatory, the former requires some explanation.

Compound metadata fields are a type of metadata attribute that consists of multiple sub-attributes. They are used to provide more detailed and specific information about a particular metadata attribute. For example, the "creator" metadata field is a compound field that consists of multiple sub-attributes, such as "name", "affiliation", and "ORCID ID". By using compound metadata fields, data managers can provide a more complete and accurate description of the dataset or data files, making them more discoverable and reusable. Compound metadata fields can also help to ensure that metadata is collected consistently across different datasets or data files. It is important to define the sub-attributes of a compound metadata field and provide clear guidance on how to complete them to ensure consistency and accuracy.

The ”creator” metadata field is also an example of a field that can have multiple values, albeit multiple combinations of “name”, “affiliation” and “ORCHID ID”.

Finally, there are a few special information sheets. Firstly, there is a specific tab in the template to collect information about the person or persons that have put together the metadata. Secondly, there is a sheet that collects information about any and all tools used to prepopulate the template. An example is the use of metadata extraction tools to extract already available metadata from for instance a STATA (*.dta) file.

Descriptive metadata at dataset level

The CG Core aims to describe all types of information products that are published by the different CGIAR centers and entities. There are many benefits of having a clear and harmonized way to describe information products (CGIAR:

- better interoperability
- better transparency
- easy monitoring

Most of the CG core metadata fields can be found in the template and many in the descriptive metadata at dataset level. Some have been omitted because they are provided at a later stage in the data management process, such as the persistent identifier. Some of the descriptive metadata are linked more to data file level than to a data set level and therefore that metadata is collected elsewhere in the template. Some of the key CG core metadata fields that are not included in the template are marked in **red**. The reason for not including them is also provided. The CG Core metadata fields included are marked in **blue**. The fields that are underlined have a link to the official URL of that concept.

Metadata Properties defined in the CG core include:

identifier is defined as an unambiguous reference to the resource within a given context. The recommended best practice is to identify the resource by using a DOI or a handle. This is done at publication time and therefore is not collected with the metadata template.

title is defined as a name given to the resource, following standard title formatting for capitalization and punctuation.

subtitle is defined as a brief statement that provides context or clarification to the title. Subtitles can provide more detail about the resource or highlight a specific aspect of its content, making it easier for users to understand and navigate the resource. In metadata, including a subtitle can improve discoverability and search engine optimization, as well as aid in linking related resources.

description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource. Descriptive details significantly improve discoverability via search engines such as Google and Bing, and will aid interlink between related resources at the meta-search/indexer level. Descriptions can be provided in multiple languages if appropriate and available.

type is a key core CG metadata field and is defined as the nature of the resource. For data sets as we describe them here the term is **Dataset As A Collection Of Data**. Therefore we do not collect that piece of metadata in the dataset level descriptive part of the template. We actually include it in the technical metadata part for individuals files.

language if the resource is defined at the file level and not the dataset level.

license is defined as a legal document giving official permission to do something with the resource. Under this heading we also include Data Transfer and Use Agreements (DTUAs)that do not fall under the list of licenses from SPDX. As is the case with human subjects research data.

rightsHolder is defined as A person or organization owning or managing rights over the resource.

The term accessRights is defined as Information about who can access the resource or an indication of its security status. Access Rights may include information regarding access or restrictions based on privacy, security, or other policies. Values should come from the following list: "open", "restricted".

issued is another field that is not included in the template as it refers to the date when the information product was created in its final form to be published.

embargoDate is relevant in cases when the information product has an embargo this date indicates when it would be available.

creator is used to be used to link the resource to one or more creator or contributor objects. This is a compound concept consisting of:

- Name
- Affiliation
- OrcidID

If a different ID needs to be added both the ID name as well as the unique ID associated with that ID.

contributor is used for Organisation, or service making contributions to the information product. For people we use the creator field.

- Type is a controlled vocabulary, including new terms such as "*One CGIAR Initiative*"
- Name

Temporal coverage is identified by the **start date** and the **end date** of the data and the **temporal resolution** identifies over what time frame the data uses (yearly, monthly, weekly, daily, hourly, etcetera)

spatial or geographic coverage. We recognize a couple of different classification principles.

- AEZ: Agro-ecological zone
- Region, especially but not limited to CGIAR regions. Regions consist of one or more countries
- Country
- Administrative unit (official)
    - Unit type
    - Unit name
- If it is a different spatial classification provide the following information
    - Spatial classification name
    - Spatial classification URL
    - Specific ID

An example of a different spatial classification is the FPU used in the IMPACT model.

subject, concept or **keyword** is defined as a unit of thought describing the subject of the information product. It is a compound term consisting of the the vocabulary on which the keyword is based, the label of the keyword and the URI of te keyword.

There are a few specific socio-economic metadata fields that are missing from the CG-core metadata schema, which are sometimes also of relevance to other domains.

Unit of analysis refers to what unit of analysis is relevant for the data. Sometimes this is apparent from the spatial coverage, but not always. Especially when it is not apparent from the spatial coverage this should be filled in. possible values could be e.g. plot, household, individual, community.

Time method refers to whether the data is longitudinal, time series or a panel dataset.

Data source refers to how the data was collected. Typical values include: survey, census, official statistics, observation, model output, machine generated.

Frequency refers to how often the data is/was collected, especially relevant if not apparent from the temporal resolution mentioned earlier.

If the data is primary human subjects research data or human subjects related research data, **response rate**, is a valuable attribute.

Sampling frame should always be included.

Domains refers to the scientific domains relevant for the data, typically this would be concepts such as economics, health, social sciences, agricultural sciences, meteorology.

The **impact area focus**, is a specific CGIAR tag related to the five impact areas of the CGIAR. Data that is especially relevant for a specific impact area, should be tagged as such.

The next section in the descriptive metadata refers to ethics approval related to the data. This is especially relevant for primary human subjects research data and primary human subjects related research data.

The first field is whether or not and if so what type of review was conducted. Valid values are:

0. no;
1. exempt;
2. expedited;
3. full

If ethics review was conducted, provide the name of the Institutional review board (IRB) or institutional research ethics committee (IREC). The registration country, organization and registration code refer to identification of the IRB is it is officially registered. Reference code refers to the reference code provided by the IRB with the approval. Research can be approved by multiple IRBs.

If the data contains sensitive information including but not limited to sensitive information from sensitive questions, such as gender related questions, women's empowerment, crime and corruption or other legally challenging questions, or politically sensitive questions?

The next section is a compound variable capturing any key publications related to the data set if relevant.

The final section of the dataset level refers to model output or metrics generated with a model, tool, or algorithm.

Technical metadata at data file level

The technical metadata at file level provides information about individual files that make up a data set. Technical metadata provides information about the structure, format, and location of the data files, as well as details about their processing and storage. This section of the metadata template is crucial for ensuring that data can be effectively managed, accessed, and reused.

The sheet is divided into 3 parts. Each line references a specific file in the dataset. The three parts are in the columns.

The first part columns A and B serve as identification for the specific data file in order to link it to the information in the descriptive metadata at file level and the structural metadata. The next part consisting of columns C through O, consist of the purely technical metadata. In columns P through T cover versioning and quality control.

The identification section provides basic information about the file, such as its location, name, size, and storage location. This information is used to uniquely identify the file and link it to other metadata about the file, such as its description and structure. The first two columns of the technical metadata template are the file location and file name.

The file location refers to the specific folder or location on a shared drive (such as Sharepoint, OneDrive, GoogleDrive, or Dropbox) where the file is stored. This information is important for identifying the location of the file in case it needs to be accessed or retrieved in the future.

The file name column is where the name of the file is recorded. This information is important for identifying and referencing the specific file within a dataset or project. It is recommended to provide a descriptive and meaningful name for the file that accurately reflects its contents and purpose.

The technical metadata section is an essential part of the documentation of a file. It provides detailed information about the file, which is crucial for understanding and working with the data contained within it.

The first aspect covered in this section is file size. The size of the data file is an important factor to consider, particularly for large data files. It is recommended to document the file size in the technical metadata to provide users with an understanding of the amount of storage required for the file. Additionally, the storage location of the file should be documented with the relevant URI, if known. This information is crucial for locating and accessing the file in the future, and ensuring that it is properly stored and managed.

Encryption: This field indicates whether the data file is encrypted or not. If the data file is encrypted, please specify the encryption method used. If not encrypted, this field should be marked as "none". Additionally, include information about any other security measures that have been taken to protect the data file.

The "Access requirements" field refers to any requirements or restrictions related to accessing and using the data file. This could include login credentials, permission levels, or any other access requirements needed to use the data file. Providing this information is important for ensuring that only authorized individuals are able to access and use the data, and that any necessary security measures are in place. It

is important to specify any access requirements clearly to avoid unauthorized access and ensure the confidentiality and integrity of the data.

The data format of the file is an essential aspect to understand when working with the data. This field describes the format of the data file, such as text, audio, or video. Additionally, for unstructured data sets, the recording or transcription method should be documented, such as handwritten notes, digital audio recording, or speech-to-text transcription.

The file format is an important field to document, especially if it is not crystal clear from the file extension. This field indicates the file format and distinguishes between different types of files, such as data files, support documentation, ETL procedures (e.g., STATA do files), and metadata files. The file type is also important to document, and distinguishes between different types of data files, such as tabular, multi-index, unstructured, script, and OIMS-compatible metadata files.

Data file structure is another crucial aspect to document, particularly for tabular data files. This field describes the structure of the data file and can include details such as the number of rows and columns, variable types, and variable names.

The “structural metadata” field indicates whether the file has associated structural metadata, which is crucial for understanding the data contained within the file. The metadata location indicates where the structural metadata of this file is located, distinguishing between metadata located in this template, separate data dictionary OIMS-compatible JSON, or other data dictionaries. The metadata location field refers to the location of metadata specifically associated with the data file, as opposed to other metadata that may be relevant to the dataset as a whole.

Finally, the location of back-ups and the frequency of back-ups during the data collection process should be documented if applicable. This information is important for ensuring data integrity and reliability.

The versioning and quality control section provides information on the history of the data file and any quality control measures that have been taken to ensure the accuracy and consistency of the data.

• Data file version: This field provides information about the version of the data file, particularly relevant when there are multiple versions of the same data file.

• Data quality checks: This field includes information about how the data was checked for completeness, accuracy, and consistency. It can also provide details about any data cleaning or transformation procedures that were performed to improve data quality.

• Data processing steps: This field includes information about how the data was processed, such as the algorithms or models used and the parameters that were used in the processing. This information can help ensure transparency in the data processing procedures.

• Data file dependencies: This field indicates whether there are any other data files that are needed to use the current data file. This information can help users understand any dependencies between data files.

• Data file history: This field provides information on changes made to the data file relative to the previous version. This information can help track changes to the data over time and ensure that users are working with the most up-to-date version of the file.

By filling out these technical metadata fields, data managers can ensure that the data is properly documented and can be effectively managed and reused. The CG-Core metadata standard provides a comprehensive framework for technical metadata, including the fields outlined above. Additional resources for technical metadata include Duke et al. (2014) and the National Science Foundation's Data Management Plan guidance (National Science Foundation 2015)

Descriptive metadata at data file level

The descriptive metadata at file level in the template is a crucial part of the documentation of the data file. It provides information about the content and context of the file, which is essential for understanding and working with the data contained within it. The descriptive metadata template includes several important fields, as outlined below:

Identification was discussed in the previous chapter and serves to ensure that the files between sheets are linked correctly:

File location: This field refers to the specific folder or location on a shared drive (such as Sharepoint, OneDrive, GoogleDrive, or Dropbox) where the file is stored. This information is important for identifying the location of the file in case it needs to be accessed or retrieved in the future.

File name: This field records the name of the file. This information is important for identifying and referencing the specific file within a dataset or project.

Descriptive metadata fall into key descriptive metadata for any file and descriptive metadata that are relevant for specific files. The general descriptive metadata include:

Description of the file: This field provides a brief description of the content and purpose of the file. It can include information such as the type of data, the research question it addresses, or the specific section of the questionnaire covered by the file.

if relevant for the data file: The purpose or research question that the data file addresses could be included in the descriptive metadata. This information can be helpful for understanding the context of the data and how it was collected or analyzed.

Language: This field indicates the language of the data contained within the file.

Origin: This field indicates the data source or origin and can be useful for tracking where the data came from and who is responsible for it. Sometimes this information is provided at data set level if identical for all files in the data set.

Specific descriptive metadata depend on the type of data. If we are dealing with unstructured or semi-structured data, especially related to anthropological and sociological research.

Interview or observation protocol: This field is especially relevant for unstructured data sets and indicates the method used to collect the data, such as structured interview, semi-structured interview, ethnographic observation, or field notes. Note that, field notes are independent of scientific domain.

Interview setting: This field indicates the setting in which the interview or observation was conducted, such as in-person, phone, or video conference.

Languages spoken: This field indicates the languages spoken by the participants in the data collection process.

If the data set is GIS data, then certain: geospatial reference system: If the data file includes GIS data, this field indicates the geospatial reference system used.

Projection: This field indicates the projection used for GIS data.

If the data is related to image file, specific information is needed.

image resolution: If the data file includes an image file, this field indicates the resolution of the image.

Image color space: This field indicates the color space used for the image file.

Other image technical data: This field can be used to include any additional technical data related to the image file.

By filling out these descriptive metadata fields, data managers can ensure that the data is properly documented and can be effectively managed and reused. It is important to provide as much information as possible to ensure that the data can be accurately interpreted and used by other researchers.

Structural metadata

Structural metadata refers to the description of the organization of data files and their internal structures. It provides information on the data type, format, and the relationships between different elements in the data, enabling data discovery, interpretation, and analysis. Structural metadata is essential for data reuse, as it allows users to understand the composition of the data and how it can be queried and manipulated.

In practical terms, structural metadata includes information on the structure of data files, such as column headers, data types, and units of measurement. It may also include information on how different data files are related to each other, such as foreign keys or common identifiers. Structural metadata is often stored in a data dictionary or schema, which provides a formalized representation of the data structure and is used to facilitate data integration and interoperability.

The Structural Metadata sheet in the metadata template captures information about the structure of the data, including the arrangement of the data within files and tables, and the specific characteristics of the variables. This information is critical for interpreting the data and for ensuring that the data can be integrated with other datasets.

The structural metadata is divided into several sections, including Identification, Structural metadata, ethics related issues and quality assurance. The structural meta data section is divided into a number of sub-sections: Basic Structural Metadata, Compound and Multiple Value Variables, Measurement, Keys, Controlled Vocabularies and Ontologies. Each section and sub-section provides essential information about the dataset that is necessary for understanding the data and ensuring its effective management, interpretation, and reuse.

Identification

Identification: This section provides basic information about the data file, such as its location, name, and storage location. It also includes a field for the sheet name and table number if the data is in an Excel workbook with multiple sheets or contains multiple tables.

Sheet: If the data is in an Excel file with multiple sheets, enter the name of the sheet that contains the relevant structural metadata information. If there is only one sheet, leave this field blank.

Table: If the file contains multiple tables, enter the name of the table. You also need to specify how different tables are separated. This could be in an Excel file where tables have names.

If the file is a multi-index data file the metadata attribute fields **multi index 1** and **multi-index 2** provide the relevant information. Multi-index 1 indicates if the variable or index is in a row column or the cells of the table. Multi-index 2 indicates the or column number of the index being described in this record.

Basic structural metadata

The first sub-section of the main structural metadata section provides technical information about the data fields, including the variable name, label, description, format, data type, and missing values. It is the type of information that can often be extracted from existing data files such STATA *.dta files. This is in columns I through N.

The variable name refers to the name of the variable as it appears in the data file. The variable label provides a short description of the variable, while the variable description provides a more detailed explanation of the variable.

The variable format determines how variables are displayed, such as date formats or numeric formats. The data type of the variable can be chosen from the following options: text, number (integer or float), enumeration (factor or code), or phone number. The missing values field indicates how missing values are displayed in the data.

Compound and multiple value variables

The second sub-section provides information if the variable is complex. There are two types of technical complexity. The first is if a number of variables belong together intrinsically. An example is if the data file contains a value and a variable for the unit of measurement and if necessary a conversion factor. This is a compound variable. Another example is a household roster that is included in the main household file with multiple household members and their characteristics placed sequentially in different variables. In the latter case we are dealing with a compound variable that exists multiple times. We can also have a variable that is not compound but for which multiple answers are possible. Please provide unique IDs to know how to link the variables.

Measurement

The third sub-section, the "Structural metadata: measurement" section in the metadata template captures information about the units of measurement, the method used to obtain the value of the variable, and the protocol of the measurement method. This information is critical for understanding the accuracy and reliability of the data.

The "Unit_of_measurement" field provides information about the units used to measure the variable. This information is important for ensuring that the data can be interpreted correctly and used in subsequent analyses. Examples of common units of measurement include meters, kilograms, liters, and seconds.

The "Method" field indicates how the value of the variable was obtained. This information is important for understanding the accuracy and reliability of the data. Possible methods include direct measurement, recall, estimate, opinion, or expert judgment.

The "Protocol" field provides information about the protocol used to measure the variable. This can include details about the specific instruments or procedures used, or any standards or guidelines that were followed during the measurement process. This information can help ensure the transparency and reproducibility of the data.

It is important to provide as much detail as possible in this section to ensure that the data can be effectively managed, understood, and reused.

Keys

The "keys" field in the structural metadata template indicates whether a given variable is a key or not, and if so, what type of key it is. A key is a field or combination of fields that is used to uniquely identify a record or row in a table or dataset.

There are three possible options for this field:

- If the variable is not a key, mark it as 0.
- If the variable is a primary key, mark it as 1. A primary key is a field or combination of fields that uniquely identifies a record in a table and is used as a reference for foreign keys in other tables.
- If the variable is a foreign key, mark it as 2. A foreign key is a field in one table that refers to the primary key of another table, establishing a link between the two tables.

Including information about keys in the structural metadata is important for understanding the relationships between different tables or datasets, and for facilitating data integration and interoperability. It can also help ensure data accuracy and consistency, as keys provide a way to enforce data integrity constraints and prevent duplicates or inconsistent records.

Controlled vocabularies and ontologies

The controlled vocabulary section of the structural metadata at variable level captures information about the enumerations or codes used in the data. Enumerations are a set of predefined values that can be assigned to a variable to indicate its possible values. This section of the metadata template provides a standardized approach for describing these enumerations to ensure that they are clear, accurate, and reusable.

The first five columns of the section describe the controlled vocabularies of the values in the field. The Enumeration column captures the name of the enumeration used for the variable. The Enumeration_description column provides a brief description of the enumeration. The vocabulary ID column captures the unique identifier for the controlled vocabulary used for the enumeration. The Enumeration_description_link column provides a link to a file that contains more information about the local controlled vocabularies used as enumerations/coding. Finally, the Comment column allows for any additional information about the enumeration to be added.

The last four columns relate to ontology terms and concepts. The Ontology_name column captures the name of the ontology that the term is associated with. The Ontology_uri column captures the URI (unique resource identifier), such as the web address of the ontology term. The Ontology_term column captures the actual ontology term that is associated with the variable. Finally, the Notes column allows for any additional notes about the ontology terms to be added.

Providing this information in a standardized format makes it easier for other researchers to understand the enumerations used in the data, and to reuse the data for other purposes. This can improve the interoperability and accessibility of the data, and ensure that it is used appropriately and effectively.

Ethics

The Ethics section of the structural metadata is important for indicating any sensitivities related to the data that could have potential ethical implications. This includes identifying any personally identifiable information (PII), geospatial data, or questions that may be sensitive in nature.

The sensitivity field provides information about the type of sensitivity that may be present in the data. This may include PII, geospatial data, or sensitive questions related to gender, politics, crime, or corruption. If the data contains PII, the PII field indicates whether it is direct, indirect, or not PII.

The sensitive question category field allows the user to identify any specific types of questions that may be sensitive in nature. This information can be helpful for understanding the potential implications of the data and for ensuring that the data is handled appropriately.

The SensitivityDescription field provides a brief description of the sensitivity and why it may be important to consider in the analysis or use of the data.

The geospatial field indicates the level of granularity of the geospatial data in the dataset. This information can be important for understanding any potential privacy concerns related to the data.

The derived data field indicates whether the data is derived from the original raw or cleaned data. If the data is derived, the RDM_protocol and RDM_method fields provide information on the specific protocols and research data management methods used to create the derived dataset.

Finally, the LinkedVariableIDs field allows users to identify any linked variable IDs that may be associated with the data. This information can be helpful for understanding the relationships between different variables in the dataset and for facilitating data integration and interoperability.

Data quality

This sub-section provides information on data quality assessment, including fields for accuracy and precision. Additional fields can be added as needed to provide more information about data quality.

The Data Quality Assessment field can be used to document any relevant information about the quality of the data. This could include information about data cleaning procedures, data validation and verification methods, or any other relevant information about the data quality.

Accuracy and precision are two important measures of the quality of data.

Accuracy refers to how close a measurement is to its true value. In other words, it measures the extent to which the data represents the reality it is meant to represent. Accuracy is typically measured by comparing the data to a known standard or reference.

Precision, on the other hand, refers to how consistent and reproducible a measurement is. In other words, it measures the degree of variation within a set of data. A highly precise measurement is one where the values are clustered closely together and there is little variation between them.

Precision can be expressed in various ways, depending on the type of data being collected. For example, for decimal values, precision may be expressed in terms of the number of decimal places, while for integer values, precision may be expressed in terms of the number of significant digits.

In the metadata template, we suggest including a "precision" field for any numeric values, where the precision can be expressed in a way that is appropriate for the type of data being collected. For example, if the data includes measurements of length, the precision might be expressed in terms of the number of decimal places (e.g. "precision: 0.01 meters").

It's important to note that the precision value should always be defined in a way that is consistent with the measurement method and the level of accuracy of the instrument used to collect the data. If the

precision is not known or cannot be accurately determined, it's better to omit this field rather than providing inaccurate information (Michener et al 2017)

In data management, accuracy and precision are often used to describe the quality of data. Accuracy can be affected by various factors such as measurement bias, instrumentation errors, and data entry errors, while precision can be affected by factors such as sampling variability, instrument calibration, and data processing methods.

When defining the accuracy attribute in the metadata, it is important to specify the method used to measure accuracy, as well as any sources of potential error or bias. The precision attribute should specify the level of precision achieved in the data and any methods used to ensure consistency and reproducibility.

Overall, accuracy and precision are both important measures of data quality, and ensuring accurate and precise data is critical for producing reliable and useful research results.

In summary, the Structural Metadata sheet is used to provide detailed information about the variables in the dataset. By providing this metadata, researchers can ensure that the data is well-documented and easily understood, which can help to ensure the integrity and usefulness of the data for future research.

As with the other metadata sheets in the template, it is important to provide as much detail as possible in the Structural Metadata sheet, using the standardized vocabulary and formats provided in the template. This can help to ensure that the metadata is consistent and easily interpretable by others who may want to use the data in the future.

Controlled vocabularies

Controlled vocabularies are a set of predefined terms used to describe a particular concept in a consistent and standardized manner. In many datasets, controlled vocabularies are used to describe variables such as locations, socio-economic characteristics, and other relevant information. The use of controlled vocabularies can improve the consistency and accuracy of data across different studies, making it easier to compare and integrate datasets.

However, different studies may use different controlled vocabularies, which can hinder the interoperability and reusability of data. To address this issue, researchers can map the idiosyncratic controlled vocabularies in specific datasets to standard terms, which can increase the interoperability of data across different studies. This process of mapping can help to reduce ambiguity and ensure that data from different sources can be integrated and analyzed together.

In addition to facilitating data integration, the use of controlled vocabularies and mapping can also improve the efficiency of data analysis. By using standardized terms, researchers can reduce the time and effort required to clean, process, and analyze data. This can lead to more accurate and reliable results and facilitate the reproducibility of research.

Controlled vocabularies are an essential component of metadata and data management, as they help standardize the terms used to describe variables or categories, making them more interoperable and reusable across different datasets. In order to facilitate the use and mapping of controlled vocabularies, we use the following approach. Consisting of two linked tables. The first table is a list of controlled vocabularies used in the dataset. The second is a list of controlled vocabulary codes or terms.

The controlled vocabulary list is a sheet that contains information about the different controlled vocabularies used in the data set. Each vocabulary is identified by a unique vocabulary ID, and is described in the vocabulary Description column. The vocabulary Source column provides information about the source of the vocabulary, such as a codebook or external reference. If a mapping exists of the vocabulary to a standard vocabulary, the standard used to map is indicated in the Standard Used to Map column and the URL of the mapping file is provided.

The controlled vocabulary term list is a sheet that contains information about the different terms used in the controlled vocabularies. Each term is identified by a vocabulary ID and a code or identifier. The description column provides a brief description of the term. If applicable, the ontology or standard that the term is associated with is provided in the Ontology or Standard column, along with the specific ontology term that the controlled vocabulary term maps to and the URI of the ontology term.

By organizing controlled vocabularies and providing a way to map them to standard terms, data managers can increase the interoperability and reusability of data across different datasets. The use of standard terms can also help with data integration and comparison across different studies, making it easier for researchers to build on previous work and conduct meta-analyses.

To create a table and list of controlled vocabularies, data managers can use spreadsheet software, such as Microsoft Excel or Google Sheets. The table and list can be included in the metadata for the dataset, making it easily accessible to anyone who wants to use or reference the data.

In summary, the use of controlled vocabularies and mapping can improve the consistency, accuracy, interoperability, and reusability of data across different studies. These benefits can lead to more efficient and effective research and analysis, ultimately advancing the field of science.

Controlled vocabulary list

This sheet contains a list of controlled vocabularies used in the dataset, along with their corresponding vocabulary ID and other relevant information. Each row represents a specific controlled vocabulary, and the columns provide information about the vocabulary's ID, description, source, standard used to map (if applicable), and URL of the mapping file (if applicable).

Vocabulary ID: A unique identifier for the vocabulary.

Vocabulary Description: A brief description of the vocabulary.

Vocabulary Source: The source of the vocabulary, such as a codebook or external reference.

Standard Used to Map: If a mapping exists of the vocabulary to a standard vocabulary which standard is it.

URL of Mapping File: Provide the URL of the mapping file (if applicable).

Controlled vocabulary terms

This sheet contains a list of terms used in the dataset, along with their corresponding controlled vocabulary ID and other relevant information. Each row represents a specific term, and the columns provide information about the term's code or identifier, description, ontology or standard (if applicable), ontology term (if applicable), and URI of the ontology term (if applicable).

Code: The code or identifier for a specific term within the controlled vocabulary.

Description: A brief description of the term.

Ontology or Standard: If applicable, the ontology or standard that has a term that the controlled vocabulary term can be associated with.

Ontology Term: If applicable, the specific ontology term that the controlled vocabulary term maps to.

Ontology Term URI: If applicable, the URI of the ontology term.

Final remarks

The data collected with this template can be used to tag resources with metadata in accordance with One CGIAR standards. Moreover, by using this template, the data can be parsed easily and converted into an OIMS-compatible JSON file for enhanced interoperability and reusability (Wilkinson et al. 2016)

Metadata plays a crucial role in ensuring that research data is discoverable, accessible, and reusable over time. As data is increasingly being recognized as a valuable resource that can be used for multiple purposes, it is important to have proper metadata to accompany it. Metadata provides context and information about the data, including its purpose, structure, content, quality, and the methods used to create it. This information is essential for other researchers who want to use the data, as it allows them to understand the data and its potential limitations.

In addition, metadata is also important for data archiving, preservation, and sharing. By providing detailed information about the data, metadata helps to ensure that the data can be effectively stored, preserved, and shared over time. Metadata also allows for data reuse, as it enables other researchers to understand and use the data in new ways. This can lead to new discoveries and insights that would not have been possible without the proper metadata.

This user guide provides guidance on how to create metadata for research data. It includes instructions on the different types of metadata, such as descriptive, administrative, and structural metadata, as well as guidance on how to create each type of metadata. The guide also provides instructions on how to create a metadata template, which can be used to ensure that metadata is consistent and complete across different datasets.

Overall, the user guide emphasizes the importance of metadata in ensuring that research data is discoverable, accessible, and reusable over time. By following the guidance provided in this user guide, researchers can create effective metadata that will help to ensure the long-term usability and accessibility of their data.

References

Borgman, Christine L. 2012. "The Conundrum of Sharing Research Data." *Journal of the American Society for Information Science and Technology* 63 (6): 1059–78. https://doi.org/https://doi.org/10.1002/asi.22634.

Choudhury, G. S., & Shreeves, S. L. (2013). The importance of metadata in research data management and sharing. Journal of the Medical Library Association: JMLA, 101(4), 337–341. https://doi.org/10.3163/1536-5050.101.4.015

Dappert, A., Klar, J., & Lewis, S. (2017a). Metadata for research data: Current practices and trends. International Journal of Digital Curation, 12(1), 67-91.

Dappert, A., Dröge, E., Fraas, C., & Lautenschlager, M. (2017b). Technical metadata for digital preservation: A review of key standards and guidelines. D-Lib Magazine, 23(5/6), n.p. https://doi.org/10.1045/may2017-dappert

Devarakonda, R., Palanisamy, G., Greenberg, J., & Wilson, B. (2013). Technical metadata for digital preservation. Data Science Journal, 12, WDS31–WDS43. https://doi.org/10.2481/dsj.WDS-041

Drachsler, H., Verbert, K., Santos, O. C., & Manouselis, N. (2015). Pan-European survey on the uptake and impact of Open Educational Resources in adult education. Open Praxis, 7(4), 331–348. https://doi.org/10.5944/openpraxis.7.4.267

Duke, C. S., Porter, S. S., Bales, M. E., McClure, R. S., & Calvert, S. G. (2014). Metadata quality in institutional repositories: A field study. Information Research, 19(4), n.p. https://doi.org/10.47989/irisic2154

Ghosh, S. (2014). Data interoperability: Why it's essential, and how we can achieve it. Journal of eScience Librarianship, 3(2), e1062. https://doi.org/10.7191/jeslib.2014.1062

Koskela, M., Persson, A., & Mäkelä, M. (2016). Importance of metadata in data management and sharing: Practices and perspectives of research staff. Journal of Documentation, 72(5), 960-980. https://doi.org/10.1108/JD-02-2016-0022

Kruseman, Gideon. 2022. "A Flexible, Extensible, Machine-Readable, Human-Intelligible, and Ontology-Agnostic Metadata Schema (OIMS)." *Frontiers in Sustainable Food Systems* 6. https://doi.org/10.3389/fsufs.2022.767863 .

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. Ecological applications, 7(1), 330-342.

Michener, W. K., Jones, M. B., (2012). Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology & Evolution, 27(2), 85-93. https://doi.org/10.1016/j.tree.2011.11.016

National Science Foundation. (2015). Chapter II. C. 2. j. Data management plan. https://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp

Recker, J., Li, Y., Li, Y., & Li, Y. (2019). FAIR data management: The basics and beyond. Journal of Business Research, 100, 470-481.

Rugg, A., & McInerney, P. (2012). Structural metadata for digital content: A review of the research landscape. Journal of Documentation, 68(4), 520-534. https://doi.org/10.1108/00220411211244554

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. PLoS ONE, 6(6), e21101. https://doi.org/10.1371/journal.pone.0021101

United Nations. (2019). Archiving and preserving electronic records: A review of the international landscape. https://unpan1.un.org/intradoc/groups/public/documents/un/unpan051987.pdf

Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (March): 160018. http://dx.doi.org/10.1038/sdata.2016.18.