



Metadata Editor

User and Practice Guide

World Bank, Office of the Chief Statistician

Version: 1.0

Generated on: May 20, 2025

Prepared by: Office of the Chief Statistician

For online version, visit github.com/worldbank/metadata-editor-docs.

© 2025 World Bank. Licensed under [CC BY 4.0](#).

About

The Metadata Editor is an open-source application designed to streamline the documentation of data of diverse types, including statistical indicators, survey microdata, geographic raster and vector datasets, documents, images, videos, scripts, and more. It enables users to generate rich, structured metadata that complies with international metadata standards and schemas. Additionally, the application is designed for extensibility, allowing users to integrate new standards, schemas, and updated versions as they emerge.

The Metadata Editor is a tool used by data curators to generate rich and structured metadata. These metadata, typically saved as JSON or XML files, serve as input to data cataloguing and dissemination systems.

The Metadata Editor can be deployed as a stand-alone desktop application or as a server-based tool, supporting collaborative metadata curation within research institutions, statistical agencies, data repositories, and digital libraries. When operated on a server, it facilitates multi-user collaboration, version control, and centralized metadata management, enhancing the efficiency and consistency of documentation workflows.

Built with a modular architecture, the Metadata Editor is developed using PHP and leverages APIs for seamless integration with other data systems. It incorporates Python libraries such as Pandas and GeoPandas to support the ingestion of datasets in various formats, extract metadata automatically, and generate summary statistics, making metadata creation both efficient and precise.

Data types and metadata standards supported

The Metadata Editor supports the documentation of a wide range of structured data types, using established metadata standards and schemas:

- **For structured data**
 - **Microdata:** Unit-level data on individuals, households, facilities, establishments, or other entities, derived from surveys, censuses, administrative records, or sensors. The Metadata Editor supports documentation using the Data Documentation Initiative (DDI) Codebook standard (version 2.5); see <https://ddialliance.org/ddi-codebook>.
 - **Indicators and databases of indicators:** Summary measures derived from observed data, often stored as time series in databases. The Metadata Editor supports a bespoke World Bank metadata schema, built by compiling metadata elements used in multiple leading indicator databases.
 - **Geographic datasets and geographic data services:** Data describing geographic locations, boundaries, and earth surface characteristics, provided as raster or vector datasets or as web services. The Metadata Editor supports ISO 19139 (and related ISO 19110/19115) metadata standards. See <https://www.iso.org/standard/67253.html>
 - **Statistical tables:** Aggregated statistical information presented in cross-tabulations, such as those in statistical yearbooks and census reports. The Metadata Editor supports a specific metadata schema developed by the International Household Survey Network (IHSN).
- **For unstructured data**
 - **Text:** Collections of documents (e.g., books, reports, manuals) form corpora that can be structured using natural language processing (NLP). The Metadata Editor supports a schema combining elements from Dublin Core (<https://www.dublincore.org/>), MARC21 from the US Library of Congress (<https://www.loc.gov/marc/bibliographic/>), and BibTeX (<https://www.bibtex.org/>).
 - **Images:** Digital images can be analyzed with machine learning techniques for classification and object detection. The Metadata Editor provides two metadata schema options: Dublin Core with imageObject elements from

Schema.org and IPTC (<https://iptc.org/>). Video Recordings: Speech-to-text algorithms allow the automatic transcription of video and audio recordings, making them discoverable and analyzable as structured data. The Metadata Editor supports a schema incorporating elements from Dublin Core and videoObject from Schema.org (<https://schema.org/>).

- **Videos:** the Metadata Editor uses a metadata schema built on videoobject from schema.org.
- **Research projects and scripts:** Research projects and related scripts for data transformation, analysis, and visualization are essential for reproducibility and transparency. The Metadata Editor includes a dedicated schema for documenting research projects and scripts.

In addition to these metadata standards and schemas, the Metadata Editor allows exporting metadata templates as SDMX Metadata Structure Definitions and the metadata themselves as SDMX metadatasets. An option to export metadata to schema.org (<https://schema.org/>), Croissant (<https://github.com/mlcommons/croissant>), and DCAT (<https://www.w3.org/TR/vocab-dcat-3/>) is also provided for some types of data.

The technical documentation of the metadata standards supported by the Metadata Editor is available at <https://worldbank.github.io/metadata-schemas/> where structures and information on each metadata element are provided.

Licenses and disclaimer

Software license

The Metadata Editor is published as open-source software under the MIT License (<https://opensource.org/license/mit>) with additional terms for international organizations:

MIT LICENSE

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

ADDITIONAL TERMS FOR INTERGOVERNMENTAL ORGANIZATIONS

Where the Licensor is an intergovernmental organization (IGO), the following additional terms shall apply:

1. **Privileges and Immunities.** Nothing in this License shall constitute or be interpreted as a waiver, express or implied, of any privileges and immunities that apply to the Licensor pursuant to international law, treaties, or the Licensor's constituent documents, including immunity from jurisdiction and execution.
2. **Dispute Resolution.** Any dispute arising under this License that involves the Licensor shall be resolved, to the extent not settled amicably, through the following procedures:

- First, by non-binding mediation conducted in accordance with rules designated by the Licensor, or, if none, the UNCITRAL Mediation Rules;
 - Failing settlement within 45 days, by final and binding arbitration in accordance with the UNCITRAL Arbitration Rules then in force, by a sole arbitrator, conducted in English, seated at the Licensor's headquarters, and conducted remotely when practicable.
- 3. Interpretation under International Law.** This License shall be interpreted in accordance with general principles of international law, including those reflected in the Berne Convention (1971), the WIPO Copyright Treaty (1996), and the Universal Copyright Convention (1971). No provision shall be interpreted in a manner that derogates from the legal status of the Licensor under international law.
- 4. No Waiver.** No provision of this License shall be deemed waived and no breach consented to unless expressly agreed in writing and signed by the Licensor.
- 5. No Ongoing Support or Maintenance.** Nothing in this License shall be construed to create an obligation on the part of the Licensor to provide updates, maintenance, support or any form of continued service for the software once released.
- 6. Entire Agreement.** This License constitutes the entire agreement between the Licensee and the Licensor concerning the Work and supersedes any additional terms submitted by You. No modification shall be effective unless agreed in writing by both parties.
- 7. Severability.** If any provision of this License is held to be unenforceable, it shall be interpreted to give effect to its purpose to the extent permitted by applicable law, and the remainder of the License shall remain in full force and effect.
- 8. Precedence.** In the event of any inconsistency between these Additional Terms for IGOs and the terms of the MIT License, these Additional Terms for IGOs shall prevail.

User Guide license

This User Guide is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0). See <https://creativecommons.org/licenses/by-nc-nd/4.0/>



Disclaimer

The Metadata Editor is provided "as is", without any warranties or guarantees, express or implied. The World Bank makes no representations regarding the accuracy, reliability, completeness, or suitability of the software for any particular purpose. Users assume full responsibility for the installation, configuration, and use of the application. The World Bank shall not be liable for any direct, indirect, incidental, consequential, or special damages arising from the use or inability to use the software. By using the Metadata Editor, users acknowledge and accept that they are solely responsible for ensuring that their use of the application complies with applicable laws, policies, and data governance requirements. Nothing herein shall constitute or be considered to be a limitation upon or a waiver of the privileges and immunities of any of the member institutions of The World Bank Group, which are specifically reserved.

Acknowledgments

The Metadata Editor application was developed by Mehmoond Asghar (Senior Data Engineer, World Bank, lead developer) with Olivier Dupriez (Deputy Chief Statistician, World Bank, project manager). The User Guide was written by Olivier Dupriez and Mehmoond Asghar.

The application was in part inspired by the Nesstar Publisher software application (by the Norwegian Social Science Data Archive).

The application benefited from feedback from many colleagues who reviewed and tested it.

Rationale and objective

The volume and diversity of data available to researchers, policymakers, and analysts are expanding rapidly. However, despite the proliferation of data resources, many valuable datasets remain underutilized due to challenges in discovering, understanding, and effectively using them. Without proper documentation, even high-quality data can be difficult to interpret, integrate, or trust. To fully realize the potential of data, it must be findable, interpretable, accessible, and reusable. High-quality metadata is essential to achieving these objectives.

Metadata is structured information that describes, explains, locates, and facilitates the retrieval, use, and management of data resources (National Information Standards Organization, 2004). It serves as the foundation of data discovery, enabling search engines, cataloging systems, and knowledge management platforms to efficiently index and retrieve datasets. Beyond discovery, metadata enables users to assess the relevance, quality, and fitness-for-purpose of a dataset before investing time and resources in its use. Well-structured metadata is also essential for making data AI-ready, as it enhances machine interpretability and supports automation in data processing, integration, and analysis.

However, producing high-quality metadata is a complex task. Data repositories, statistical agencies, research institutions, and other data-producing organizations face challenges in maintaining consistency, interoperability, and completeness in metadata documentation. Adopting internationally recognized metadata standards and schemas is critical to ensuring metadata quality. These standards—developed by expert communities for different data types—enable structured, detailed, and interoperable metadata that enhances data usability. Metadata compliance with established standards also supports interoperability across data management and dissemination systems, streamlining data integration and exchange.

To address these challenges, the **Metadata Editor** provides an open-source solution for documenting diverse data types using internationally recognized metadata standards (metadata schemas). Developed by the World Bank, the Metadata Editor is designed to:

- **Ensure flexibility** – allowing users to develop and customize metadata templates tailored to specific needs, including domain-specific controlled vocabularies.
- **Enhance usability** – providing an intuitive interface that supports both technical and non-technical users in metadata creation and management.
- **Foster collaboration** – supporting multi-user workflows with advanced permission settings in its web-based version, enabling teams to work efficiently on metadata documentation.
- **Improve data quality and interoperability** – ensuring that metadata aligns with established standards, facilitating data integration across systems, and supporting data governance and quality assurance frameworks.
- **Prepare data for the AI era** – enabling structured, machine-readable metadata that improves the discoverability, interpretability, and automation of data processing, making datasets more useful for AI applications and advanced analytics.

By streamlining metadata creation and ensuring adherence to international standards, the Metadata Editor aims to contribute to the modernization of data dissemination. It enables organizations to provide and disseminate credible, high-quality, and accessible data.

Introduction to metadata standards

Data producers who seek to ensure that their data are credible, discoverable, visible, and usable must provide data users with rich and structured metadata—information that elucidates the context, quality, and characteristics of the underlying data. Comprehensive and structured metadata enhance the trustworthiness of the data but also enable advanced search functionalities that empower users to locate, assess, utilize, and repurpose the data effectively and responsibly.

Metadata: definition and types

Metadata refers to information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage data. It may also include information that describes how data should be published, managed, and displayed.

Metadata can be categorized into several types, each serving a specific purpose in enhancing data transparency, discovery, management, and usability.

- **Metadata intended to be shared with data users**
 - **Cataloguing metadata** is used to uniquely identify and differentiate datasets within a collection or catalog. It includes key elements such as the dataset's title, unique identifier, version number, and publication date. Cataloguing metadata functions as a bibliographic record, enabling efficient dataset discovery, retrieval, and tracking within a repository or catalog.
 - **Descriptive and reference metadata** provide complementary details about a dataset's context, content, and creators, aiding users in understanding its origin, purpose, and applicability. *Descriptive metadata* focuses on identifying and describing the dataset, including elements such as title, creators and contributors (identifying individuals or organizations involved in dataset production), abstract, description, keywords, topics covered, and geographic and temporal coverage. *Reference metadata* offers deeper context about the dataset's creation, including the rationale for its collection, methodologies used for data collection and processing, and quality-related information. Together, descriptive and reference metadata enable secondary users to assess the dataset's relevance, understand its scope, and ensure its appropriate use.
 - **Structural metadata** defines the organization and format of the dataset, and the relationships between its components. Structural metadata ensures that datasets can be effectively analyzed and integrated into workflows. Structural metadata may include a *data dictionary* for microdata (providing a detailed list and description of data files and variables), a *data structure definition* outlining the modeling and structure of the data for indicators, or a *feature catalog* for geographic datasets.
- **Metadata intended for catalog and system administration**
 - **Administrative metadata** provides essential information for managing datasets, including details about their origin, acquisition date, and access rights (e.g., licensing and permissions). It defines the terms of use, specifying who can access the data, for what purposes, and under what conditions. Additionally, administrative metadata includes operational details necessary for IT staff to manage the dataset effectively, such as file formats, storage requirements, and system dependencies. This metadata type ensures that datasets are managed securely, efficiently, and in compliance with access policies.

Structured metadata

Structured metadata refers to metadata that is organized in a predefined and consistent format, often using standardized structures or schemas to ensure uniformity and interoperability.

Structured metadata helps maximizing the utility of data assets, as its predictable format enables advanced functionalities that are otherwise difficult to achieve:

- **Supporting the development of advanced search and discovery tools**, allowing users to more efficiently locate and explore relevant datasets based on various parameters. It allows the development of advanced search tools and enables the optimization of search results ranking algorithms (e.g., by boosting the importance of selected components of the metadata).
- **Facilitating seamless interaction** through application programming interfaces (APIs), enabling integration with external applications and workflows.
- **Improving the effectiveness of quality assurance processes** by standardizing how data descriptions are evaluated for completeness, consistency, and accuracy.
- **Enhancing interoperability across systems**. Structured metadata enables automated harvesting and extraction of information across different platforms, fostering connections and relationships between assets stored in various systems. This interconnected approach improves data accessibility, consistency, and integration.
- **Supporting long-term data preservation**, as its clear organization aids in maintaining accessibility and usability over time.

Structured metadata contains *metadata elements*, each representing a specific piece of information about a dataset (metadata elements can for example be the dataset title or geographic coverage, producer, a variable name or variable label, etc.) Elements within all metadata standards are organized into groups and sub-groups. For instance, the DDI Codebook metadata standard, which is utilized for documenting microdata, comprises four primary sections: document description (`doc_desc` ; metadata about the metadata), study description (`study_desc` ; details regarding the study itself), file description (`data_files` ; information pertaining to each data file), and variable description (`variables` ; data at the level of each variable).

```
+ "doc_desc": { ... },
+ "study_desc": { ... },
+ "data_files": [ ... ],
+ "variables": [ ... ],
```

The *study description* section of the DDI Codebook includes sub-groups of elements that document the authorship of the study (`authoring_entity`), the processes involved in implementing the study (`study_development`), and other relevant aspects.

```
- "study_desc": {
  + "title_statement": { ... },
  + "authoring_entity": [ ... ],
  + "oth_id": [ ... ],
  + "production_statement": { ... },
  + "distribution_statement": { ... },
  + "series_statement": { ... },
  + "version_statement": { ... },
  + "bib_citation": "string",
  + "bib_citation_format": "string",
  + "holdings": [ ... ],
  + "study_notes": "string",
  + "study_authorization": { ... },
  + "study_info": { ... },
  + "study_development": { ... },
  + "method": { ... },
  + "data_access": { ... }
},
```

Each metadata element within a metadata standard or schema encompasses the following attributes:

- **Key:** This denotes the standardized name assigned to the element and must remain constant (the key cannot be edited or translated).
- **Type:** Metadata elements can be various types like text (string), numbers (numeric), arrays, or Booleans (logical). If an element is an array, its sub-elements can be any of the other types. Arrays are used for elements with sub-elements, like a country element including a name and a code. The type is defined by the standard and cannot be altered.
- **Label:** A brief title for the element. While each element or sub-element has a default label, which can be modified to fit an organization's specific terminology.
- **Repeatable status:** Shows if the element can accept multiple entries. For instance, the "nation" element used to describe the geographic coverage of a dataset is *repeatable* since a dataset can possibly cover multiple countries, but the "title" element is *not repeatable* as a dataset is expected to have one and only one primary title. The status of an element is defined in the metadata standard and cannot be altered.
- **Required status:** Specifies whether the element is *required* (i.e., mandatory) or *optional*. This indicates whether metadata for a dataset that does not have content for this element should be validated or not. If the element is *required*, metadata with no content for this element will result in a schema validation error. A "recommended" status may also be included. Although it is advised to populate as many metadata elements as feasible, standards should allow for incomplete metadata. For that reason, very few elements are usually marked as "required". Required elements typically include a unique identifier (preferably a global unique identifier like a Digital Object Identifier/DOI), the dataset title, the authoring entity, and the geographic coverage.
- **Description:** Every element in a metadata standard comes with a default description which serves as instructions to data curators. These descriptions and guidelines can be edited as needed.
- **Controlled vocabularies:** *Controlled vocabularies* consist of pre-defined code lists which define the list of acceptable values for a metadata element. Using controlled vocabularies fosters consistency and coherence in metadata. Controlled vocabularies may be part of some metadata standards, or (more frequently) provided in *metadata templates* specific to an organization.

Metadata standards

A metadata schema or metadata standard comprises an organized set of clearly defined metadata elements designed for documenting a dataset, accompanied by rules and instructions to ensure their uniform and consistent implementation. Adopting metadata standards represents a pragmatic and efficient approach to foster the completeness, usability, discoverability, and interoperability of the metadata. In addition to the benefits provided by structuring metadata, the adoption of standardized metadata structures offers the following advantages:

- **Completeness of metadata:** Metadata standards and schemas provide exhaustive checklists of required, recommended, and optional information for documenting datasets. Adherence to these standards ensures that essential information is systematically included, mitigating the risk of oversight.
- **Interoperability:** Metadata standards promote interoperability among data catalogs, enabling seamless information exchange through automated harvesting and synchronization mechanisms.

Each primary type of development data (microdata, indicator, geographic dataset, and others) uses a specific metadata standard. The Metadata Editor supports the following types of data: microdata, indicators (or time series) and databases of indicators, geographic datasets, publications related to data products (documents), images, videos, and scripts used for data processing, editing, analysis, and other data transformations. Metadata standards have been independently developed for each data type. Ideally, common elements among various standards (such as those used to capture the title and identifier of the documented resource) would be consistently defined. However, such consistency is not always guaranteed as some metadata standards have been developed independently from each other.

Metadata standards supported by the Metadata Editor

The schemas or standards implemented in the Metadata Editor are the following:

| Data type | Standard |
|----------------------------------|--|
| Microdata | Data Documentation Initiative 2.5 (Codebook) |
| Geographic datasets and services | ISO 19110, ISO19115, ISO19119, ISO 19139 |
| Time series, Indicators | Custom-designed schema |
| Documents | Dublin Core Metadata Initiative (DCMI), MARC |
| Statistical tables | Custom-designed schema |
| Photos / Images | IPTC (for advanced use) or Dublin Core augmented |
| Videos | Dublin Core augmented with VideoObject from schema.org |
| Programs and scripts | Custom-designed schema |

The Metadata Editor can also export metadata to schema.org, Croissant, and DCAT formats, and to SDMX compliant metadata for indicators.

How to produce standard-compliant metadata?

Standard-compliant metadata can be produced and edited through three main approaches: **using a Metadata Editor** software application, **programmatically** using a language like R or Python, or through **other metadata-enabled software** applications. Each approach offers distinct advantages depending on the level of automation, user expertise, and the specific type of data being documented.

Using a metadata editor

A metadata editor is a specialized software application designed to facilitate the creation, editing, and management of structured metadata in a user-friendly manner. This approach is particularly beneficial for non-programmers and for data types that require significant manual curation, such as microdata.

The World Bank Metadata Editor provides an intuitive environment for documenting various types of data, including indicators, geographic datasets, microdata, and scripts. Users begin by selecting the relevant data type and choosing from predefined metadata templates based on internationally recognized standards. The Metadata Editor then automatically generates metadata entry forms tailored to the chosen template, allowing users to input metadata systematically.

Key features of the Metadata Editor include:

- **Template-based metadata entry** – ensuring standardization and consistency.
 - **Support for multiple output formats** – metadata can be saved in standard-compliant JSON or XML formats, ready for integration into data catalogs or conversion to PDF, HTML, or other formats.
 - **Automated metadata extraction** – for certain data types, such as microdata and geographic datasets, metadata embedded within data files (e.g., variable names, variable labels, and value labels in microdata files, or features in geographic vector datasets) can be automatically extracted, significantly reducing manual effort of generating detailed metadata.
 - **Collaboration and multi-user workflows** – enabling teams to work on metadata documentation efficiently.
 - **Option to lock and version metadata** – allowing implementation of quality assurance and governance for the review and use of the metadata.
-

Programmatically

Metadata can be generated programmatically using R or Python (or similar language). Metadata files are JSON or XML files which correspond closely to R lists or Python dictionaries structured. The programmatic approach offers significant advantages for users with programming expertise, including:

- **Automation and scalability in metadata creation** – metadata generation can be scripted, making it highly efficient for large datasets or frequently updated data.
- **Automation in metadata maintenance** – the programmatic approach allows metadata to be edited in a dynamic, scripted manner.
- **Integration with data pipelines** – programmatic metadata creation allows seamless embedding into existing data processing workflows.

- **Machine-learning-based metadata enrichment** – advanced techniques, such as natural language processing (NLP) and machine learning, can be applied to automatically extract, enhance, or classify metadata elements.

Metadata generated programmatically can be saved as JSON or XML, or published in various applications. They can be published directly in the Metadata Editor, which can then serve as a central repository for the metadata.

To be compatible with the Metadata Editor, the metadata must be generated in compliance with the schemas outlined in [ReDoc URL].

Other metadata-enabled software applications

Several data management and survey processing tools integrate metadata standardization features. While these applications typically generate only a subset of required metadata elements, they offer valuable automated metadata extraction capabilities.

Examples include:

- **Survey Solutions (World Bank)** – allows exporting metadata compliant with the DDI Codebook metadata standard for survey data.
- **CsPro (U.S. Census Bureau)** – includes a function for exporting structured metadata, primarily focused on file- and variable-level documentation.
- **QGIS** - Exports metadata for geographic datasets, compatible with the ISO19139 standard.

While these tools generate essential metadata elements, they may not provide complete metadata coverage needed for full documentation. The metadata exported from such applications can be further edited, enriched, and completed using the Metadata Editor, ensuring full compliance with metadata standards.

Choosing the right approach

The choice of metadata generation method depends on the nature of the data, user expertise, and the level of automation required:

- For non-technical users or for datasets requiring manual curation, the Metadata Editor provides an accessible, structured environment.
- For organizations managing large datasets or requiring automated workflows, programmatic metadata generation (or a combination of programmatic metadata generation and editing of the metadata in the Metadata Editor) is highly efficient.
- For users leveraging metadata from existing data management tools, the Metadata Editor can complement and enhance the metadata to ensure full documentation.

Installation

The Metadata Editor can be installed on a server or on a stand-alone personal computer (PC). For installation on a PC, see the section *Installation on a personal computer*.

Installing the Metadata Editor on a web server provides several advantages:

- **Collaboration:** Multiple users can work on metadata simultaneously.
- **Scalability:** Allows for efficient handling of larger datasets and multiple concurrent users.
- **Accessibility:** Users can access the application from any device with a browser.
- **Centralization of metadata:** If the Metadata Editor runs on an organization's server (typically, on its intranet), all metadata will be stored in a central location. the Metadata Editor can therefore operate as a corporate metadata repository.
- **Permission control:** Running the application on a server allows configuration of an organization's user authentication system.
- **Security:** Centralizing the generation and storage of the metadata on a server facilitates the management of backups and implementation of other security measures.

Installation on a server

Hardware requirements

Ensure your server meets the following minimum specifications:

- CPU: Dual-core processor or higher
- RAM: At least 4 GB (8 GB recommended for larger datasets)
- Storage: Minimum 20 GB of free disk space

Software requirements

To run the Metadata Editor, you will need:

OS:

- Windows
- Linux
- MacOS

Web server:

- Apache 2.4 or later
- IIS 6/7.x or later
- NGINX

- PHP: Version 8.1 or later, with the following required extensions:
 - xsl
 - xml
 - mbstring
 - mysqli

Database:

- MySQL 8.x or MariaDB. The database is not provided in the Metadata Editor package. The selected database application must be downloaded from its respective repository or website.

Python:

- Python 3.12 or later is required for running the FastAPI backend for data import/export and generating summary statistics for Stata, SPSS, and CSV files.

Downloading and installing the Metadata Editor

Required components

To install the Metadata Editor, you need:

- Metadata Editor: PHP application with a MySQL/MariaDB database.
- PyDataTools: Python backend API for data import/export and summary statistics.

Folder structure

After installation, your directory structure should look like this:

```
metadata_editor
|
+--editor
+--pydatatools
```

Downloading the source code

- *Option 1: Using Git*

Navigate to the web server directory where you want to install the project and run:

```
$ mkdir metadata_editor
$ cd metadata_editor

$ git clone https://github.com/ihsn/editor
$ git clone https://github.com/mah001/pydatatools
```

- *Option 2: Using Zip Packages*

Download the zipped packages and extract them to create the required folder structure:

- Metadata editor: <https://github.com/ihsn/editor/archive/refs/heads/main.zip>
- PyDataTools: <https://github.com/mah0001/pydatatools/archive/refs/heads/main.zip>

Configuring the database

Step 1: Create Database and User

Use command line, PHPMyAdmin or any other database client tool to connect to your database server.

To connect via command line:

```
$ mysql -u root -p
```

Create database and user account:

- Create a new database (e.g., metadata_editor).

```
CREATE DATABASE metadata_editor;
```

Create a database user with a secure password.

```
CREATE USER 'editor_user'@'localhost' IDENTIFIED BY 'replace-this-with-password';
```

Grant this user access to the metadata_editor database.

```
GRANT ALL PRIVILEGES ON metadata_editor.* TO 'editor_user'@'localhost';
FLUSH PRIVILEGES;
```

Step 2: Update Database Configuration

- Navigate to the `editor/application/config/` folder.
- Copy or rename `database.sample.php` to `database.php`.
- Open `database.php` in a text editor and update the following fields:
 - `hostname` - ip address or the machine name where database is hosted
 - `username` - database user name
 - `password` - database password
 - `database` - database name

php

```
$db['default'] = array(
    'dsn' => '',
    'hostname' => 'localhost',
```

```
'username' => 'nada_user',
'password' => '<db-pass-here>',
'database' => 'metadata_editor',
'dbdriver' => 'mysqli',
'dbprefix' => '',
'pconnect' => FALSE,
'db_debug' => FALSE,
'cache_on' => FALSE,
'cachedir' => '',
'char_set' => 'utf8',
'dbcollat' => 'utf8_general_ci',
'swap_pre' => '',
'encrypt' => FALSE,
'compress' => FALSE,
'stricton' => FALSE,
'failover' => array(),
'save_queries' => TRUE,
'prefix_short_words'=>TRUE
);
};
```

Save the file.

Setting folder permissions

Run the following commands to set read/write permissions for the folders where the data will be stored:

```
$ chmod -R 775 datafiles files logs
```

Running the installer

- Open a web browser and navigate to the Editor installation URL. For example: <http://your-domain/editor-folder-name>, or <http://localhost/editor-folder-name>.

Metadata Editor Installer

Server information

PHP version: **8.4.5**
 DB version: **8.4.4 - connection was successful!**
 Web server: **Apache/2.4.63 (Unix) PHP/8.4.5 OpenSSL/3.4.1**

Required PHP Extensions

| Extensions | Enabled |
|---------------------|------------|
| xsl | yes |
| xml | yes |
| simplexml | yes |
| xmlreader | yes |
| gd (optional) | yes |
| zip (optional) | yes |
| mbstring (optional) | yes |
| mysqli | yes |

Other PHP.INI Settings

| Setting | Value | Recommended |
|---------------------|---------|--|
| file_uploads | Enabled | Enabled |
| post_max_size | 2000M | 800M |
| upload_max_filesize | 2000M | 800M |
| date.timezone | UTC | See how to configure and select the right timezone |

Folder READ/WRITE/DELETE permissions

| Folder | Read/Write | Delete |
|---------------------|------------|------------|
| Catalog (datafiles) | yes | yes |
| Log (./logs/) | no | no |

[Install Database](#)

- Check that all settings are marked with a green tick and fix any that are not on your webserver before running the installer.
- Click on the [Install Database](#) button and complete the form to create an initial Site Administrator account.

Metadata Editor Installer

Create administrator account

Database and Tables were created successfully!

First Name*

Last Name*

Email Address*



Password*



Password Confirm*



Create account

⚠ Note: Use a complex password (at least 12 characters, including uppercase, numbers, and special characters) to enhance security.

Installing and configuring PyDataTools (Python/FastAPI)

- **Step 1: Install Python** Download and install Python 3.12 from <https://www.python.org/downloads/>.
- **Step 2: Install dependencies** Navigate to the pydatatools folder and run:

```
$ pip install -r requirements.txt
```

python

- **Step 3: Run the FastAPI service** To start the FastAPI service, run:

```
$ nohup python -m uvicorn main:app --reload --host 0.0.0.0 --port 8000 &
```

```
INFO:     Will watch for changes in these directories: ['/Volumes/webdev/pydatatools']
INFO:     Unicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
INFO:     Started reloader process [32013] using StatReload
INFO:     Started server process [32037]
INFO:     Waiting for application startup.
Starting FIFO worker
INFO:     Application startup complete.
```

The Metadata Editor and PyDataTools should now be operational.

Email configurations

For the Metadata Editor to function correctly it is important that this step be completed.

Many of the functions within the Editor, such as registration, forgot password require that the Editor be able to send emails to users.

The email settings are stored in the config file `application/config/email.php`. The following settings are required to configure email:

```
$config['useragent']      = 'PHPMailer';
$config['protocol']       = 'smtp';
$config['smtp_host']      = 'outlook.office365.com';
$config['smtp_auth']       = true;
$config['smtp_user']       = 'your-email-address@outlook.com'; //email or username
$config['smtp_email']      = 'your-email-address@outlook.com'; //email address to send from
$config['smtp_pass']       = 'your-email-account-password';
$config['smtp_port']       = 1025;
$config['smtp_crypto']     = 'tls';
```

- **SMTP_HOST:** Your SMTP server host name
- **SMTP_AUTH:** Set it to true if your SMTP server requires authentication (username and password)
- **SMTP_USER:** Email address or user name. If your email server require a username for authentication, type the username, otherwise fill in with the email address
- **SMTP_EMAIL:** Email address same as in SMTP_USER
- **SMTP_PASS:** Email password, if required. Otherwise, leave it empty
- **SMTP_PORT:** Port number used by your email server. e.g. 25, 587, 443
- **SMTP_CRYPTO:** Encryption to use. Options are `tls`, `ssl` or leave it empty if none is required

Once you have updated the configurations, save the file.

Test the email settings

The quickest way to test if the email settings are working is to use the "forgot password" option from the user login page.

- If you are already Logged in, log out and then go to the forgot password page.
- Enter the administrator or any other accounts email address that you know have an account in the Editor.

- If the email failed, you will see an error message indicating that the email was not sent.

Test email settings using site admin tool

The Editor site administration includes a page to test email settings. Open the site administration, go to `settings` menu and open the last option `SMTP settings` and click on `Test email configurations`.

If you have filled in the email configurations (`application/config/email.php`), the page will read all settings from there. Otherwise, fill the email settings and press the button `Send email` to see if your settings are correct. The page will print detailed messages for both success and failed emails.

Enable backups

The Metadata Editor will operate as a central metadata repository. For a production level installation, it is therefore essential that you implement a proper backup system, with automatic backups as appropriate.

User login and profile

Access to the Metadata Editor is managed by the system administrator(s) within your organization. User accounts can be created individually or by configuring the organization's authentication system to automatically register all members as authorized users (using a Single Sign-On - SSO).

Logging In and Out

To access the Metadata Editor (login):

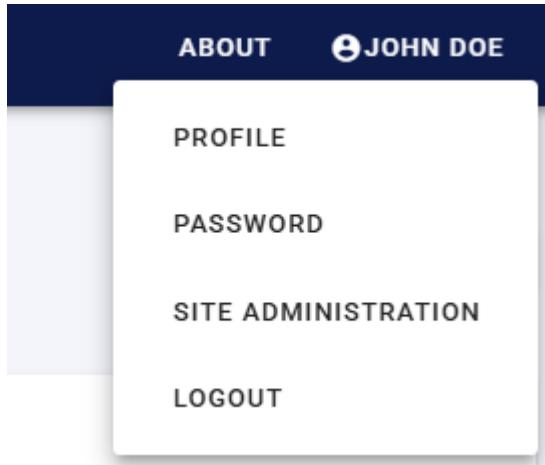
- Open the application URL provided by your organization.
- Enter your email or username and password to log in.

To logout:

Click on [Logout](#). Note that you will be automatically logged out after [X minutes] of inactivity.

Editing your user profile, password, and API key

After logging in, click on your username in the top menu to open a dropdown menu with the following options:



- [Profile](#) and API keys

Clicking on [Profile](#) opens a page displaying your profile details with an option to edit them.

The header bar is dark blue with white text. It features the "Metadata Editor" logo on the left, followed by "About" and "English" with a dropdown arrow. On the right, there's a user icon and "John Doe" with a dropdown arrow.

Profile

| | |
|---------------|----------------------|
| John Doe | Edit |
| Name | John Doe |
| Email Address | johndoe@ihsn.org |
| Company | |
| Phone | |
| Country | - |

API keys

[Generate API key](#)[Delete](#)

You can also generate an API key from this page. An API key is required only if you plan to manage metadata programmatically using the Metadata Editor's API with a programming language such as R or Python.

⚠ Important: Keep your API key confidential. It is equivalent to a password to which your profile and permissions are associated. **If you suspect that your API key has been compromised, delete it immediately and generate a new one.** Refer to the section *The Metadata Editor API* for more details.

- [Password](#)

Opens a page where you can change your password.

- [Site administration](#)

Available only to users with administrator credentials. Opens a dashboard displaying information about the projects you can access.

Quick start: Overview

The Quick Start examples provide a fast way to familiarize yourself with the Metadata Editor and experiment with creating standard-compliant metadata. Designed as step-by-step walkthroughs, these examples guide you through the core features of the application, allowing you to practice metadata creation with pre-provided sample files. Each example focuses on a specific data type supported by the Metadata Editor. We provide a dedicated Quick Start example for each supported data type:

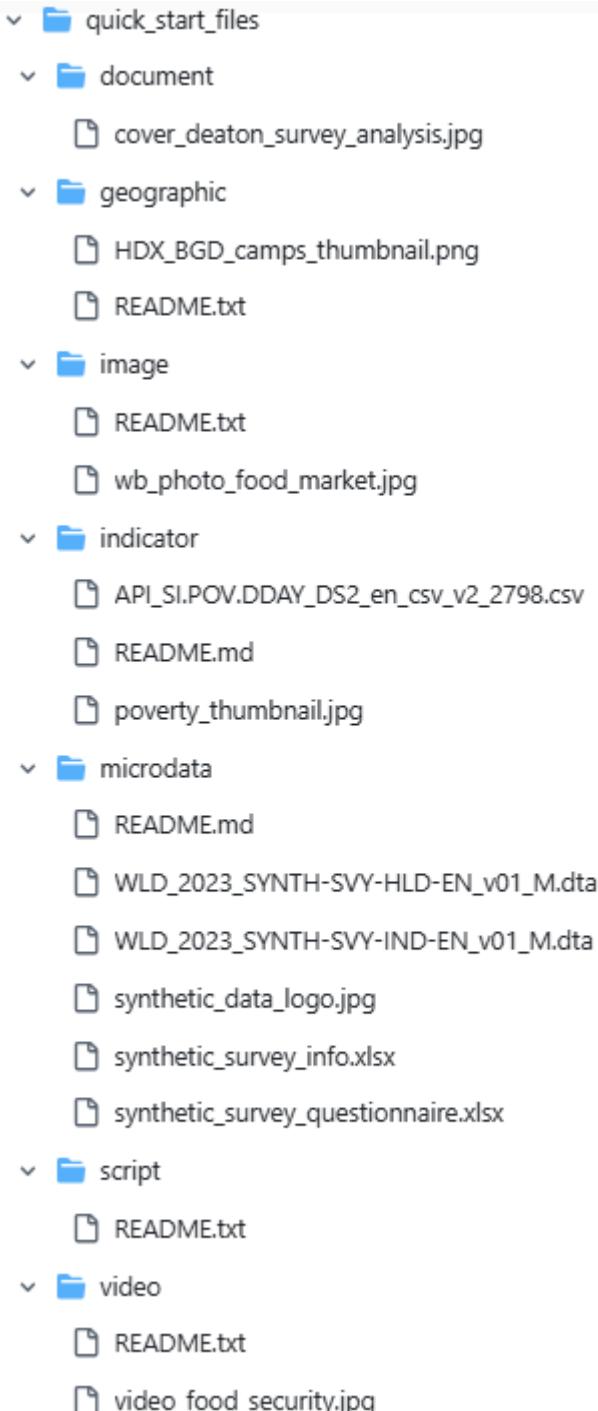
- **Document** – Documenting a publication (in this case, a book).
- **Microdata** – Documenting a survey dataset (a synthetic survey dataset composed of a household-level and individual data files provided in Stata format).
- **Indicator and database** – Documenting a statistical indicator and its source database.
- **Geographic dataset** – Documenting a vector dataset (provided in shapefile format) and a raster dataset (provided in geotiff format).
- **Scripts** – Documentation of a research project with multiple related scripts.
- **Image** – Documenting a digital photograph from a World Bank photo album.
- **Video** – Documenting a video.

These examples are intentionally simplified to provide an overview of the application's core functionalities and do not cover all available features. For a more comprehensive understanding, refer to the *Documenting Data* chapters. All necessary materials for replication are made openly accessible from the Metadata Editor Github repository.

If you have an installed NADA catalog with administrator credentials, you can also practice publishing metadata in an online catalog. NADA is an open source cataloguing application developed by the World Bank. Metadata generated by the Metadata Editor can be directly published in NADA. Refer to the NADA User Guide for instruction on how to install NADA and obtain a NADA API key that would allow you to publish metadata in it.

Before starting, please ensure that:

- You have the Metadata Editor installed on your system (server or PC), with an *Editor* role that allows you to create new projects.
- You have downloaded the required example files from ... Insert Download Link ... and saved them in a dedicated folder on your PC (the files do not have to be on the server that hosts the Metadata Editor). If you unzip the file provided on GitHub, you will have a folder with the following structure (which is not imposed by the Metadata Editor; all you need is access to the files):



- You have access to a test NADA catalog with the necessary credentials to publish metadata (this is only needed if you want to experiment publishing metadata in a NADA catalog).

Once ready, follow the step-by-step instructions provided in each example to practice creating and publishing metadata using the Metadata Editor.

Quick-start: Document

In this example, we will document a book titled *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy* authored by Angus Deaton and published in 2019. This book is freely available from the World Bank's Open Knowledge Repository at <http://hdl.handle.net/10986/30394>.

In this Quick-start exercise, we assume that you want to publish information on the book in a data catalog, with a link to the World Bank's Open Knowledge Repository (i.e., we assume that you not plan to make the book directly available from your catalog). The only file you need to reproduce this Quick-Start example is the image file of the book's cover page (file .../quick_start_files/document/cover_deaton_survey_analysis.jpg), although you may use another image file of your choice.

This Quick Start section does not include detailed guidance on documenting publications. For comprehensive instructions, see the chapter **Documenting a publication or report**.

Step 1: Create a new project and add a thumbnail

To begin, open the Metadata Editor in your web browser (the URL is determined by where you installed the application), and log in with your username and password. The *My projects* page will be displayed, showing all projects you have previously created and those that have been shared with you by other data curators, if any. If you are using the application for the first time and no project has been shared with you, the project list will be empty.

Click on **CREATE NEW PROJECT** and select *Document* when prompted to indicate that the resource you will be documenting is a document (a publication or report).

Create new project

 Microdata

 Timeseries

 Timeseries database

 Document

 Table

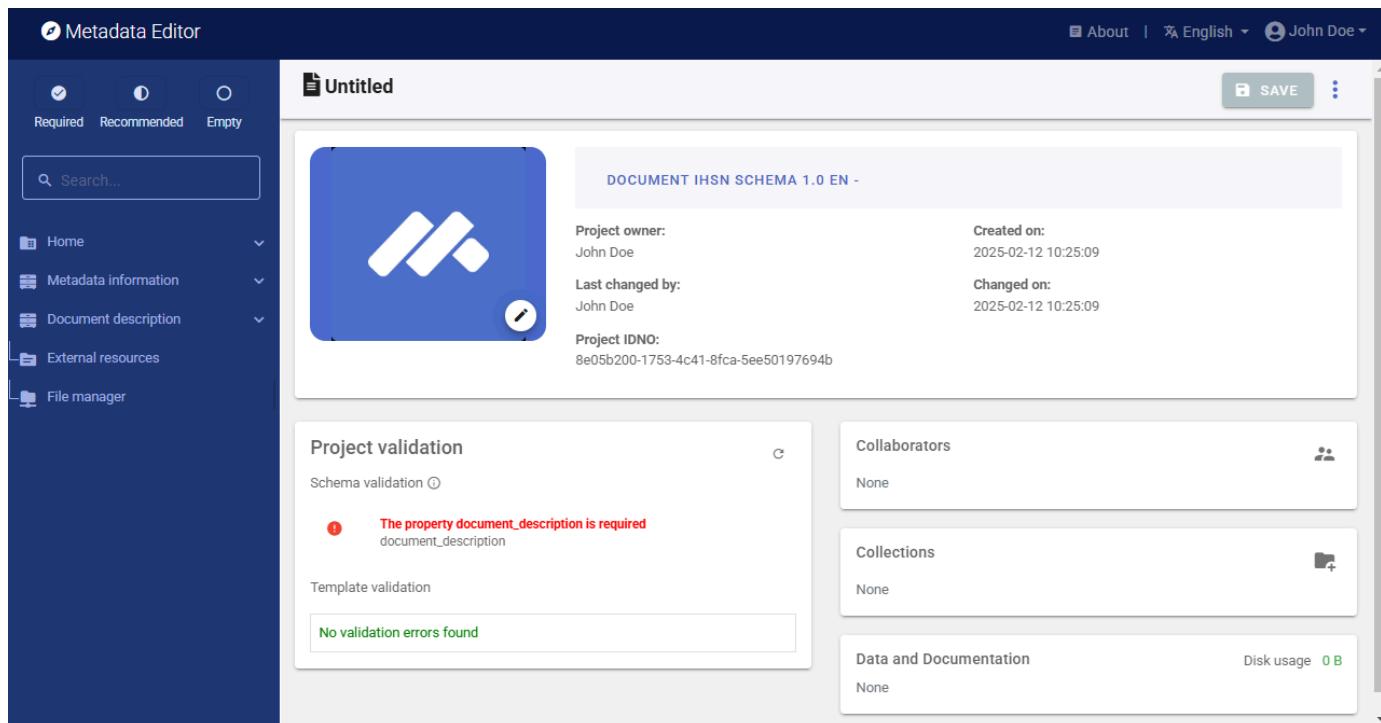
 Image

 Script

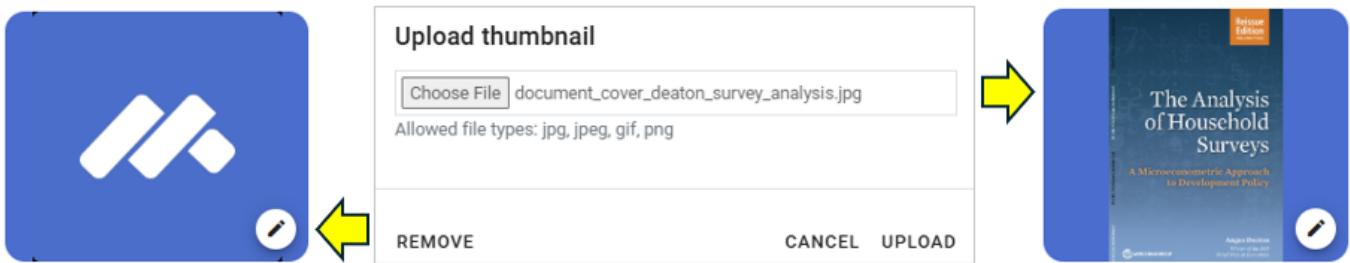
 Video

 Geospatial

A new project page will open in a new tab.



You will use the cover page of the book as a thumbnail. Note that providing a thumbnail is not required, but recommended. The thumbnail will be displayed in the Metadata Editor project list, and in the NADA catalog if the metadata is published in NADA. The cover page has been captured and saved as *cover_deaton_survey_analysis.jpg* (you can capture a document cover page with any screen capture tool, and save the image file in JPEG or PNG format). Click on the  icon in the screenshot image, and select the image file when prompted.



Documenting a dataset (or a document in this case) consists of entering metadata in metadata entry forms defined by a *metadata template*. When you create a new project, a default template is automatically selected. We will use this template, so there is no need to switch template.

Step 2: Enter metadata

In the navigation tree, select *Metadata information / Information on metadata* to enter information on who documented the publication and when. All information in this section is optional. Enter your name (as metadata producer) and the date of the day in ISO format YYYY-MM-DD (this is the date when the metadata, not the book, was produced). Then click on **SAVE**.

| Name | Abbreviation | Affiliation | Role |
|----------|--------------|-------------|------|
| John Doe | | | |

You can now start entering the metadata related to the book itself in the *Document Description* section. In the navigation tree, first select *Title statement* and enter the required **primary ID**, a unique identifier of your choice, e.g., JD_DOC_001 (if you want to publish the document in a NADA catalog, make sure that the identifier is not used by another user or for another project). Also enter the optional **Other identifiers** of the document, including the DOI (10.1596/ 978-1-4648-1331-3) and the ISBN (978-1-4648-1352-8). Then enter the **Title** of the book (*The Analysis of Household Surveys*) which is a required element, and its **Subtitle** (*A Microeconometric Approach to Development Policy*). Note that another option is to enter the title and subtitle as a single element under **Title**.

The Analysis of Household Surveys

Title statement

Primary ID * JD_DOC_001

Other identifiers

| Type | Identifier |
|-------------------|---------------------------|
| DOI | 10.1596/978-1-4648-1331-3 |
| ISBN (electronic) | 978-1-4648-1352-8 |

Title * The Analysis of Household Surveys

Subtitle A Microeconometric Approach to Development Policy

Alternate title

Translated title

Then proceed with the other sections in the navigation tree and fill out the following elements.

- **Author:** Angus Deaton
- **Publication date*** (in ISO format YYYY-MM-DD): 2019-01-16
- **Abstract:** Two decades after its original publication, The Analysis of Household Surveys is reissued with a new preface by its author, Sir Angus Deaton, recipient of the 2015 Nobel Prize in Economic Sciences. This classic work remains relevant to anyone with a serious interest in using household survey data to shed light on policy issues. This book reviews the analysis of household survey data, including the construction of household surveys, the econometric tools useful for such analysis, and a range of problems in development policy for which this survey analysis can be applied. The author's approach remains close to the data, using transparent econometric and graphical techniques to present data in a way that can clearly inform policy and academic debates. Chapter 1 describes the features of survey design that need to be understood in order to undertake appropriate analysis. Chapter 2 discusses the general econometric and statistical issues that arise when using survey data for estimation and inference. Chapter 3 covers the use of survey data to measure welfare, poverty, and distribution. Chapter 4 focuses on the use of household budget data to explore patterns of household demand. Chapter 5 discusses price reform, its effects on equity and efficiency, and how to measure them. Chapter 6 addresses the role of household consumption and saving in economic development. The book includes an appendix providing code and programs using STATA, which can serve as a template for the users' own analysis.
- **Language:** English (code EN)
- **Rights:** CC BY 3.0 IGO
- **Document type:** Book (select from drop down)
- **Keywords:** Enter one keyword per row: household survey; survey design; data collection; economic development; development policy
- **Topics:** Enter one topic per row and only fill in column Topic): Development Patterns and Poverty; Living Standards; Poverty Assessment; Poverty and Policy; Statistical & Mathematical Sciences

This information can be entered in the Metadata Editor template in the following elements:

| From World Bank | In the metadata template (Document description) |
|------------------|---|
| Publication date | Date / Date published |
| Author | Authors and contributors / Authors |
| Document type | Content description / Document type |
| Language | Content description / Language |
| Abstract | Content description / Abstract |
| Keywords | Content description / Keywords |
| Topics | Content description / Topics |
| Rights | Access and rights / Rights |

After entering all available information, click on [SAVE](#). Click on *Preview* in the navigation tree to view all information you have entered so far.

The screenshot shows the Metadata Editor interface. The left sidebar has buttons for Required, Recommended, and Empty, and a search bar. The navigation tree on the left includes Home, Preview, Metadata information, Document description (which is expanded), External resources, and File manager. The main content area shows the title 'The Analysis of Household Surveys'. Under 'Information on metadata', there is a table for 'Metadata producers' with one row for 'John Doe'. Below it, 'Production date' is listed as 2025-02-12. Under 'Document description', the 'Title statement' section contains a 'Primary ID' of JD_DOC_001. The 'Other identifiers' section lists a DOI of 10.1596/ 978-1-4648-1331-3 and an ISBN (electronic) of 978-1-4648-1352-8. A 'Title' field is also present.

Step 3: Provide a link to the document

The next step is to add a link (URL) to the book. Another option would be to upload the book in PDF; but in this exercise, we assume you do not want to distribute the book but only to list it in a catalog with a link to an external repository. This information, along with any other related files and links you may want to attach to the metadata, is referred to as *External resources*. To add external resources, click on *External resources* in the navigation tree and click on [CREATE RESOURCE](#).

The screenshot shows the 'External resources' section of the Metadata Editor. At the top right are 'SAVE' and three-dot buttons. Below is a search bar and a sidebar with categories: Home, Preview, Metadata information, Document description, External resources (which is selected), and File manager. The main area displays 'External resources' with '0 resources'. Buttons for 'CREATE RESOURCE' and 'IMPORT' are at the bottom right.

This will open a new resource page where you can describe the resource. Most elements are optional, but at a minimum, you should enter the **Resource type** (select *Document*, *Technical* from the drop down), the **Title** including the subtitle (*The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*), the **Author** (Angus Deaton), and the **Date** (the date of publication, in ISO format: 2019-01-16). To provide users with access to the book, add the external link (<http://documents.worldbank.org/curated/en/593871468777303124>) in **Resource attachment**.

The screenshot shows the 'Create new resource' form for 'The Analysis of Household Surveys'. The sidebar is identical to the previous screenshot. The main form has a 'Table of content' section with an empty text area, a 'Resource attachment' section with a file input field showing 'File: No file attached', and a URL input field containing 'http://documents.worldbank.org/curated/en/593871468777303124'. At the top right are 'SAVE*' and 'CANCEL' buttons.

Then click **SAVE**. The document will now be listed as an external resource.

With this, you have completed the documentation of the book. The *My Projects* page will show this new entry. You may at any time go back to it to edit or complete the metadata.

Step 4: Export and publish metadata

In the *Project* page, a menu of options is available to you.

| Project | Metadata |
|--|--|
| <input checked="" type="checkbox"/> Export package (ZIP) | <input checked="" type="checkbox"/> Apply default values from template |
| <input type="checkbox"/> Export JSON | <input type="checkbox"/> Import project metadata |
| <input type="checkbox"/> Publish to NADA | <input type="checkbox"/> Import external resources |
| <input type="checkbox"/> PDF documentation | External resources |
| <input type="checkbox"/> Change log | <input type="checkbox"/> Export RDF/XML |
| | <input type="checkbox"/> Export RDF/JSON |

- **Export package (ZIP)**

This option will allow you to generate a ZIP file containing all metadata and resources related to the project. This package can be shared with others, who can import it in their own Metadata Editor.

- **Export JSON**

Export metadata to JSON will generate a JSON file containing the metadata. The option is provided to include all elements or only the non-private ones. The JSON file will look like this:

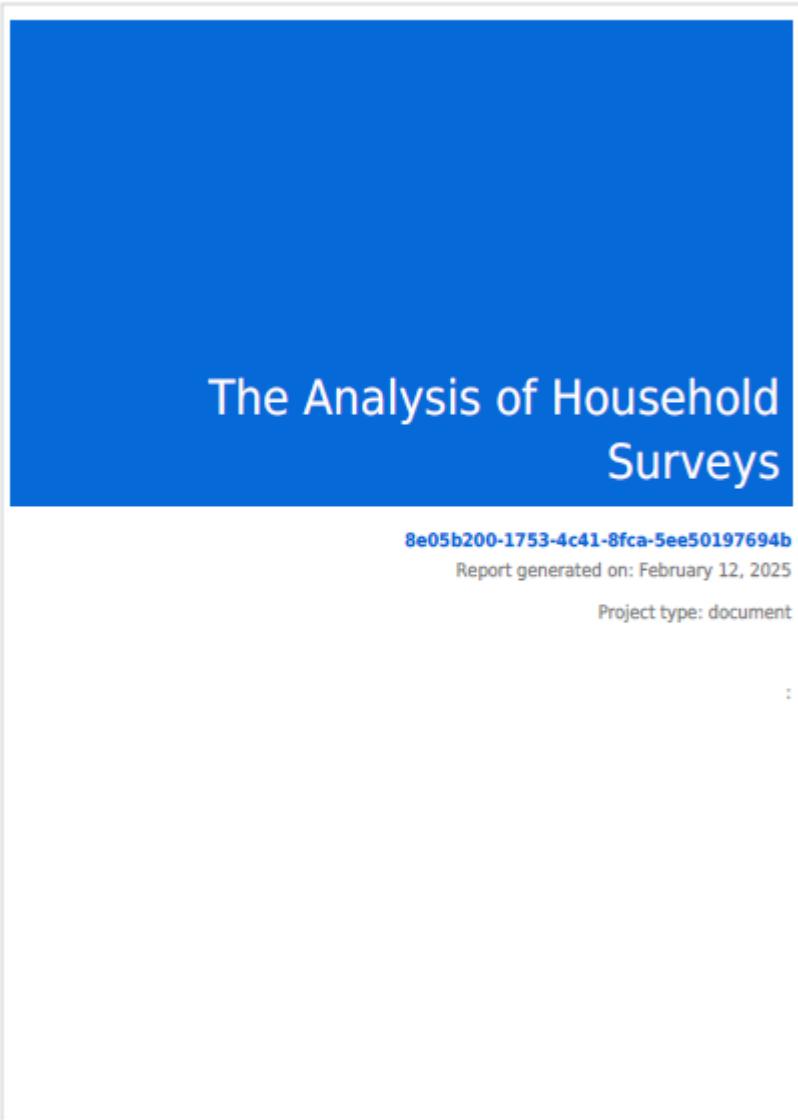
```
pretty-print 
[{"type": "document",
"idno": "8e05b200-1753-4c41-8fca-5ee50197694b",
"metadata_information": {
"producers": [
{
"name": "John Doe"
}
],
"production_date": "2025-02-12"
},
"document_description": {
"title_statement": {
"idno": "ID_DOC_001",
"title": "The Analysis of Household Surveys",
"sub_title": "A Microeconometric Approach to Development Policy"
},
"identifiers": [
{
"identifier": "10.1596/ 978-1-4648-1331-3",
"type": "DOI"
},
{
"identifier": "978-1-4648-1352-8",
"type": "ISBN (electronic)"
}
],
"authors": [
{
"first_name": "Angus",
"last_name": "Deaton"
}
],
"date_published": "2019-01-16",
"abstract": "Two decades after its original publication, The Analysis of Household Surveys is reissued with a new preface by its author, Sir Angus Deaton, recipient of the 2015 Nobel Prize in Economic Sciences. This classic work remains relevant to anyone with a serious interest in using household survey data to shed light on policy issues. This book reviews the analysis of household survey data, including the construction of household surveys, the econometric tools useful for such analysis, and a range of problems in development policy for which this survey analysis can be applied. The author's approach remains close to the data, using transparent econometric and graphical techniques to present data in a way that can clearly inform policy and academic debates. Chapter 1 describes the features of survey design that need to be understood in order to undertake appropriate analysis. Chapter 2 discusses the general econometric and statistical issues that arise when using survey data for estimation and inference. Chapter 3 covers the use of survey data to measure welfare, poverty, and distribution. Chapter 4 focuses on the use of household budget data to explore patterns of household demand. Chapter 5 discusses price reform, its effects on equity and efficiency, and how to measure them. Chapter 6 addresses the role of household consumption and saving in economic development. The book includes an appendix providing code and programs using STATA, which can serve as a template for the users' own analysis."
},
"languages": [
{
"name": "English",
"code": "EN"
}
],
"rights": "CC BY 3.0 IGO",
"type": "Book [book]",
"keywords": [
{
"name": "household surveys"
}
]
}
```

- **Export RDF/XML and Export RDF/XML**

These options allow you to export the metadata related to external resources in JSON or XML format.

- **PDF documentation**

A PDF version of the metadata can be automatically created. Select *PDF documentation* then click on **GENERATE PDF**. When the PDF is generated, click on **DOWNLOAD PDF**. You will obtain a bookmarked PDF file with all entered metadata.



- ***Publish to NADA***

If you have a NADA catalog and the credentials to publish content in it, you can also [Publish to NADA](#). Select a configured NADA catalog, select the options as shown in the screenshot below, and click [PUBLISH](#).

The screenshot shows the 'Metadata Editor' application interface. On the left is a sidebar with navigation links: Home, Metadata information, Document description, External resources, and File manager. A search bar is at the top of the sidebar. The main content area has a title 'The Analysis of Household Surveys'. At the top right are 'About', 'English', and 'John Doe' buttons, along with a 'SAVE' button.

Publish to NADA

Publish project directly to a NADA catalog

Catalog Configure new catalog

Demo catalog - <https://nada-demo.ihsn.org/index.php/catalog>
({"id": "4", "title": "Demo catalog", "url": "https://nada-demo.ihsn.org/index.php/catalog", "user_id": "11"})

Project options

| Option | Value |
|------------------------------|---------|
| Overwrite if already exists? | Yes |
| Publish | Publish |
| Data access | |
| Collection | N/A |

External resources
Select external resources to publish

Overwrite resources
1 resources found 1 selected

| Title | Type |
|---|-------------------------------|
| <input checked="" type="checkbox"/> The Analysis of Household Surveys: A Microeconometric Approach to Development Policy http://documents.worldbank.org/curated/en/593871468777303124 | Document, Technical [doc/tec] |

Options

- Publish project
- Publish thumbnail
- External resources (1)

PUBLISH

The book will now be listed and made discoverable in the NADA catalog.

Demo NADA Catalog

Data Catalog

[Home](#) [Catalog](#) [Collections](#) [Citations](#) [How to?](#) [Login](#)[Home](#) / Central Data Catalog

Keywords...

[Search](#)

All

Microdata

Geospatial

Time series

Tables

Documents

Images

Videos

Scripts

10

2

10

18

8

6

5

1

Years



Showing 1-8 of 8

Popularity



Countries

 **Impact of COVID-19 on Learning : Evidence from Six Sub-Saharan African Countries**

Burkina Faso, Ethiopia, Malawi...and 3 more, 2021

Hai-Anh H. Dang, Gbemisola Oseni Siwatu, Alberto Zezza, Kseniya Abanokova

ID: WB_159274 Last modified: Sep 11, 2021 Views: 1626

 **The Analysis of Household Surveys : A Microeconometric Approach to Development Policy**

World, 2019

Angus Deaton

ID: WB_AD_001 Last modified: Sep 11, 2021 Views: 1540

 **Disability Measurement in Household Surveys : A Guidebook for Designing Household Survey Questionnaires**

World, 2020

Marco Tiberti, Valentine Costa



Quick Start: Indicator

In this example, we will document an indicator produced by the World Bank, and published in the Bank's Poverty and Inequality Platform (PIP) and World Development Indicators (WDI): the *Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)*. The objective here is to document the indicator for cataloguing, not to publish the data.

The metadata we will enter in the Metadata Editor is the metadata published by the World Bank at <https://data.worldbank.org/indicator/SI.POV.DDAY?locations=US> (downloaded on 12 February 2025). We will complement this metadata with information extracted from the data itself, e.g., to obtain the geographic and time coverage. Last, we will document the fact that this indicator is one of the Sustainable Development Goals monitoring indicators.

The only files you need to reproduce this Quick-start example are the image file that will be used as thumbnail (.../quick_start_files/indicator/poverty_thumbnail.jpg) and the CSV file (quick_start_files/indicator/SI.POV.DDAY_countries_data) that contains the list of countries and years for which the data are available in the WDI database (CSV file extracted from an Excel file downloaded on 13 February 2025 from the World Bank website at <https://api.worldbank.org/v2/en/indicator/SI.POV.DDAY?downloadformat=excel>).

This Quick Start section does not include detailed guidance on documenting indicators and time series. For comprehensive instructions, see the chapter **Documenting indicators and databases**.

Step 1: Create a new project

To begin, open the Metadata Editor in your web browser (the URL is determined by where you installed the application), and log in with your username and password. The *My projects* page will be displayed, showing all projects you have previously created and those that have been shared with you by other users of the Metadata Editor, if any. If you are using the application for the first time and no project has been shared with you by other users of the Metadata Editor, the project list will be empty.

| Title | Owner | Last modified | Modified | Actions |
|---|----------|---------------|------------|---------|
| Popstan Synthetic Household Survey 2023 d1cff314-d3f0-4ea7-bdd6-b39966721646 | John Doe | John Doe | 2025-02-12 | ⋮ |
| The Analysis of Household Surveys 8e05b200-1753-4c41-8fc4-5e50197694b | John Doe | John Doe | 2025-02-12 | ⋮ |

Click on [CREATE NEW PROJECT](#) and select *Indicator* when prompted to indicate the type of resource you will be documenting.

Create new project

 Microdata

 Timeseries

 Timeseries database

 Document

 Table

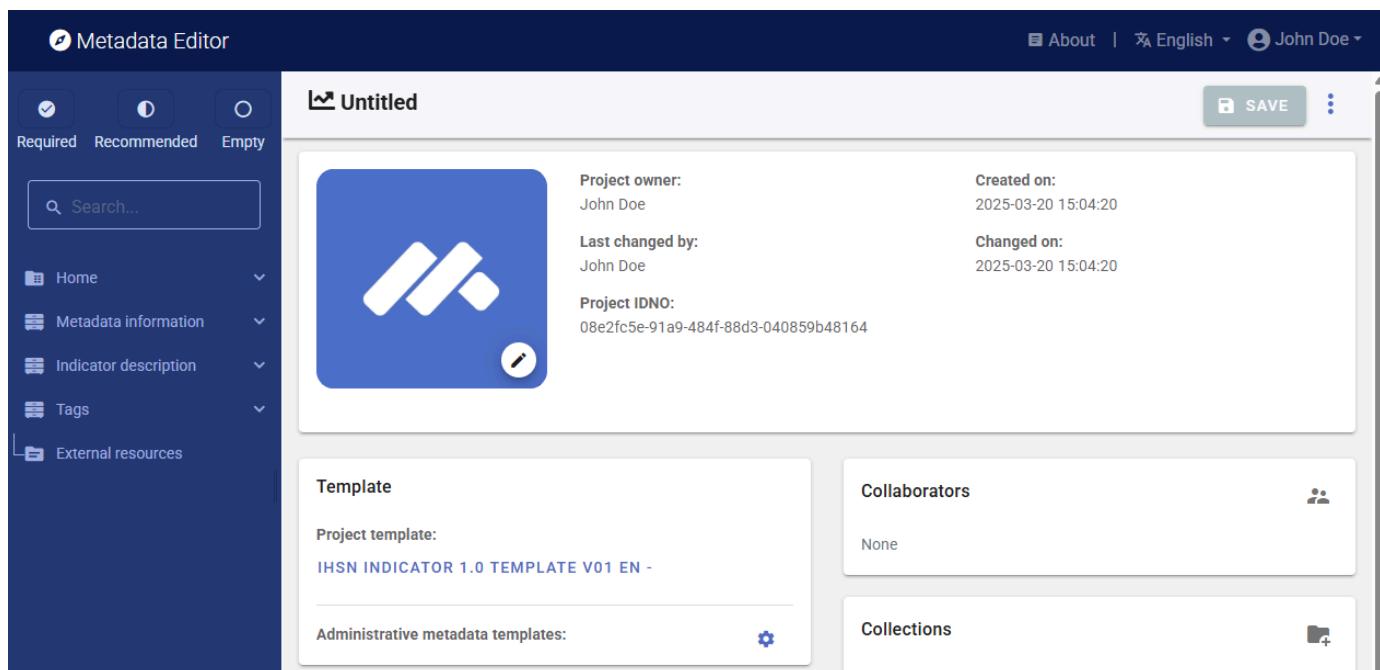
 Image

 Script

 Video

 Geospatial

A new Project page will open in a new tab.

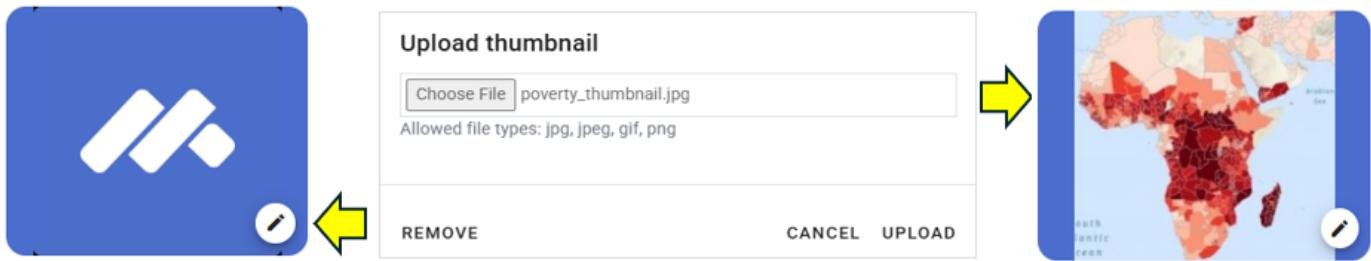


The screenshot shows the Metadata Editor interface for creating a new project. The top navigation bar includes links for About, English, and user John Doe. The main area displays a project titled "Untitled". The project details are as follows:

- Project owner: John Doe
- Created on: 2025-03-20 15:04:20
- Last changed by: John Doe
- Changed on: 2025-03-20 15:04:20
- Project IDNO: 08e2fc5e-91a9-484f-88d3-040859b48164

The left sidebar contains a search bar and navigation links for Home, Metadata information, Indicator description, Tags, and External resources. The right sidebar includes sections for Collaborators (None) and Collections.

You can use the image *poverty_thumbnail.jpg* as a thumbnail, or your own JPG or PNG file. The thumbnail will be displayed in the Metadata Editor My projects list, and in the NADA catalog if the project is published in NADA. Click on the  icon in the screenshot image, and select the image file when prompted.



Documenting an indicator consists of entering information (metadata) about the indicator in metadata entry forms defined by a *metadata template*. When you create a new project, a default template is automatically selected. We will use the template named *IHSN DDI 2.5 Template v01 EN*. If this is the template that shows in your screen, no action is needed. Otherwise, click on the template name, and select the template *IHSN DDI 2.5 Template v01 EN* in the list that will appear, then click **APPLY**.

Step 2: Enter metadata

In the navigation tree, select *Metadata information / Information on metadata* to enter optional elements used to capture information on who documented the indicator and when. Enter your name (as **Metadata producer**) and the **date** of the day in ISO format YYYY-MM-DD. This is the date when the metadata, not the indicator, was produced. Then click on **SAVE**.

You can now start entering the metadata related to the indicator itself in the "Indicator description" section. In the navigation tree, first select **Title statement** and enter the required **Primary ID**, a unique identifier of your choice, e.g., JD_IND_001 (if you want to publish the indicator in a NADA catalog, make sure that the identifier is not used by another user or for another project). Also enter the (optional) **Other identifiers** for the indicator (enter the WDI indicator identifier: SI.POV.DDAY), and the indicator **title**, which is a required element. Then click **SAVE**.

You can now proceed with the other sections in the navigation tree and fill out the relevant metadata elements using the following information extracted from the World Bank website:

- **Name:** Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)

- **Source:** World Bank, Poverty and Inequality Platform. Data are based on primary household survey data obtained from government statistical agencies and World Bank country departments. Data for high-income economies are mostly from the Luxembourg Income Study database. For more information and methodology, please see pip.worldbank.org.
- **License:** CC BY-4.0
- **Development Relevance:** The World Bank Group is committed to reducing extreme poverty to 3 percent or less, globally, by 2030. Monitoring poverty is important on the global development agenda as well as on the national development agenda of many countries. The World Bank produced its first global poverty estimates for developing countries for World Development Report 1990: Poverty (World Bank 1990) using household survey data for 22 countries (Ravallion, Datt, and van de Walle 1991). Since then there has been considerable expansion in the number of countries that field household income and expenditure surveys.
- **General Comments:** The World Bank's internationally comparable poverty monitoring database now draws on income or detailed consumption data from more than 2000 household surveys across 169 countries. See the Poverty and Inequality Platform (PIP) for details (pip.worldbank.org).
- **Limitations and Exceptions:** Despite progress in the last decade, the challenges of measuring poverty remain. The timeliness, frequency, quality, and comparability of household surveys need to increase substantially, particularly in the poorest countries. The availability and quality of poverty monitoring data remains low in small states, countries with fragile situations, and low-income countries and even some middle-income countries. The low frequency and lack of comparability of the data available in some countries create uncertainty over the magnitude of poverty reduction. Besides the frequency and timeliness of survey data, other data quality issues arise in measuring household living standards. The surveys ask detailed questions on sources of income and how it was spent, which must be carefully recorded by trained personnel. Income is generally more difficult to measure accurately, and consumption comes closer to the notion of living standards. And income can vary over time even if living standards do not. But consumption data are not always available: the latest estimates reported here use consumption data for about two-thirds of countries. However, even similar surveys may not be strictly comparable because of differences in timing or in the quality and training of enumerators. Comparisons of countries at different levels of development also pose a potential problem because of differences in the relative importance of the consumption of nonmarket goods. The local market value of all consumption in kind (including own production, particularly important in underdeveloped rural economies) should be included in total consumption expenditure but may not be. Most survey data now include valuations for consumption or income from own production, but valuation methods vary.
- **Long Definition:** Poverty headcount ratio at \$2.15 a day is the percentage of the population living on less than \$2.15 a day at 2017 purchasing power adjusted prices. As a result of revisions in PPP exchange rates, poverty rates for individual countries cannot be compared with poverty rates reported in earlier editions.
- **Periodicity:** Annual
- **Related Source Links:** World Bank, Poverty and Inequality Platform: pip.worldbank.org
- **Short Definition:** Poverty headcount ratio at \$2.15 a day is the percentage of the population living on less than \$2.15 a day at 2017 international prices.
- **Statistical Concept and Methodology:** International comparisons of poverty estimates entail both conceptual and practical problems. Countries have different definitions of poverty, and consistent comparisons across countries can be difficult. Local poverty lines tend to have higher purchasing power in rich countries, where more generous standards are used, than in poor countries. Since World Development Report 1990, the World Bank has aimed to apply a common standard in measuring extreme poverty, anchored to what poverty means in the world's poorest countries. The welfare of people living in different countries can be measured on a common scale by adjusting for differences in the purchasing power of currencies. The commonly used \$1 a day standard, measured in 1985 international prices and adjusted to local currency using purchasing power parities (PPPs), was chosen for World Development Report 1990 because it was typical of the poverty lines in low-income countries at the time. As differences in the cost of living across the world evolve, the international poverty line has to be periodically updated using new PPP price data to reflect these changes. The last change was in September 2022, when we adopted \$2.15 as the international poverty line using the 2017 PPP. Poverty measures based on international poverty lines attempt to hold the real value of the poverty line constant across countries, as is done when making comparisons over time. The \$3.65 poverty line is derived from typical national poverty lines in countries classified as Lower Middle Income. The \$6.85 poverty line is derived from typical national poverty lines in countries classified as Upper Middle Income. Early editions of World Development Indicators used PPPs from the Penn World Tables to convert values in local currency to equivalent purchasing power measured in U.S dollars. Later editions used 1993, 2005, and 2017 consumption PPP estimates produced by the World Bank. The current extreme poverty line is set at \$2.15 a day in 2017 PPP terms,

which represents the mean of the poverty lines found in 15 of the poorest countries ranked by per capita consumption. The new poverty line maintains the same standard for extreme poverty - the poverty line typical of the poorest countries in the world - but updates it using the latest information on the cost of living in developing countries. As a result of revisions in PPP exchange rates, poverty rates for individual countries cannot be compared with poverty rates reported in earlier editions. The statistics reported here are based on consumption data or, when unavailable, on income surveys.

- **Topic:** Poverty; Poverty rates
- **Unit of Measure:** %

Additional useful information

- **Database ID:** WDI
- **URL of CC BY-4.0 license:** <https://creativecommons.org/licenses/by/4.0/>
- **SDG Goal, Target Corresponding to the Indicator:**
 - **Framework:** Sustainable Development Goals (SDG)
 - **Custodian:** United Nations
 - **Goal 1:** No poverty (End poverty in all forms by 2030)
 - **Target 1.1:** By 2030, eradicate extreme poverty for all people everywhere, currently measured as people living on less than \$1.25 a day
 - **Indicator 1.1.1:** Proportion of the population living below the international poverty line by sex, age, employment status and geographical location (urban/rural)
- **Time Coverage:** 1963 to 2023 (as of February 2025)
- **Geographic Coverage:** 266 countries and regions; the list (names and codes) will be copy/pasted from the CSV file provided.

With this information in hand, you can now start documenting the indicator. The content of the WDI metadata and the additional elements can be entered in the Metadata Editor template in the following elements:

| From World Bank | In the metadata template (Indicator description) |
|----------------------------|--|
| ID | Title statement / Primary ID |
| WDI indicator ID | Title statement / Other identifiers |
| Name | Title statement / Name |
| Source | Sources, concepts and methods / Notes on data source |
| License | Access and use / License |
| Development Relevance | Sources, concepts and methods / Relevance |
| General comments | Sources, concepts and methods / Notes on data source |
| Limitations and Exceptions | Quality / Limitations |
| Long Definition | Sources, concepts and methods / Definition long |
| Periodicity | Description / Methodology |

| | |
|---------------------------|---|
| From World Bank | In the metadata template (Indicator description) |
| Related source links | Sources, concepts and methods / Data source |
| Short definition | Sources, concepts and methods / Definition short |
| Stat, Concept and Method. | Sources, concepts and methods / Methodology |
| Topic | Description / Topics |
| Unit of measure | Description / Measurement unit |
| Database ID | Indicator description / Title statement / Database ID |
| URL of CC BY-4.0 license | Access and use / License (URL) |
| SDG framework | Standards and frameworks / Frameworks / Name + Abbreviation |
| SDG custodian | Standards and frameworks / Frameworks / Custodian |
| SDG goal | Standards and frameworks / Frameworks / Goal ID + Name + Description |
| SDG target | Standards and frameworks / Frameworks / Target ID + Name + Description |
| SDG indicator | Standards and frameworks / Frameworks / Indicator ID + Name + Description |
| Time coverage | Geographic and time coverage / Time coverage |
| Geographic coverage | Geographic and time coverage / Countries |

Note: The geographic coverage for the indicator consists of a large number of countries (266 countries and economies), which must be itemized in the metadata. Entering the country names and codes manually would be tedious. Instead, you may use the copy/paste functionality of the Metadata Editor. The list of countries and country codes is available in the CSV file *SI.POV.DDAY_countries_data.csv*. Copy this list (only including the country names and codes). Select the [Geographic and time coverage / Countries' element in the navigation tree, then click on the copy/paste icon \(triple dots\)](#). Select Paste (Replace) or Paste (Append). The 266 countries and economies will now be in the metadata.

| Countries ? | Name | Code |
|---|------|------|
|  | | |

A screenshot of the Metadata Editor interface. A context menu is open over a table row containing the data for 'Aruba'. The menu options are: Copy, Paste (Replace), Paste (Append), and Undo paste.

| Name | Code |
|-----------------------------|------|
| Aruba | ABW |
| Africa Eastern and Southern | AFE |
| Afghanistan | AFG |
| Africa Western and Central | AFW |

Step 3: Add information on related resources

Once you have entered the metadata, you can finalize the documentation of the indicator by documenting and attaching external resources. External resources include all materials you want to make accessible to users when you publish the indicator in a catalog. In this example, we will add one external resource: a link to the World Bank Poverty and Inequality Platform website.

To create an external resource, click on "External resources" in the navigation tree and then click on [Create resource](#). Select the resource type from the drop down (in this case, *Web Site*), give it a short title (*Poverty and Inequality Platform (PIP)*), and enter the URL (<https://pip.worldbank.org/home>). Then click [SAVE](#). You will now have an external resources listed.

A screenshot of the Metadata Editor showing the creation of an external resource. The left sidebar shows the navigation tree with 'External resources' selected. The main area shows an 'Edit resource' form for a resource titled 'Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)'. The form fields are: Resource type: Web Site; Resource format: (empty); Title: Poverty and Inequality Platform (PIP); Author: World Bank. The 'SAVE' button is at the top right of the form.

Step 4: Export and publish metadata

In the *Project* page, a menu of options is available to you.

The screenshot shows the Metadata Editor interface with the following details:

- Top Bar:** Shows 'Metadata Editor' in the title bar, 'About', 'English', and a user profile for 'John Doe'.
- Indicator Card:** Displays the indicator 'Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)' with a map visualization and the text 'IHSN INDICATOR 1.0 TEMPLATE V01 EN -'.
- More Options:** A red circle highlights the three-dot 'More options' button in the top right corner of the indicator card.
- Sidebar Options:**
 - Project:**
 - Export package (ZIP)
 - Export JSON
 - Export MSD (SDMX/XML 3.0)
 - Export MetadataSet (SDMX/JSON ...)
 - Publish to NADA
 - PDF documentation
 - Change log
 - Metadata:**
 - Apply default values from template
 - Import project metadata
 - Import external resources
 - Export RDF/XML
 - Export RDF/JSON

- **Export package (ZIP)**

This option will allow you to generate a ZIP file containing all metadata and resources related to the project. This package can be shared with others, who can import it in their own Metadata Editor.

- **Export JSON**

Export metadata to JSON will generate a JSON file containing the metadata. The option is provided to include all elements or only the non-private ones. The JSON file will look like this:

```
Pretty-print 
{
  "type": "timeseries",
  "idno": "4333f56f-6614-4470-b4c3-d84c699495ce",
  "changed": "1739991893",
  "changed_utc": "2025-02-19T19:04:53+00:00",
  "created": "1739395113",
  "created_utc": "2025-02-12T21:18:33+00:00",
  "created_by": "11",
  "changed_by": "11",
  "metadata_information": {
    "producers": [
      {
        "name": "John Doe"
      }
    ],
    "prod_date": "2025-02-12"
  },
  "series_description": {
    "framework": [
      {
        "name": "Sustainable Development Goals",
        "abbreviation": "SDG",
        "custodian": "United Nations",
        "goal_id": "1",
        "goal_description": "End poverty in all forms by 2030. ",
        "goal_name": "No poverty",
        "target_id": "1.1",
        "target_name": "By 2030, eradicate extreme poverty for all people everywhere, currently measured as people living on less than $1.25 a day",
        "indicator_id": "1.1.1",
        "indicator_name": "Proportion of the population living below the international poverty line by sex, age, employment status and geographical location (urban/rural)"
      }
    ],
    "alternate_identifiers": [
      {
        "identifier": "SI.POV.DDAY",
        "name": "WDI",
        "database": "World Development Indicators"
      }
    ],
    "name": "Poverty headcount ratio at $2.15 a day (2017 PPP) (% of population)",
    "definition_long": "Poverty headcount ratio at $2.15 a day is the percentage of the population living on less than $2.15 a day at 2017 purchasing power adjusted prices. As a result of revisions in PPP exchange rates, poverty rates for individual countries cannot be compared with poverty rates reported in earlier editions.",
    "definition_short": "Poverty headcount ratio at $2.15 a day is the percentage of the population living on less than $2.15 a day at 2017 international prices.",
    "relevance": "The World Bank Group is committed to reducing extreme poverty to 3 percent or less, globally, by 2030. Monitoring poverty is important on the global development agenda as well as on the national development agenda of many countries. The World Bank produced its first global poverty estimates for developing countries for World Development Report 1990: Poverty (World Bank 1990) using household survey data for 22 countries (Ravallion, Datt, and van de Walle 1991). Since then there has been considerable expansion in the number of countries that field household income and expenditure surveys."
  }
}
```

- **Export MSD (SDMX/XML 3.0)**

Export a metadata structure definition compliant with the SDMX 3.0 standard.

- **Export MetadataSet (SDMX/JSON)**

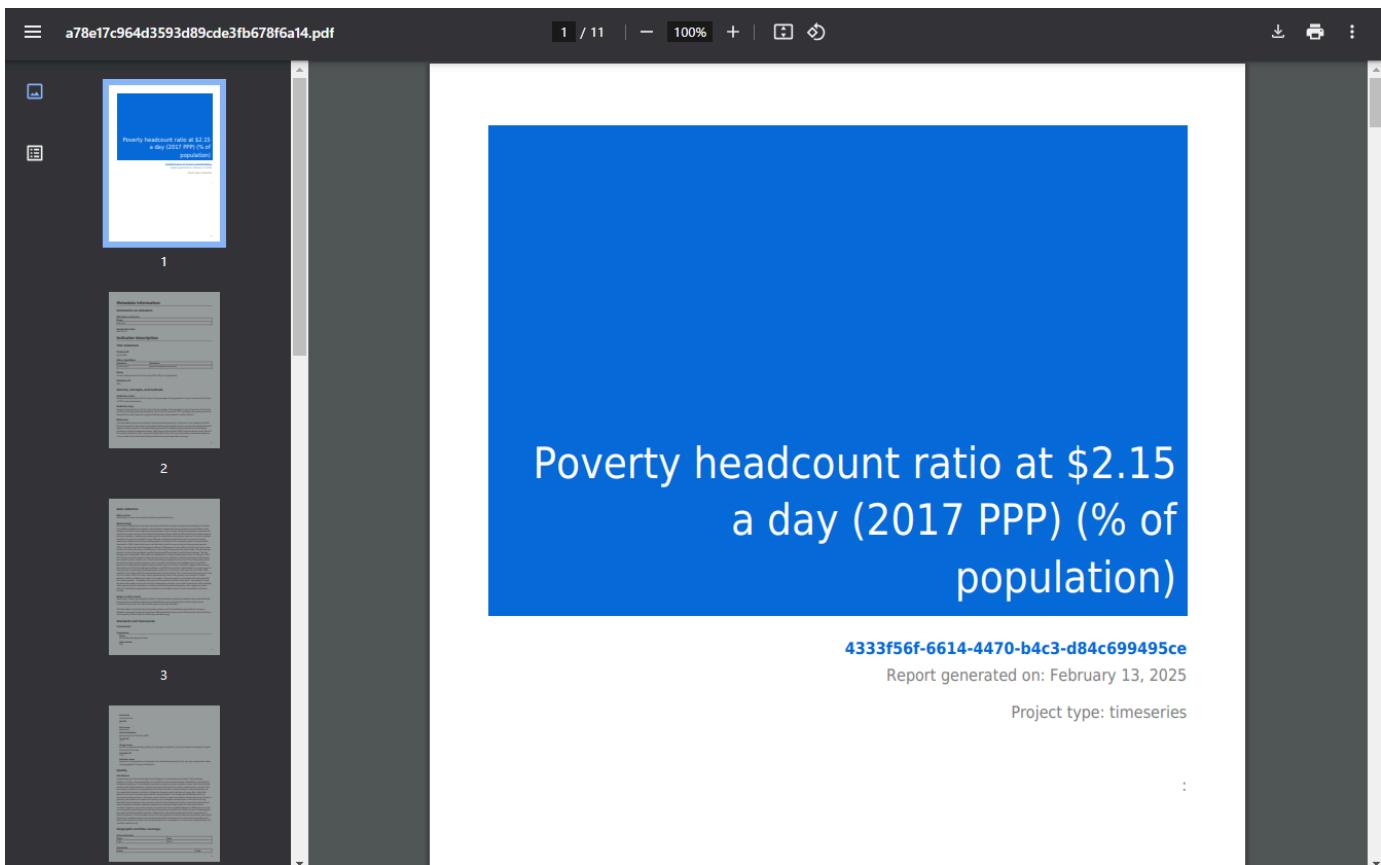
Export a metadataset compliant with the SDMX 3.0 standard.

- **Export RDF/XML and Export RDF/XML**

These options allow you to export the metadata related to external resources in JSON or XML format.

- **PDF documentation**

A PDF version of the metadata can be automatically created. Select PDF documentation then click on [GENERATE PDF](#). When the PDF is generated, click on [DOWNLOAD PDF](#). You will obtain a bookmarked PDF file with all entered metadata.



- **Publish to NADA**

If you have a NADA catalog and the credentials to publish content in it, you can also [Publish to NADA](#). Select a configured NADA catalog, select the options as shown in the screenshot below, and click [PUBLISH](#).

The screenshot shows the 'Metadata Editor' application with a dark blue theme. On the left is a sidebar with navigation links: Home, Preview, Metadata information, Indicator description, Tags, External resources (with 'Poverty and Inequality Platform (PIP)' selected), and File manager.

The main content area is titled 'Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)'. It includes a 'Catalog' section where 'NADA demo - https://nada-demo.ihsn.org' is selected. Below this are 'Project options' and 'External resources' sections. Under 'Project options', 'Overwrite if already exists?' is set to 'Yes', 'Publish' is set to 'Publish', and 'Data access' is 'Data not available - [data_na]'. Under 'External resources', 'Overwrite resources' is selected, and one resource ('Poverty and Inequality Platform (PIP)') is listed. At the bottom, under 'Options', 'Publish project' is checked. A large blue 'PUBLISH' button is at the bottom right.

The indicator will now be listed and made discoverable in the NADA catalog, with a link to the World Bank PIP platform.



Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)

1963 - 2023

Reference ID JD_IND_001

Metadata [JSON](#)

CREATED ON

Feb 13, 2025

LAST MODIFIED

Feb 13, 2025

PAGE VIEWS

1

SERIES DESCRIPTION[Overview](#)[Geographic information](#)[License](#)[Metadata production](#)

Overview

SERIES UNIQUE ID

JD_IND_001

SERIES NAME

Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)

DATABASE ID

WDI

IDENTIFIERS

| ID | Name | Database |
|-------------|------|------------------------------|
| SI.POV.DDAY | WDI | World Development Indicators |

SERIES UNIT OF MEASURE

Percentage (%)

DEFINITION SHORT

Poverty headcount ratio at \$2.15 a day is the percentage of the population living on less than \$2.15 a day at 2017

Quick start: Microdata

In this example, we will document a survey dataset (microdata) using the DDI Codebook metadata standard.

This Quick Start section does not include detailed guidance on documenting microdata. For comprehensive instructions, see the chapter **Documenting Data – Microdata**.

The dataset we will document is a synthetic household survey dataset created for a fictional country. The files needed to reproduce the example are provided in folder `.../quick_start_files/microdata`. Download and save the content of this folder in a local folder of your choice. The materials provided include:

- Two data files in Stata 17 format. All variables and values have been labeled in the Stata files. The two data files are related. Variable `hid` (an identifier unique to each household) provides a key that allows merging the data files.
 - `WLD_2023_SYNTH-SVY-HLD-EN_v01_M.dta` : a household-level data file containing 49 variables and 8,000 observations.
 - `WLD_2023_SYNTH-SVY-IND-EN_v01_M.dta` : an individual-level data file containing 27 variables and 32,396 observations.
- Survey questionnaire and survey documentation, in MS-Excel format.
 - A simplified survey questionnaire (file `synthetic_survey_questionnaire.xlsx`, with sheets *Household form EN* for variables collected at the household level, and *Individual form EN* for variables collected at the individual level.)
 - A simplified technical report, with information on the sampling design (file `synthetic_survey_info.xlsx`, sheet *Survey info*)
- Other: `synthetic_data_logo.jpg` (a logo for the survey, in JPG format)

Step 1: Create a new project

To begin, open the Metadata Editor in your web browser (the URL is determined by where you installed the application), and log in with your username and password. The My projects page will be displayed, showing all projects you have previously created and those that have been shared with you by other data curators, if any. If you are using the application for the first time and no project has been shared with you, the project list will be empty. In the example below, we see that one project was previously created.

Metadata Editor

About | English | John Doe

My projects

PROJECTS **COLLECTIONS** **TEMPLATES** **ADMINISTRATIVE METADATA**

CREATE NEW PROJECT IMPORT

Search... Recent ↑

Showing 1 - 1 of 1 projects

| | Title | Owner | Last modified | Modified | Actions |
|--------------------------|---|----------|---------------|------------|---|
| <input type="checkbox"/> |  The Analysis of Household Surveys 8e05b200-1753-4c41-8fca-5ee50197694b | John Doe | John Doe | 2025-02-12 |  |

Click on **CREATE NEW PROJECT** and select *Microdata* when prompted to indicate the type of resource you will be documenting (a survey dataset).

Create new project

 **Microdata**

 **Timeseries**

 **Timeseries database**

 **Document**

 **Table**

 **Image**

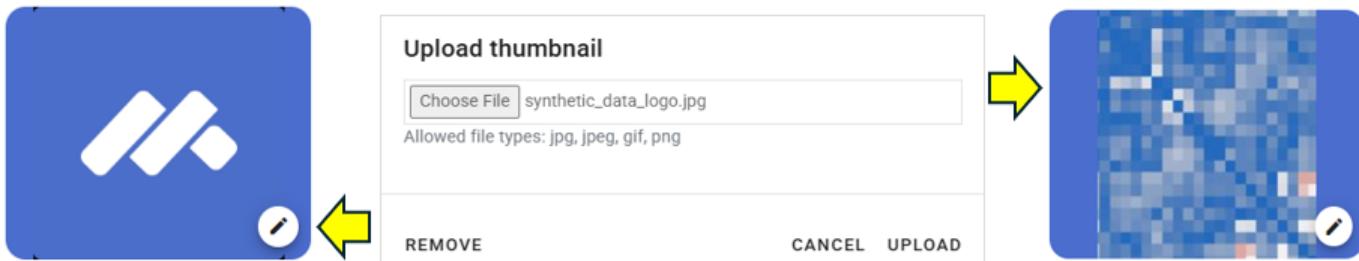
 **Script**

 **Video**

 **Geospatial**

A new project page will open in a new tab.

We will use an image as thumbnail for the project. The thumbnail is optional; it will be displayed in the Metadata Editor project list, and in the NADA catalog if the metadata is published in NADA. For a survey or census dataset, you will typically use the logo of the survey or census as thumbnail. You can change the image that will be used as a thumbnail for the project (optional). You may use the provided image (`synthetic_data_logo.jpg`) or another jpg or png file of your choice. Click on the `edit` icon and select the image file when prompted.



Documenting a dataset (or a document in this case) consists of entering metadata in metadata entry forms defined by a *metadata template*. When you create a new project, a default template is automatically selected. We will use this template, so there is no need to switch template.

##Step 2: Add descriptive metadata

In the section *Document description / Metadata preparation*, enter the (optional) information on the metadata. Enter your name (as `Metadata producer`) and the `date` of the day in ISO format YYYY-MM-DD. This is the date when the metadata, not the dataset, was produced. Then click on `SAVE`.

The screenshot shows the 'Metadata preparation' section of the Metadata Editor. The left sidebar has a tree view with 'Home', 'Document description', 'Metadata preparation' (which is expanded), 'Metadata producers', 'Production date', 'Document ID', 'DDI metadata version', 'Study description', and 'Tags'. The main area has a title 'Untitled'. It contains a 'Metadata preparation' section with a note 'Metadata producers', a table with one row for 'John Doe', a 'Production date' field set to '2025-02-12', and a 'Document ID' field. A note below the production date says: 'The date the DDI metadata document was produced (not the date it was distributed or archived), preferably entered in ISO 8601 format (YYYY-MM-DD or YYYY-MM). A validation rule can be set in user templates to enforce a date format. This is a "Recommended" element, as information on the producer and on the date of metadata production is useful for catalog administration purposes.' A 'SAVE' button is at the top right.

You can now start entering the descriptive metadata related to the survey itself in the *Study Description* section. You will find the relevant information in the Excel file *synthetic_survey_info.xlsx*. This information is also provided below:

- **Title:** Popstan Synthetic Household Survey 2023
- **Acronym:** PSHS 2023
- **Country / code:** Popstan / POP
- **Producer:** National Statistics Office (NSO)
- **Funding:** The Foundation for Synthetic Data
- **Dates of data collection:** 4 July 2023 to 27 December 2023
- **Abstract:** The 2023 Popstan Synthetic Household Survey is a periodic national welfare monitoring survey conducted by the Popstan National Statistics Office. It is used to update the national poverty profile, based on the (indexed) national poverty line calculated in 2017.
- **Series:** The 2023 survey is the 4th welfare monitoring survey conducted by the National Statistics Office. The survey is conducted every two years. Previous surveys were conducted in 2017, 2019, and 2021.
- **Universe:** Resident population with exception of homeless, nomads, and residents in institutional households
- **Geographic coverage:** National (all 10 regions)
- **Topics:** Demographics; education; labor; assets ownership; household expenditure, housing, water and sanitation.
- **Sampling:** A sample of 8,000 households was selected. The response rate was 100%.
- **Sample frame:** Synthetic population census dataset (conducted in January 2023, with January 15 as census night)
- **Sample design:** A stratified sample was drawn. The urban/rural areas of each province (geo1) were used as strata. The sample of 8,000 households was selected proportional to the size of each geo1. In each stratum, we randomly enumerate areas (Eas), and in each EA we randomly selected 25 households (uniform sample).
- **Weighting:** The sample weight was calculated for each household (variable hhweight). Considering the sample design and the response rate of 100%, no adjustment or calibration of the weights was required.
- **Mode of data collection** Multi-mode: CAPI (face-to-face, using tablets for data capture) and CATI (telephone interviews)
- **Interviewer's training:** Interviewers were trained by the staff of the National Statistics Office during a period of two weeks prior to the fieldwork. All interviewers were teachers. The training was provided at the provincial level, in the two national languages.
- **Fieldwork organization:** One interviewer was contracted per enumeration area. The interviewers conducted 4 visits to each household during the survey period. One controller was assigned to each team of 5 interviewers. One supervisor was assigned to each province.
- **Data processing:** Data were captured during data collection (face-to-face or by phone) using tablets and the Survey Solutions software application. Validation (range and consistency checks) was implemented in real time by the

application. At the completion of each one of the 4 phases of the survey, the data were processed for batch editing. ANalysis and tabulation was then made using R.

- **Standard compliance:** The questionnaire was designed in compliance with national classifications and using the survey question bank maintained by the National Statistics Office.
- **Version of dataset:** v01_2023_EN
- **Citation requirement:** Users of the dataset are requested to cite the data as follows: National Statistics Office (NSO) of Popstan. 2023. Microdata from the Synthetic Household Survey 2023 (version v01_2023_EN). Downloaded from NSO website on [date]

In the navigation tree, first select *Study description / Identification*, and enter the relevant information on the survey. First, enter the **title** of the dataset and the required **Primary ID**, a unique identifier of your choice, e.g., JD_MICRO_001 (if you want to publish the document in a NADA catalog, make sure that the identifier is not used by another user or for another project).

The screenshot shows the Metadata Editor interface. The left sidebar has a dark blue header with 'Metadata Editor' and three status indicators: 'Required' (green checkmark), 'Recommended' (yellow circle), and 'Empty' (red circle). Below this is a search bar with a magnifying glass icon. The main navigation tree on the left is expanded to 'Study description' > 'Identification', which includes 'Title', 'Subtitle', 'Alternate title', 'Translated title', 'Primary ID', 'Other identifiers', 'Study type', and 'Series information'. The right panel shows the 'Popstan Synthetic Household Survey 2023' record. Under 'Identification', the 'Title' field contains 'Popstan Synthetic Household Survey 2023', 'Primary ID' contains 'JD_MICRO_001', and 'Study type' contains 'Socio-Economic/Monitoring Survey'. There is also a table for 'Other identifiers' with a '+ ADD ROW' button. The top right of the interface has 'About', 'English', and 'John Doe' dropdowns, along with a 'SAVE' button and other icons.

Then browse the navigation tree and find the most appropriate elements for each piece of information contained in the Excel file; populate other sub-sections of the *Study description* as relevant (see table below). Not all fields are expected to be filled, but all available information must be captured in the Metadata Editor. The (?) icon next to each element label will display a description of the element. Click **SAVE** periodically.

The metadata elements in the template correspond to the content of the Excel file *synthetic_survey_info.xlsx* as follows:

| In synthetic_survey_info.xlsx | In the metadata template (Study description) |
|-------------------------------|---|
| Title | Identification / Title |
| Acronym | Identification / Alternate title |
| Country / code | Universe and geographic coverage / Country |
| Producer | Producers and sponsors / Primary producer |
| Funding | Producers and sponsors / Funding agencies |

| In synthetic_survey_info.xlsx | In the metadata template (Study description) |
|-------------------------------|--|
| Dates of data collection | Data collection / Dates of data collection |
| Abstract | Overview / Abstract |
| Series | Identification / Series information |
| Universe | Universe and geographic coverage / Universe |
| Geographic coverage | Universe and geographic coverage / Geo. coverage |
| Topics | Scope / Topics (itemize the list) |
| Sampling | Sampling / Response rates |
| Sample frame | Sampling / Sample frame / Name |
| Sample design | Sampling / Sampling procedure |
| Weighting | Sampling / Weighting |
| Mode of data collection | Data collection / Mode of data collection |
| Interviewer's training | Data collection / Collector training |
| Fieldwork organization | Data collection / Control operations |
| Data processing | Data processing / Data processing |
| Standard compliance | Quality standards / Standard compliance |
| Version of dataset | Version / Version name |
| Citation requirement | Data access / Citation requirement |

Step 3: Import and document the data files (section Study Description)

When all available descriptive information is entered in *Study description*, click on *Data files* in the navigation bar. On the *Data files* page, click [IMPORT FILES](#), select the two Stata data files to be imported (*WLD_2023_SYNTH-SVY-HLD-EN_v01_M.dta* and *WLD_2023_SYNTH-SVY-IND-EN_v01_M.dta*), and click [IMPORT](#).

The Editor will import the data files, and extract all available metadata from the files (variable list, names, variable labels, value labels). It will also generate summary statistics.

The *Data files* page will now display your two files, which are also listed in the navigation bar under *Data files*.

| File# | File name | Variables | Cases | Modified | Data |
|-------|--|-----------|-------|------------|---|
| F1 | WLD_2023_SYNTH-SVY-HLD-EN_v01_M WLD_2023_SYNTH-SVY-HLD-EN_v01_M.csv 1.57 MB | 49 | 8000 | 2025-02-12 | CLEAR DATA EDIT DELETE EXPORT |
| F2 | WLD_2023_SYNTH-SVY-IND-EN_v01_M WLD_2023_SYNTH-SVY-IND-EN_v01_M.csv 2.42 MB | 27 | 32396 | 2025-02-12 | CLEAR DATA EDIT DELETE EXPORT |

You can preview the data by clicking *Data*, but note that the data cannot be edited in the Metadata Editor.

You can also view summary statistics for all variables by clicking on **Variables** for a selected file in the navigation tree, and selecting tab **STATISTICS**.

You will now document the data files they contain. First, click on the filename of a data file in the navigation tree, add a brief description of the file, and click **SAVE** (you must save before moving to another page). Do that for both data files.

Metadata Editor

Popstan Synthetic Household Survey 2023

Data file: WLD_2023_SYNTH-SVY-HLD-EN_v01_M

File name: WLD_2023_SYNTH-SVY-HLD-EN_v01_M

Description: Contains all variables at the household level

Producer:

Data checks:

SAVE CANCEL

Step 4: Document the variables

Next, we will add or edit the variable-level information. In the navigation bar, select *Variables* for one of the data files (you will do that for both data files).

| Variables | Search |
|-------------|-------------------------------------|
| V1 hid | Unique household identifier |
| V2 geo1 | Geographic area - Admin 1 |
| V3 geo2 | Geographic area - Admin 2 |
| V4 ea | Enumeration area |
| V5 urbrur | Residence (urban/rural) |
| V6 hhsize | Household size |
| V7 statocc | Status of occupancy of the dwelling |
| V8 rooms | Number of rooms in dwelling |
| V9 bedrooms | Number of bedrooms in dwelling |
| V10 floor | Main material of the floor |
| V11 walls | Main materials of the walls |
| V12 roof | Main materials of the roof |

Variable categories:

| Value | Label |
|-------|-------|
| 1 | Rural |
| 2 | Urban |

+ ADD ROW

Variable information:

- Is weight:
- Interval type: Discrete
- Type:

The page displays the list of variables in the selected file, along with multiple options to edit and complete the metadata related to each variable. On this page, you can:

- Edit the variables labels.
- Edit the value labels (for discrete/categorical variables only).
- If necessary, delete variables.
- Tag one or multiple variable(s) as being sample weight(s).

- Add metadata related to the variable (literal question, interviewer instructions, derivation and imputation, and more) in the **DOCUMENTATION** tab.
- Identify values to be treated as *missing value*. The system missing values in Stata or SPSS will be automatically identified as *missing*. However, in some cases, one (or multiple) values may be used to represent missing values (e.g., "99" may have been used as a code to represent missing or unknown for a variable such as age).
- Set the weighting coefficient (if relevant) to be applied to generate weighted summary statistics.
- Select the summary statistics to be included in the metadata (in tab **STATISTICS**).

Note that you cannot rename the variables in the Metadata Editor, as this is considered as a change of data. If you need to change your data (renaming variables, creating new ones, deleting observations, or editing the data themselves), you will have to do that outside the Metadata Editor and re-import the modified data files.

Refer to the section on Microdata Documentation for a detailed description of all available options.

We will assume for this example that all variable and value labels as extracted from the data files do not need to be edited. You can browse the list of variables to check that their *type* has been properly detected when the data were imported. For instance, the variable *hhsize* (household size) in file *WLD_2023_SYNTH-SVY-HLD-EN_v01_M* has been imported as a discrete variable. This may be fine, but you may prefer to declare it as a continuous variable, which would allow you to generate weighted means in the Metadata Editor. To do this, change the *Interval type* from "Discrete" to "Continuous."

You will now add metadata for each variable. Most of the relevant metadata at the variable level (other than the metadata extracted automatically from the imported data files) will typically be found in the survey questionnaire and in the interviewer manual. For derived variables, the relevant information may be found in other technical documents. The **DOCUMENTATION** tab displays the metadata for the variable(s) selected in the list of variables. The following variable-level information will be added:

- **Universe** of the variable
- **Pre-question**, **Literal question**, and **Post-question** as formulated in the questionnaire (for collected variables, not for derived variables)
- **Derivation** or **Imputation** method (for derived variables)

The screenshot shows the Metadata Editor interface for the Popstan Synthetic Household Survey 2023. The left sidebar lists data files: WLD_2023_SYNTH-SVY-HLD-EN_v01_N and WLD_2023_SYNTH-SVY-IND-EN_v01_M. The main area displays four variables: walls, roof, water, and piped_water, each with its main material listed. The 'DOCUMENTATION' tab is selected for the 'roof' variable. The documentation panel includes sections for 'Universe', 'Questions and instructions', 'Pre-question text', 'Literal question', and 'Post question text'. A right-hand sidebar shows a table of values and labels, and settings for 'Interval type' (Discrete), 'Type' (Numeric), and 'Min', 'Max', 'Decimal points' (set to 1.0, 9.0, and 1 respectively). A note at the bottom of the documentation panel states: 'Record observation. If observation is not possible, ask the respondent to determine the material of the roof.'

Useful tip: When the information to be entered in the Metadata Editor is the same for multiple variables, you can enter it for all relevant variables in a same file at once. To do that, select multiple variables (using the Shift or Ctrl key) and enter the information in the relevant element(s). For example, the three variables related to education in the *individual* dataset

have the same universe ("Population aged 6 and above"). The three variables can be selected, and the information entered in *Universe* will then be automatically applied to the three variables.

After entering all available variable-level metadata, you may want to set and apply a sample weight to generate weighted summary statistics. In the dataset, two variables are used as weighting coefficients: *hhweight* (found in both data files) and *popweight* (found only in the household-level file). In both files, select the variables used as sample weights and click on **Is weight** in the *Variable information* frame. An icon 'w' will be added in the variable list to indicate this special status for the selected variables.

The screenshot shows the 'Variable Information' frame on the right side of the interface. A red circle highlights the 'Is weight' checkbox, which is checked for the 'popweight' variable. The 'popweight' variable is also highlighted with a red box in the main variable list. The variable list includes V47 (quint_rur), V48 (hhweight), and V49 (popweight).

Once the weighting variables have been identified and set as weights, you may use them to generate weighted summary statistics. Proceed as follows:

- In the household-level file:
 - Select variables 2 to 44 (using the Shift key to select a block; do not select the variable *hid* as it would not make sense to generate weighted summary statistics for the unique household identifier). Then, in the **WEIGHTS** tab, select variable *hhweight* as weighting coefficient.
 - Select variables 45 to 47 (quintiles), and select *popweight* as weighting coefficient.
- In the individual-level file:
 - Select variables 3 to 26, and in tab **WEIGHTS** select *hhweight* as the weighting coefficient.

You will see that red icons have been added in the list of variables, indicating that the summary statistics for these variables need to be refreshed.

| | | | |
|-----|----------------|-----------------------------|--|
| V15 | occupation | Main occupation | |
| V16 | industry | Industry of main occupation | |
| V17 | migrate_recent | Recent migration | |
| V18 | disability | Has a disability | |

For each file, click on the **Refresh stats** button. The summary statistics will be recalculated to include weighted estimates.

The screenshot shows the 'Variables' list on the left. A red circle highlights the 'Refresh stats' button in the top right corner of the list area. The list includes variables V1 (hid), V2 (idno), and V3 (relation).

The summary statistics will now display both the unweighted and weighted values.

The screenshot shows the Metadata Editor interface for the Popstan Synthetic Household Survey 2023. The left sidebar includes links for Home, Preview, Document description, Study description, Tags, and Data files. Under Data files, there are two entries: WLD_2023_SYNTH-SVY-HLD-EN and WLD_2023_SYNTH-SVY-IND-EN, each with Variables and Data sub-options. The main workspace is titled 'Popstan Synthetic Household Survey 2023'. It displays a table of variables (V9 to V13) with their descriptions. A yellow arrow points from the 'Variables' section to the 'STATISTICS' tab. Another yellow arrow points from the 'STATISTICS' tab to the 'Frequencies' table. A third yellow arrow points from the 'Frequencies' table to the 'Summary statistics' table. The right side features a 'Variable categories' section with a table of values and labels, and a 'Variable information' panel with a 'Is weight' toggle.

The summary statistics that are displayed in the Metadata Editor will be part of the saved metadata. As a last step in documenting variables, you should browse the variables to verify that the selected summary statistics for each variable are indeed meaningful. Statistics like *mean* or *standard deviation* should not be included for categorical variables.

Step 5: Add information on related resources

Once you have entered the variable-level metadata for both files, you can finalize the documentation of the dataset by documenting and attaching *external resources* to the survey metadata. External resources include all materials that you want to make accessible to users when you publish the dataset in a catalog. This may include the microdata files if you want to disseminate them (openly or under restrictions). In this example, we will load the dataset and the two Excel files that contain respectively the questionnaire and the information on the survey. The two Stata files can be uploaded as one single ZIP file. Note that you could provide the data in more than one format, for example, you could export them to CSV and SPSS formats and include data files in these formats as external resources, for user convenience.

To create external resources, click on *External resources* in the navigation tree and then click on [Create resource](#). For each resource, select the file type ("Document, Questionnaire" for the questionnaire, "Document, Technical" for the survey info, and "Microdata" for the data files), give each resource a short title (e.g., "Dataset in Stata 12 format" for the zipped data file), and upload the corresponding file in the *Resource attachment* part of the metadata entry page. Then click [SAVE](#). You will now have three external resources listed.

Step 6: Export and publish metadata

In the *Project* page, a menu of options will be available to you.

- **Export package (ZIP)**

This option will allow you to generate a ZIP file containing all metadata and resources related to the project, including the data files if you have not removed them. This package can be shared with others, who can import it in their own Metadata Editor.

- **Export DDI Codebook**

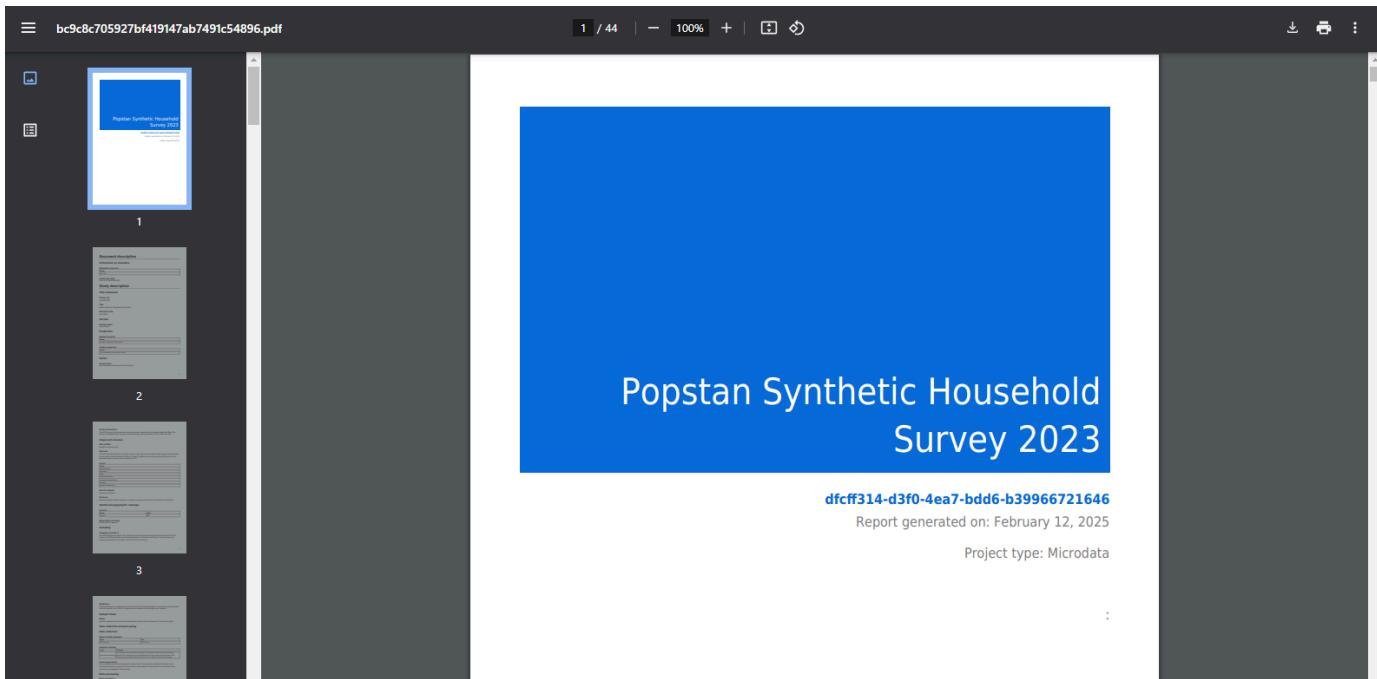
Export metadata compliant with the DDI Codebook as an XML file.

- **Export RDF/XML and Export RDF/XML**

These options allow you to export the metadata related to external resources in JSON or XML format.

- **PDF documentation**

A PDF version of the metadata can be automatically created. Select PDF documentation then click on **GENERATE PDF**. When the PDF is generated, click on **DOWNLOAD PDF**. You will obtain a bookmarked PDF file with all entered metadata.

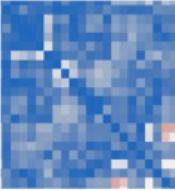


- **Publish to NADA**

If you have a NADA catalog and the credentials to publish content in it, you can also "Publish to NADA". Select a configured NADA catalog, select the options as shown in the screenshot below, and click **PUBLISH**.

The screenshot shows the Metadata Editor interface with the project "Popstan Synthetic Household Survey 2023" selected. On the left, there is a sidebar with navigation links like Home, Preview, Document description, Study description, Tags, Data files, Variable groups, External resources (Survey questionnaire, Survey information, Full dataset in Stata 17 format), and File manager. The main area shows the "Catalog" configuration, where a demo catalog URL is entered. Below that, "Project options" are set, including "Overwrite if already exists?" (Yes) and "Publish" (selected). Under "External resources", there is a table showing selected resources: "Title" and "Survey questionnaire", both categorized as "Document, Questionnaire [doc/qst]". A "Overwrite resources" toggle is also present.

The dataset will now be listed and made discoverable in the NADA catalog. The microdata will be available to users, under the access policy you selected.



Popstan Synthetic Household Survey 2023

Popstan, 2023 [GET MICRODATA](#)

| | |
|--------------|--|
| Reference ID | JD_MICRO_001 |
| Producer(s) | National Statistics Office (NSO) |
| Metadata | DDI/XML JSON |

CREATED ON
Feb 12, 2025
LAST MODIFIED
Feb 12, 2025
PAGE VIEWS
701
DOWNLOADS
20

[STUDY DESCRIPTION](#) [DATA DICTIONARY](#) [DOWNLOADS](#) [GET MICRODATA](#)

| Identification | | | | | |
|------------------------|--|------|--------------|---------|-----|
| Version | SURVEY ID NUMBER | | | | |
| Coverage | JD_MICRO_001 | | | | |
| Producers and sponsors | TITLE | | | | |
| Sampling | Popstan Synthetic Household Survey 2023 | | | | |
| Data collection | ABBREVIATION OR ACRONYM | | | | |
| Data processing | PSHS 2023 | | | | |
| Quality standards | COUNTRY | | | | |
| Data Access | <table border="1"> <thead> <tr> <th>Name</th> <th>Country code</th> </tr> </thead> <tbody> <tr> <td>Popstan</td> <td>POP</td> </tr> </tbody> </table> | Name | Country code | Popstan | POP |
| Name | Country code | | | | |
| Popstan | POP | | | | |
| Metadata production | STUDY TYPE Socio-Economic/Monitoring Survey [hh/sems] | | | | |
| | SERIES INFORMATION The 2023 survey is the 4th welfare monitoring survey conducted by the National Statistics Office. The survey is conducted every two years. Previous surveys were conducted in 2017, 2019, and 2021. | | | | |

Quick start: Geographic dataset

In this example, we will document a geographic dataset extracted from the Humanitarian Data Exchange (HDX) website. This dataset provides an outline of camps of Rohingya refugees in Cox's Bazar, Bangladesh. The only file you need to reproduce this Quick-Start example is the image file .../image/HDX_BGD_camps_thumbnail.jpg (feel free to use any another PNG or JPG image file of your choice).

This Quick Start section does not include detailed guidance on documenting geographic data. For comprehensive instructions, see the chapter **Documenting geographic datasets and services**.

Step 1: Create a new project and add a thumbnail

To begin, open the Metadata Editor in your web browser (the URL is determined by where you installed the application), and log in with your username and password. The *My projects* page will be displayed, showing all projects you have previously created and those that have been shared with you by other data curators, if any. If you are using the application for the first time and no project has been shared with you, the project list will be empty.

| Showing 1 - 4 of 4 projects | | | | | |
|-----------------------------|--|-----------|---------------|------------|---|
| | Title | Owner | Last modified | Modified | Actions |
| <input type="checkbox"/> |  Market near Ramallah's main mosque 2c63f358-e5ef-4bb3-a654-a8968b6ba694 | John Doe | John Doe | 2025-02-15 |  |
| <input type="checkbox"/> |  Poverty headcount ratio at \$2.15 a day (2017) | .John Doe | .John Doe | 2025-02-13 |  |

Click on **CREATE NEW PROJECT** and select **Geospatial** when prompted to indicate the type of resource you will be documenting (in this case a geographic dataset).

Create new project

 Microdata

 Timeseries

 Timeseries database

 Document

 Table

 Image

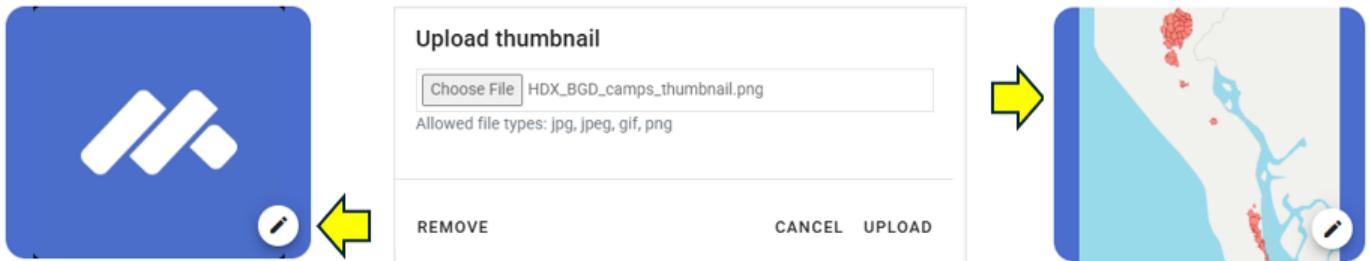
 Script

 Video

 Geospatial

A new project page will open in a new tab.

You will use the JPG file *HDX_BGD_camps_thumbnail.jpg* as a thumbnail (or you may select another JPG or PNG image file of your choice). Note that providing a thumbnail is not required, but recommended. The thumbnail will be displayed in the Metadata Editor project list, and in the NADA catalog if the project is published in NADA. Click on the  icon in the screenshot image, and select the image file when prompted.



Documenting a dataset consists of entering metadata in metadata entry forms defined by a metadata template. When you create a new project, a default template is automatically selected. We will use this template, so there is no need to switch template. The template we will use is named *Geospatial - Inspire/Gemini with additional elements* (version 1.0). It contains the metadata elements recommended by the INSPIRE directive of the European Union and the GEMINI specification from the United Kingdom, to which a few elements have been added.

Step 2: Enter metadata

On the left navigation tree, select *Document description* to enter optional elements used to capture information on who documented the dataset and when. Enter your  as metadata producer, and the  of the day in ISO format (YYYY-MM-DD). This is the date when the metadata, not the dataset, was produced. Then click on .

The screenshot shows the 'Metadata Editor' interface with the title 'Outline of camps of Rohingya refugees in Cox's Bazar, Bangladesh'. On the left, a navigation tree under 'Document description' includes 'Document title', 'Document ID', 'Metadata producers', 'Production date', and 'Version'. The 'Metadata producers' section contains a table with one row for 'John Doe'. The right side shows fields for 'Document title', 'Document ID', 'Metadata producers', 'Production date' (set to '2015-02-14'), and 'Version'. A 'SAVE' button is at the top right.

You can now start entering the metadata related to the geographic dataset itself. In the navigation tree, first select **Description / Metadata** and enter the required information on the type of resource ("Dataset") in the field **Hierarchy level** and the required **Primary ID** (a unique identifier of your choice, e.g., JD_GEO_001; if you want to publish the document in a NADA catalog, make sure that this same identifier is not used by another user or for another project).

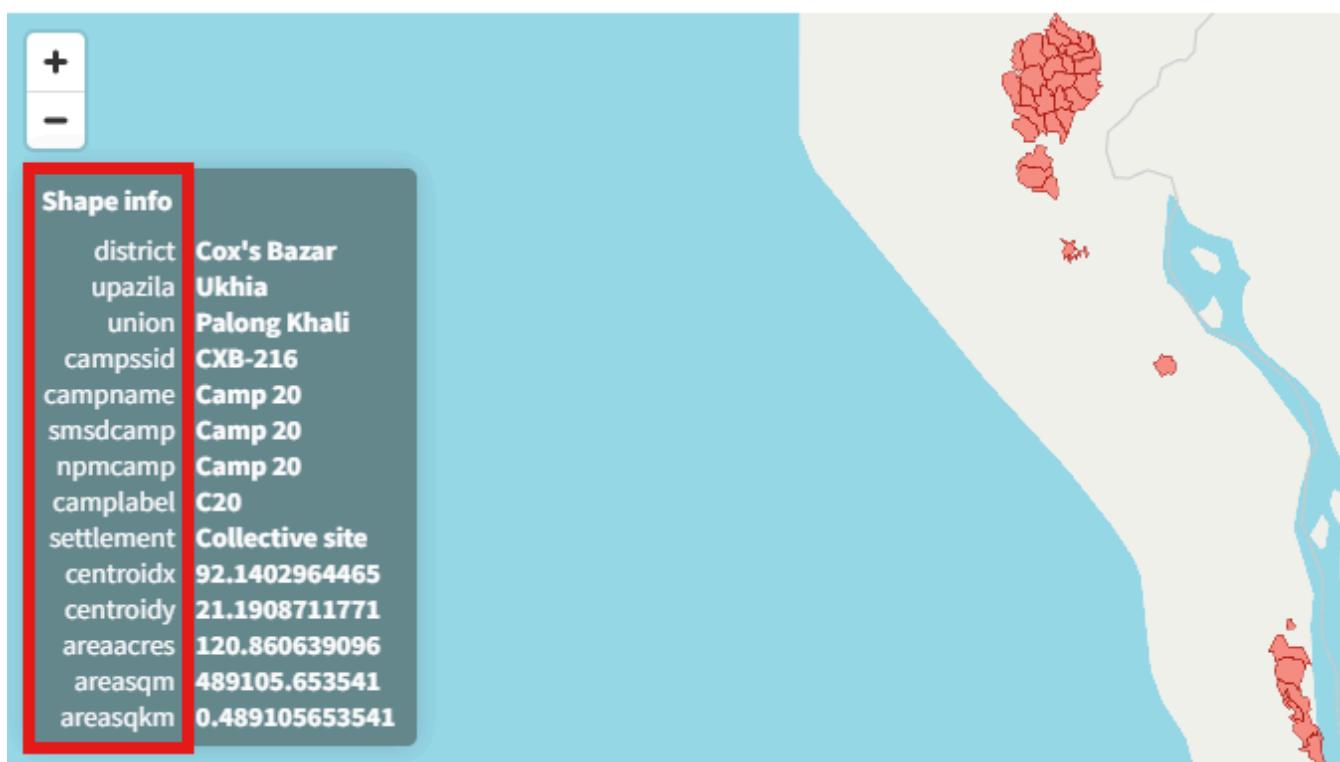
The screenshot shows the 'Metadata Editor' interface with the title 'Outline of camps of Rohingya refugees in Cox's Bazar, Bangladesh'. On the left, a navigation tree under 'Description' includes 'Metadata' (selected), 'Hierarchy level', 'Primary ID', 'Parent identifier (for series)', 'Metadata date', 'Metadata language', 'Metadata contact', 'Metadata standard name', 'Metadata standard version', 'Metadata character set', and 'Dataset URI'. The 'Metadata' section contains fields for 'Hierarchy level' (set to 'Dataset'), 'Primary ID' (set to 'JD_GEO_001'), 'Parent identifier (for series)', 'Metadata date' (set to '2025-05-08'), 'Metadata language' (set to 'English [eng]'), 'Metadata contact' (listing '1 - Metadata contact John Doe World Bank'), and a 'Metadata standard name' field. A 'Clear' button is next to the contact list. A 'Add section - Metadata contact' button is at the bottom right.

Then proceed with the other sections in the navigation tree and fill out the following elements using the information provided in the HDX website, also provided below (see *Additional information* section in web page <https://data.humdata.org/dataset/outline-of-camps-sites-of-rohingya-refugees-in-cox-s-bazar-bangladesh>). The metadata template is a bit complex, due to the complexity of the underlying ISO 19139 metadata standard.

- **Type of dataset (hierarchy level)** : Dataset

- **Metadata date:** (enter the date you are documenting the dataset)
- **Metadata language:** English
- **Metadata contact:** Enter your name and the date you are documenting the dataset.
- **Metadata standard used:** ISO 19139
- **Dataset website:** <https://data.humdata.org/dataset/outline-of-camps-sites-of-rohingya-refugees-in-cox-s-bazar-bangladesh>
- **Metadata update:** Not planned
- **Title:** Outline of camps of Rohingya refugees in Cox's Bazar, Bangladesh
- **Source:** RRRC, Inter Sector Coordination Group (ISCG), Site Management Sector, UNHCR, IOM
- **Description:** This spatial database contains the outline of the camps, settlements, and sites where Rohingya refugees are staying in Cox's Bazar, Bangladesh.
- **Time period of the dataset:** January 20, 2018 - April 05, 2024
- **Modified:** 19 May 2024
- **Expected update frequency:** As needed
- **Location:** Bangladesh
- **Topics:** Society
- **Keywords:** settlements; refugee crisis; refugees ; forced displacement ; refugee camps
- **Dataset language:** English
- **Presentation form:** Digital map
- **Status:** Completed
- **Bounding box (W/E/S/N):** 88.0844222351 ; 92.6727209818 ; 20.670883287 ; 26.4465255803
- **Geographic description:** Bangladesh
- **Temporal element:** From 2018-01-20 to 2024-04-05
- **Contributor:** Inter Sector Coordination Group (ISCG)
- **Spatial representation type:** Vector
- **Frequency of updates:** As needed
- **Methodology (lineage):** These polygons were digitized through a combination of methodologies, originally using VHR satellite imagery and GPS points collected in the field, verified and amended according to Shelter-CCCM Sector, RRRC, Camp in Charge (CiC) officers inputs, with technical support from other partners.
- **Caveats/Comments:** The camps are continuously expanding, and Camp Boundaries are structured around the GoB, RRRC official governance structure of the camps, taking into account the potential new land allocation. The database is kept as accurate as possible, given these challenges.
- **License:** Public Domain / No Restrictions (<https://data.humdata.org/faqs/licenses>)
- **Tags:** geodata ; populated places-settlements ; refugee crisis ; refugees

- **File formats:** Geodatabase; SHP; KML
- **Content of the layers:** district, upazilla, union, campssid (camp's ID), campname (camp name), smsdcamp, npmcamp, camplabel (camp label), settlement, centroidx (X coordinate of centroid), centroidy (Y coordinate of centroid), areaacres (surface area in acres), areasqm (surface area in square meters), areasqkm (surface area in square kilometers). See:



This information can be entered in the Metadata Editor as follows:

| Information from HDX | Corresponding element in the metadata template |
|------------------------|---|
| Type of dataset | Description / Metadata / Hierarchy level |
| Metadata date | Description / Metadata / Metadata date |
| Metadata language | Description / Metadata / Language |
| Metadata contact | Description / Metadata / Metadata contact |
| Metadata standard name | Description / Metadata / Metadata standard name |
| Dataset website | Description / Metadata / Dataset URI |
| Metadata update | Description / Metadata / Metadata maintenance and update frequency |
| Title | Description / Identification / Dataset identification / Title |
| Source | Description / Identification / Dataset identification / Responsible party / Org. name |

| Information from HDX | Corresponding element in the metadata template |
|----------------------------|---|
| Time period of the dataset | Description / Identification / Dataset identification / Date (creation and lastUpdate) |
| Modified | Description / Identification / Dataset identification / Date (released) |
| Description | Description / Identification / Dataset identification / Abstract |
| Topics | Description / Identification / Dataset identification / Topics |
| Keywords | Description / Identification / Dataset identification / Keywords |
| Dataset language | Description / Identification / Dataset identification / Language |
| Presentation form | Description / Identification / Dataset identification / Presentation form |
| Status | Description / Identification / Purpose, credit and status / Status |
| Bounding box | Description / Identification / Extent (g,t,v) / Geographic element / Bounding box |
| Spatial represent. type | Description / Identification / Spatial representation and resolution / Spatial represent. type |
| Expected update frequency | Description / Identification / Frequency of update / Resource maintenance |
| License | Description / Identification / Legal constraints / Use constraints |
| Methodology | Description / Data quality / Data quality / Lineage statement (scope = Dataset) |

After entering all available information, click on **SAVE**. Click on *Preview* in the navigation tree to view all information you have entered so far.

Step 3: Add external resources

You can now finalize the documentation of the dataset by documenting and attaching *external resources*. External resources include all materials (files or links) you want to make accessible to users when you publish the dataset in a catalog. In this example, we will only add one resource: a link to the HDX website.

To create an external resource, select *External resources* in the navigation tree and then click on **Create resource**. Most elements available to describe an external resource are optional, but at a minimum, you should enter the **Resource type** ("Web Site" in this case), the **Title** (*HDX Data Platform*), and the URL in the section *Resource attachment*. (<https://data.humdata.org/dataset/outline-of-camps-sites-of-rohingya-refugees-in-cox-s-bazar-bangladesh>). Then click **SAVE**. You will now have the external resource listed.

The screenshot shows the 'External resources' section of the Metadata Editor. The dataset title is 'Outline of camps of Rohingya refugees in Cox's Bazar, Bangladesh'. The sidebar on the left has 'External resources' selected. A single resource from 'HDX Data Platform' is listed, with a link to 'https://data.humdata.org/dataset/outline-of-camps-sites-of-rohingya-refugees-in-cox-s-bazar-bangladesh'. There are buttons for 'CREATE RESOURCE' and 'IMPORT' at the top right.

With this, you have completed the documentation of the dataset (note that in practice, you would generate more detailed metadata than what we did in this example). The *My Projects* page will show this new entry. You may at any time go back to it to edit or complete the metadata.

Step 4: Export and publish metadata

In the *Project* page, a menu of options is available to you.

The screenshot shows the 'Untitled' project page. The 'SAVE' button is circled in red. The sidebar contains the following options:

- Project**
 - Export package (ZIP)
 - Export JSON
 - Publish to NADA
 - PDF documentation
 - Change log
- Metadata**
 - Apply default values from template
 - Import project metadata
 - Import external resources
- External resources**
 - Export RDF/XML
 - Export RDF/JSON

• Export package (ZIP)

This option will allow you to generate a ZIP file containing all metadata and resources related to the project. This package can be shared with others, who can import it in their own Metadata Editor.

- **Export JSON**

Export metadata to JSON will generate a JSON file containing the metadata. The option is provided to include all elements or only the non-private ones. The JSON file will look like this:

Pretty-print

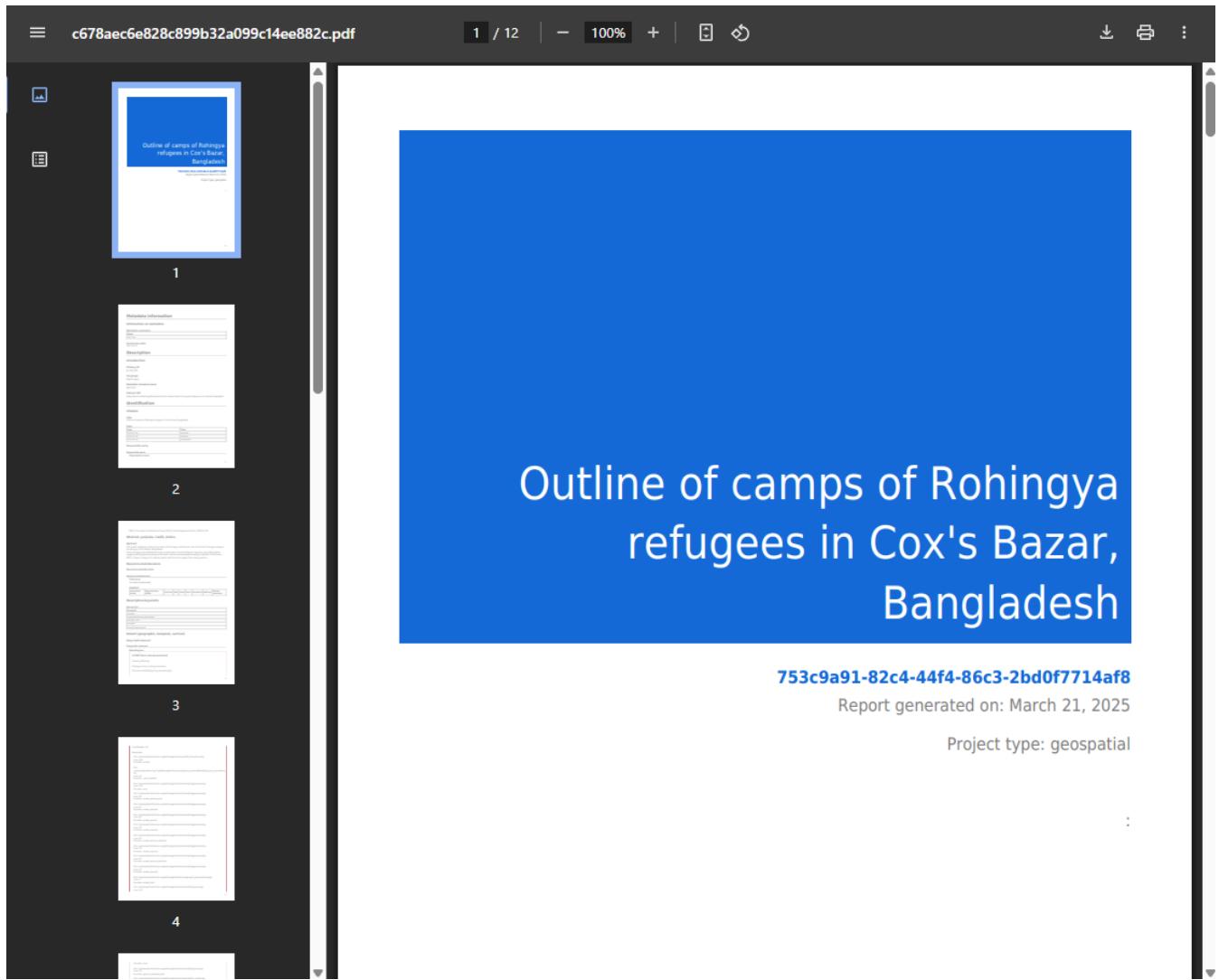
```
{
  "type": "geospatial",
  "idno": "753c9a91-82c4-44f4-86c3-2bd0f7714af8",
  "changed": "1741963168",
  "changed_utc": "2025-03-14T14:39:28+00:00",
  "created": "1739562229",
  "created_utc": "2025-02-14T19:43:49+00:00",
  "created_by": "11",
  "changed_by": "11",
  "metadata_information": {
    "producers": [
      {
        "name": "John Doe"
      }
    ],
    "production_date": "2015-02-14"
  },
  "description": {
    "idno": "JD_GEO_001",
    "language": "English [eng]",
    "dataSetURI": "https://data.humdata.org/dataset/outline-of-camps-sites-of-rohingya-refugees-in-cox-s-bazar-bangladesh",
    "identificationInfo": {
      "citation": {
        "citedResponsibleParty": [
          {
            "organisationName": "RRRC, Inter Sector Coordination Group (ISCG), Site Management Sector, UNHCR, IOM"
          }
        ],
        "date": [
          {
            "date": "2024-05-19",
            "type": "released"
          },
          {
            "date": "2018-01-20",
            "type": "creation"
          },
          {
            "date": "2024-04-05",
            "type": "lastUpdate"
          }
        ],
        "title": "Outline of camps of Rohingya refugees in Cox's Bazar, Bangladesh"
      },
      "resourceMaintenance": [
        {
          "maintenanceAndUpdateFrequency": "As needed [asNeeded]"
        }
      ],
      "resourceFormat": [
        {
          "name": "Geodatabase"
        }
      ]
    }
  }
}
```

- **Export RDF/XML and Export RDF/XML**

These options allow you to export the metadata related to external resources in JSON or XML format.

- **PDF documentation**

A PDF version of the metadata can be automatically created. Select PDF documentation then click on [GENERATE PDF](#). When the PDF is generated, click on [DOWNLOAD PDF](#). You will obtain a bookmarked PDF file with all entered metadata.



- **Publish to NADA**

If you have a NADA catalog and the credentials to publish content in it, you can also [Publish to NADA](#). Select a configured NADA catalog, select the options as shown in the screenshot below, and click [PUBLISH](#).

The screenshot shows the Metadata Editor interface with the following details:

- Left sidebar:** Shows navigation sections: Home, Preview, Document description, Description, Tags, External resources (selected), HDX Data Platform, Geospatial features, Image Gallery, and Administrative metadata.
- Top right:** Includes links for About, English, and User John Doe, along with a SAVE button and a three-dot menu.
- Main content area:**
 - Title:** Outline of camps of Rohingya refugees in Cox's Bazar, Bangladesh
 - Publish to NADA:** Publish project directly to a NADA catalog. Catalog: NADA demo - <https://nada-demo.ihsn.org>.
 - Project options:**

| Option | Value |
|------------------------------|--------------------------|
| Overwrite if already exists? | Yes |
| Publish | Publish |
| Data access | Direct access - [direct] |
| Collection | N/A |
 - External resources:** Select external resources to publish.
 - Overwrite resources:** 1 resources found, 1 selected.
 - Title:** [Title](https://data.humdata.org/dataset/outline-of-camps-sites-of-rohingya-refugees-in-cox-s-bazar-bangladesh) (Type: Web Site [web])
 - HDX Data Platform:** <https://data.humdata.org/dataset/outline-of-camps-sites-of-rohingya-refugees-in-cox-s-bazar-bangladesh>
 - Options:**
 - Publish project
 - Publish thumbnail
 - External resources (1)
 - Published project link:** https://nada-demo.ihsn.org/index.php/catalog/study/JD_GEO_001
 - PUBLISH** button.

The dataset will now be listed and made discoverable in the NADA catalog.

Demo NADA Catalog

Data Catalog

[Home](#) [Catalog](#) [Collections](#) [Citations](#) [How to?](#) [!\[\]\(ee63952771dcddb5555d089c684847fd_img.jpg\) Login](#)[Home](#) / [Central Data Catalog](#) / JD_GEO_001

Outline of camps of Rohingya refugees in Cox's Bazar, Bangladesh

Bangladesh, 2018

Reference ID JD_GEO_001

Metadata

[JSON](#)CREATED ON
Feb 14, 2025LAST MODIFIED
Feb 14, 2025PAGE VIEWS
1[STUDY DESCRIPTION](#)[DOWNLOADS](#)[GET MICRODATA](#)[Identification](#)[Spatial extent](#)[Distribution](#)[Data quality](#)[Metadata](#)

Identification

TITLE

Outline of camps of Rohingya refugees in Cox's Bazar, Bangladesh

DATE

| Date | Type |
|------------|------------|
| 2024-05-19 | released |
| 2018-01-20 | creation |
| 2024-04-05 | lastUpdate |

RESPONSIBLE PARTY

RRRC, Inter Sector Coordination Group (ISCG), Site Management Sector, UNHCR, IOM

ABSTRACT

This spatial database contains the outline of the camps, settlements, and sites where Rohingya refugees are staying in Cox's Bazar, Bangladesh.

These polygons were digitized through a combination of methodologies, originally using VHR satellite imagery and GPS points collected in the field, verified and amended according to Shelter-CCCM Sector, RRRC, Camp in Charge (CIC) officers inputs, with technical support from other partners.

Quick start: Research projects and scripts

In this Quick start example, we will document a research project titled *Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers*. The scripts related to the project are available in the following GitHub repository: <https://github.com/worldbank/double-jeopardy-in-langs>. A working paper was published on arXiv, available at <https://arxiv.org/abs/2410.10665>. The project used three datasets, one of which is available as open data.

The only file you need to reproduce this Quick-Start example is the image file `.../quick_start_files/scripts/lilm_jeopardy_research.jpg` (or select any another JPG or PNG image of your choice).

This Quick Start section does not include detailed guidance on documenting research projects and scripts. For comprehensive instructions, see the chapter **Documenting Data – Research projects and scripts**.

Step 1: Create a new project and add a thumbnail

To begin, open the Metadata Editor in your web browser (the URL is determined by where you installed the application), and log in with your username and password. The *My projects* page will be displayed, showing all projects you have previously created and those that have been shared with you by other data curators, if any. If you are using the application for the first time and no project has been shared with you, the project list will be empty.

| | Title | Owner | Last modified | Modified | Actions |
|--------------------------|--|----------|---------------|------------|----------------------------------|
| <input type="checkbox"/> |  Reshaping Food Security Analytics: An In-Depth Look at Emerging Solutions for Rapid Insights adcf62b2-2fbf-4155-ab63-36827fdcca69 | John Doe | John Doe | 2025-02-20 | <input type="button" value="⋮"/> |
| <input type="checkbox"/> |  Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population) 4333f56f-6614-4470-b4c3-d84c699495ce | John Doe | John Doe | 2025-02-20 | <input type="button" value="⋮"/> |
| <input type="checkbox"/> |  Popstan Synthetic Household Survey 2023 Generating a dozen random households | John Doe | John Doe | 2025-02-19 | <input type="button" value="⋮"/> |

Click on **CREATE NEW PROJECT** and select **Script** when prompted to indicate the type of resource you will be documenting (a research project and scripts).

Create new project

 Microdata

 Timeseries

 Timeseries database

 Document

 Table

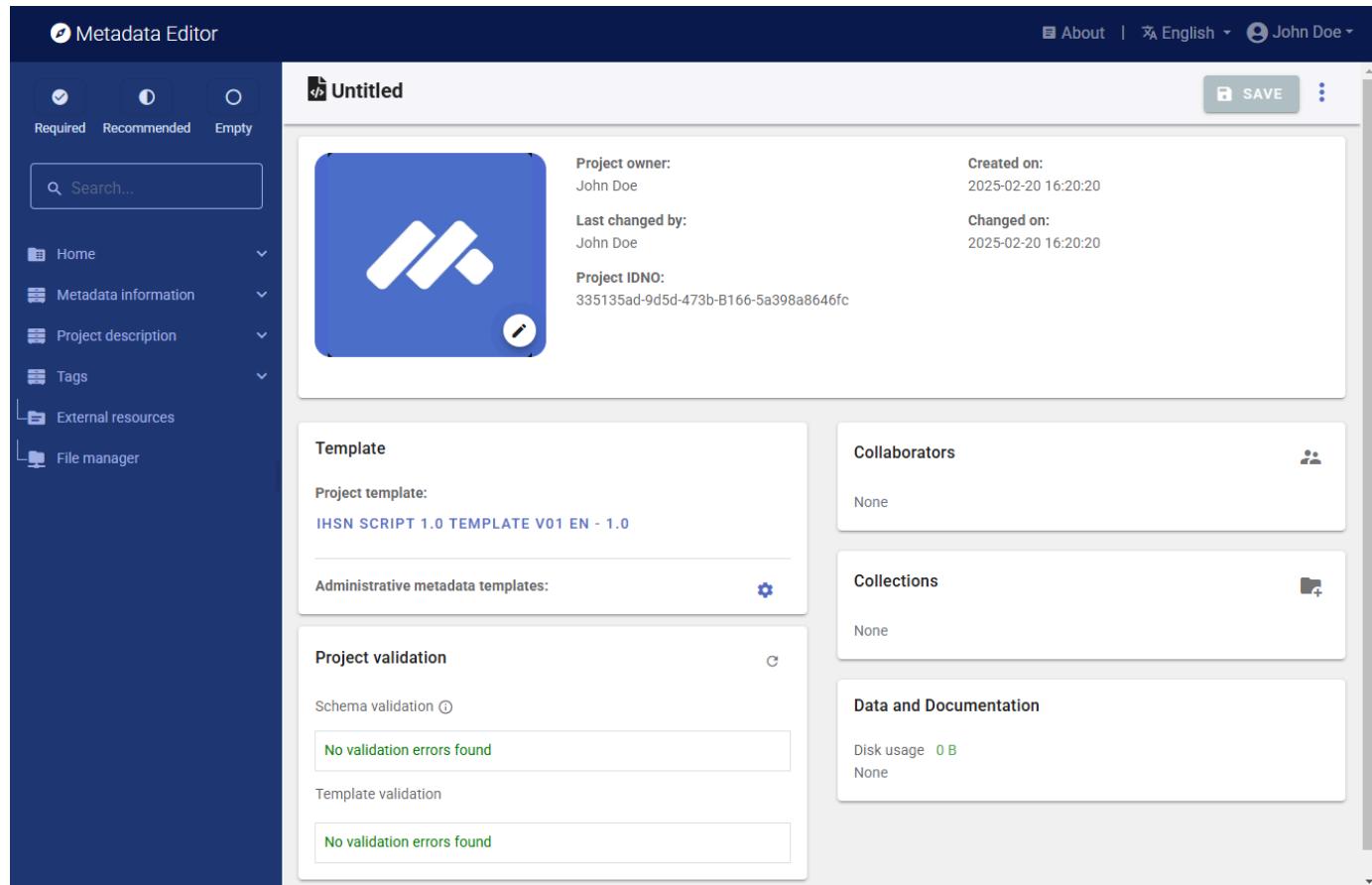
 Image

 Script

 Video

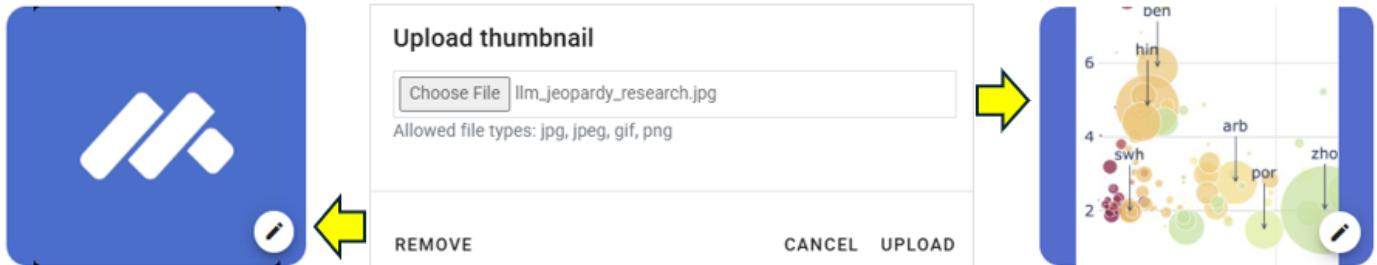
 Geospatial

A new project page will open in a new tab.



The screenshot shows the Metadata Editor application window. At the top, there's a header bar with the title 'Metadata Editor', a user profile icon, and language selection ('English'). Below the header is a sidebar on the left containing navigation links: 'Required', 'Recommended', 'Empty', 'Search...', 'Home', 'Metadata information', 'Project description', 'Tags', 'External resources', and 'File manager'. The main content area is titled 'Untitled'. It features a large thumbnail placeholder with a pencil icon, labeled 'Untitled'. To the right of the thumbnail, there are several metadata fields: 'Project owner: John Doe', 'Created on: 2025-02-20 16:20:20'; 'Last changed by: John Doe', 'Changed on: 2025-02-20 16:20:20'; and 'Project IDNO: 335135ad-9d5d-473b-B166-5a398a8646fc'. Below the thumbnail, there are sections for 'Template' (Project template: 'IHSN SCRIPT 1.0 TEMPLATE V01 EN - 1.0'), 'Collaborators' (None), 'Collections' (None), and 'Data and Documentation' (Disk usage: 0 B). The 'Template' section also includes 'Administrative metadata templates' and 'Project validation' (Schema validation: 'No validation errors found', Template validation: 'No validation errors found').

We will use the *ilm_jeopardy_research.jpg* image (or an image of your choice) as a thumbnail. Note that providing a thumbnail is not required, but recommended. The thumbnail will be displayed in the Metadata Editor project list, and in the NADA catalog if the metadata is published in NADA. Click on the  icon in the screenshot image, and select the image file when prompted.



Step 2: Enter metadata

On the left navigation tree, select *Metadata information / Information on metadata* to enter optional elements used to capture information on who documented the research project, and when. Enter your name (as **Metadata producer**) and the **date** of the day in ISO format YYYY-MM-DD. This is the date when the metadata, not the research work, was produced. Then click on **SAVE**.

You can now start entering the metadata related to the project itself. In the navigation tree, first select *Title statement* under *Project description*, and enter the required **Primary ID** (a unique identifier of your choice, e.g., JD_SCRIPT_001; if you want to publish the project in a NADA catalog, make sure that this same identifier is not used by another user or for another project). Also, enter the **title** of the project: "Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers". Click **SAVE**.

The screenshot shows the 'Title statement' section of the Metadata Editor. The left sidebar has a tree view with 'Title statement' selected. The main area contains fields for Primary ID (set to JD_SCR_001), Other identifiers (an empty table with a '+ ADD ROW' button), Title (containing the text 'Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers'), Subtitle (empty), Alternate title (empty), Translated title (empty), and Project website (set to 'https://github.com/worldbank/double-jeopardy-in-langs'). A 'SAVE' button is located in the top right corner.

Then proceed with the other sections in the navigation tree and fill out the following elements.

- **Date:** October 2024 (2024-10 in ISO format)
- **Website:** <https://github.com/worldbank/double-jeopardy-in-langs>
- **Authors:** Aivin V. Solatorio, Gabriel Stefanini Vicente, Holly Krambeck, Olivier Dupriez (all are affiliated with the World Bank)
- **Abstract (extracted from the working paper):** Artificial Intelligence (AI), particularly large language models (LLMs), holds the potential to bridge language and information gaps, which can benefit the economies of developing nations. However, our analysis of FLORES-200, FLORES+, Ethnologue, and World Development Indicators data reveals that these benefits largely favor English speakers. Speakers of languages in low-income and lower-middle-income countries face higher costs when using OpenAI's GPT models via APIs because of how the system processes the input -- tokenization. Around 1.5 billion people, speaking languages primarily from lower-middle-income countries, could incur costs that are 4 to 6 times higher than those faced by English speakers. Disparities in LLM performance are significant, and tokenization in models priced per token amplifies inequalities in access, cost, and utility. Moreover, using the quality of translation tasks as a proxy measure, we show that LLMs perform poorly in low-resource languages, presenting a double jeopardy of higher costs and poor performance for these users. We also discuss the direct impact of fragmentation in tokenizing low-resource languages on climate. This underscores the need for fairer algorithm development to benefit all linguistic groups.
- **Geographic coverage:** World (code WLD)
- **Keywords:** Large Language Models (LLMs), Low-resource languages, Inequity in access, Tokenization
- **Project output:** Working paper published on ArXiv, titled "Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers", by Aivin V. Solatorio, Gabriel Stefanini Vicente, Holly Krambeck, and Olivier Dupriez. Note that the abstract of the output is the same as the abstract for the project.
- **Data:** The list of datasets used in the project is provided in the GitHub repository:
 - **FLORES-200 and FLORES+ :** A multilingual dataset covering 100 languages, with 1,000 sentences per language. Used for evaluating translation quality and computing the tokenization premium relative to English. (URL: <https://github.com/facebookresearch/flores>)
 - **Ethnologue :** Provides linguistic data, including the number of speakers, geographic distribution, and writing systems. We use Ethnologue to estimate the number of speakers for each language. (URL: <https://www.ethnologue.com/>)

- **World Bank, World Development Indicators (WDI)** : Contains socio-economic data at the country level. Specifically, the project used the GDP per capita in current US\$ (NY.GDP.PCAP.CD) and the annual population growth rates (SP.POP.GROW) indicators to compute the population-weighted GDP for each language and for aligning population estimates to 2022 based on historical figures from Ethnologue. (URL: <https://datacatalog.worldbank.org/dataset/world-development-indicators>)
- **OpenAI GPT-4o and GTP-4 Turbo APIs** : Used to assess the reduced utility of LLMs for non-English speakers. The research project applied translation with different prompting methods to generate reference translations for FLORES sentences. The LLM translated non-English sentences into English, with the original English sentences serving as a benchmark for evaluating translation quality. (URL: <https://openai.com/api/>)
- **Software**: Python 3.0
- **Scripts**: Scripts are provided as Python notebooks (all published under Mozilla Mozilla Public License / URL: <https://www.mozilla.org/en-US/MPL/>):

 - **Tokenization of FLORES dataset** (compute-premium-costs.ipynb): Computes the tokenization premium for the FLORES dataset. The calculation of the population-weighted GDP for each language is also done in this notebook.
 - **Back-translation task for the FLORES dataset** (back-translation-task.ipynb): Generates the back-translation task for the FLORES dataset. The notebook implements the batched translation strategy for the translation task and uses the OpenAI GPT-4o API.
 - **Additional analysis of the results (analysis.ipynb)**: Notebook for additional analysis of the results. Key visualizations are generated in this notebook, including the comparison of the tokenization premiums between two different tokenizers (GPT-4o vs. GPT-4 Turbo).

- **Technology environment and requirement**: This work has been developed using a MacBook Pro with an M1 Pro processor and 64GB of RAM. No GPU is needed for the computations. Access to the OpenAI API is required.
- **Github repo**: double-jeopardy-in-langs (URL: <https://github.com/worldbank/double-jeopardy-in-langs/tree/main>)
- **Licensing**: Mozilla Public License (URL: <https://www.mozilla.org/en-US/MPL/>)

This information is entered in the metadata template as follows:

| Information | In the metadata template |
|---------------------|---|
| Identifier | Project description / Title statement / Primary ID |
| Title | Project description / Title statement / Title |
| Website | Project description / Title statement / Project website |
| Date | Project description / Version statement / Project completion date |
| Authors | Project description / Authors and contributors / Authoring entity |
| Abstract | Project description / Scope and coverage / Abstract |
| Geographic coverage | Project description / Scope and coverage / Geographic areas |
| Keywords | Project description / Scope and coverage / Keywords |
| Project output | Project description / Processes and output / Output |
| Data | Project description / Data / Datasets |
| Software | Project description / Methods, Software and scripts / Software |

| Information | In the metadata template |
|-------------|--|
| Scripts | Project description / Methods, Software and scripts / Scripts |
| GitHub repo | Project description / Methods, Software and scripts / Repository |
| Licensing | Project description / Access and rights / License |

After entering all available information, click on [SAVE](#). Click on *Preview* in the navigation tree to view all information you have entered so far.

The screenshot shows the Metadata Editor interface with a dark blue header bar. On the left, a sidebar navigation tree includes Home, Preview (which is selected), Metadata information, Project description, Title statement, Primary ID, Other identifiers, Title, Subtitle, Alternate title, Translated title, Project website, Authors and contributors, Reproducibility status, Version statement, and Project completion date. The main content area displays a project titled "Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Dis...". It shows "Metadata information" and "Information on metadata". Under "Information on metadata", there is a table for "Metadata producers" with one row for "John Doe". Below it, "Production date" is listed as "2025-03-14". In the "Project description" section, the "Title statement" is detailed with "Primary ID" (JD_SCR_001), "Title" ("Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers"), and "Project website" (<https://github.com/worldbank/double-jeopardy-in-langs>). The "Authors and contributors" section lists "John Doe" as the producer. At the top right, there are buttons for "About", "English", "John Doe", "SAVE", and a help icon. A "Download HTML" link is also visible.

Step 3: Add external resources

Once you have entered the metadata, you can finalize the documentation of the project by documenting and attaching external resources. External resources include all materials (files or links) that you want to make accessible to users when you publish the project in a catalog. In this example, we will add two external resources: a link to the research project GitHub repository, and a link to the research output (the working paper).

To create external resources, click on *External resources* in the navigation tree and then click on [Create resource](#).

- For the GitHub repository, create a new resource then select the [Resource type](#) ("Website"), give it a short [Title](#) (*GitHub repository: double-jeopardy-in-langs*), and enter the URL in [Resource attachment](#) (<https://github.com/worldbank/double-jeopardy-in-langs/tree/main>). Then click [SAVE](#).
- For the working paper, create a new resource then select the [Resource type](#) ("Document, analytical"), enter the [Title](#) (**Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers*)*, and enter the URL in [Resource attachment](#) (*<https://arxiv.org/abs/2410.10665>*). You may also enter the name of the authors, date (2024-10-14), and abstract. Then click [SAVE](#).

You will now have two external resources listed.

Step 4: Export and publish metadata

In the *Project* page, a menu of options is available to you.

| Project | Metadata |
|--|--|
| <input checked="" type="checkbox"/> Export package (ZIP) | <input checked="" type="checkbox"/> Apply default values from template |
| <input type="checkbox"/> Export JSON | <input type="checkbox"/> Import project metadata |
| <input checked="" type="checkbox"/> Publish to NADA | <input type="checkbox"/> Import external resources |
| <input checked="" type="checkbox"/> PDF documentation | External resources |
| <input type="checkbox"/> Change log | <input type="checkbox"/> Export RDF/XML |
| | <input type="checkbox"/> Export RDF/JSON |

- **Export package (ZIP)**

This option will allow you to generate a ZIP file containing all metadata and resources related to the project. This package can be shared with others, who can then import it in their own Metadata Editor.

- **Export JSON**

Export metadata to JSON will generate a JSON file containing the metadata. The option is provided to include all elements or only the non-private ones. The JSON file will look like this:

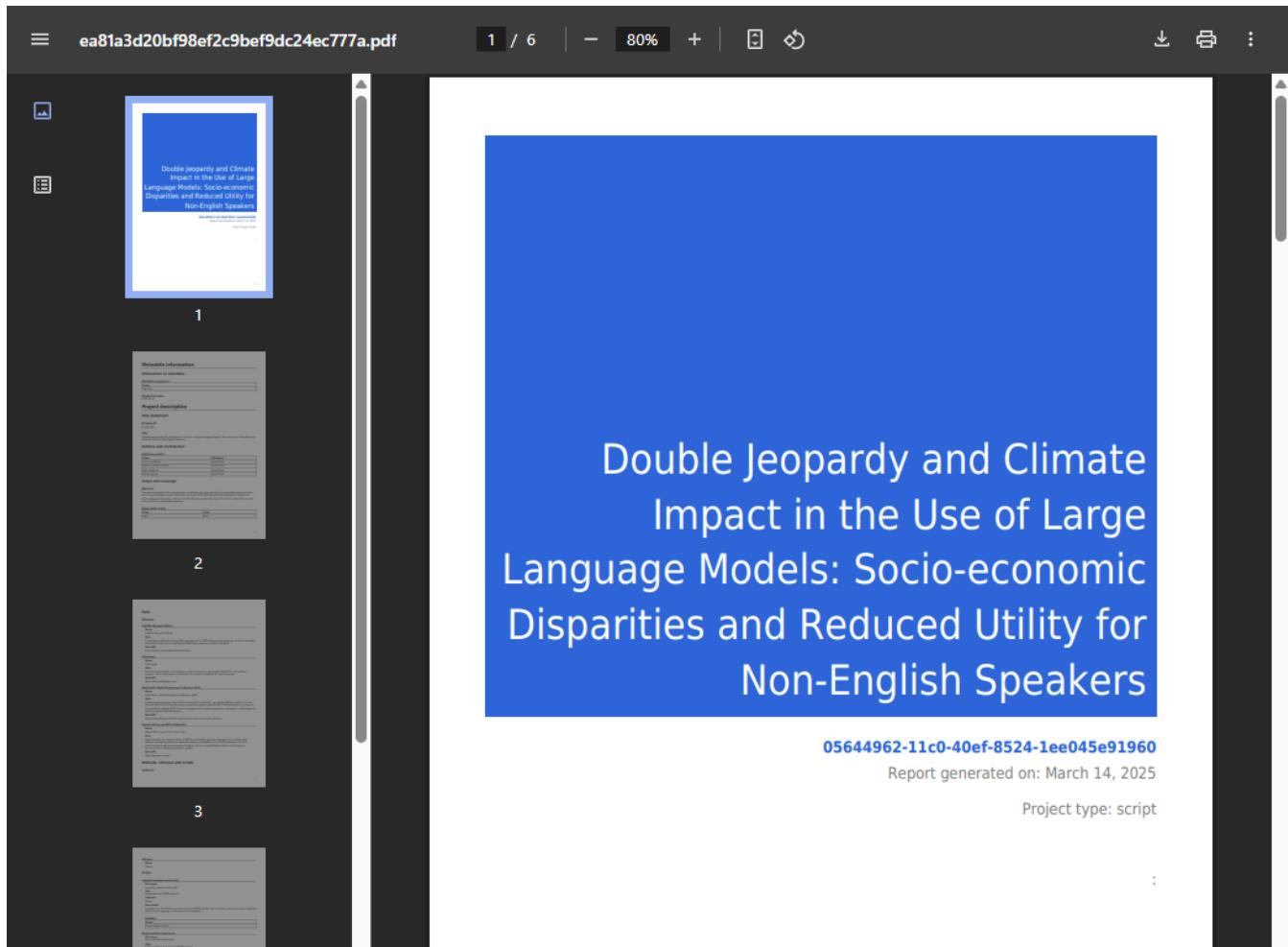
```
Pretty-print 
{
  "type": "script",
  "idno": "0564962-11c0-40ef-8524-1ee045e91960",
  "changed": "1741980469",
  "changed_utc": "2025-03-14T19:27:49+00:00",
  "created": "1740086575",
  "created_utc": "2025-02-20T21:22:55+00:00",
  "created_by": "11",
  "changed_by": "11",
  "doc_desc": {
    "producers": [
      {
        "name": "John Doe"
      }
    ],
    "prod_date": "2025-03-14"
  },
  "project_desc": {
    "title_statement": {
      "title": "Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers",
      "idno": "JD_SCR_001"
    },
    "project_website": "https://github.com/worldbank/double-jeopardy-in-l1ms",
    "authoring_entity": [
      {
        "name": "Aivin V. Solatorio",
        "affiliation": "World Bank"
      },
      {
        "name": "Gabriel Stefanini Vicente",
        "affiliation": "World Bank"
      },
      {
        "name": "Holly Krambeck",
        "affiliation": "World Bank"
      },
      {
        "name": "Olivier Dupriez",
        "affiliation": "World Bank"
      }
    ],
    "abstract": "This work investigates the socio-economic disparities and reduced utility for non-English speakers in the use of large language models (LLMs). We also use the OpenAI's GPT-4 API to assess the reduced utility of LLMs for non-English speakers.",
    "geographic_units": [
      {
        "name": "World",
        "code": "WLD"
      }
    ],
    "datasets": [
      {
        "name": "FLORES-200 and FLOREST",
        "note": "A multilingual dataset covering 100 languages with 1,000 sentences per language. Used for evaluating translation quality and computing the token"
      }
    ]
  }
}
```

- **Export RDF/XML and Export RDF/XML**

These options allow you to export the metadata related to external resources in JSON or XML format.

- **PDF documentation**

A PDF version of the metadata can be automatically created. Select *PDF documentation* then click on **GENERATE PDF**. When the PDF is generated, click on **DOWNLOAD PDF**. You will obtain a bookmarked PDF file with all entered metadata.



- **Publish to NADA**

If you have a NADA catalog and the credentials to publish content in it, you can also "Publish to NADA". Select a configured NADA catalog, select the options as shown in the screenshot below, and click PUBLISH.

Publish to NADA

Publish project directly to a NADA catalog

Catalog Configure new catalog

| |
|--|
| NADA demo - https://nada-demo.ihsn.org |
| {"id": "16", "title": "NADA demo", "url": "https://nada-demo.ihsn.org", "user_id": "11"} |

Project options

| Option | Value |
|------------------------------|--------------------------|
| Overwrite if already exists? | Yes |
| Publish | Published |
| Data access | Direct access - [direct] |
| Collection | N/A |

External resources
Select external resources to publish

Overwrite resources

2 resources found 2 selected

| Title | Type |
|---|---------------------------------|
| GitHub repository: double-jeopardy-in-llms https://github.com/worldbank/double-jeopardy-in-llms/tree/main | Web Site [web] |
| Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers https://arxiv.org/abs/2410.10665 | Document, Analytical [doc/anal] |

Options

- Publish project
- Publish thumbnail
- External resources (2)

PUBLISH

The project will now be listed and made discoverable in the NADA catalog.

! [image](img/ME_UG_v1-0-0_quick_start_script_indicator_in_NADA.png)

Quick Start: Image

In this example, we will document an image extracted from the World Bank photo collection in Flickr (<https://www.flickr.com/photos/worldbank/14131666634/in/album-72157626025379650>). This image shows a tomato stand in a market near Ramallah's main mosque. We assume that you want to publish the image in a data catalog, with a link to the World Bank's photo album.

The only file you need to reproduce this Quick-Start example is the image file .../image/wb_photo_food_market.jpg (feel free to use another image of your choice).

This Quick Start section does not include detailed guidance on documenting images. For comprehensive instructions, see the chapter **Documenting Data – Images**.

Step 1: Create a new project and add a thumbnail

To begin, open the Metadata Editor in your web browser (the URL is determined by where you installed the application), and log in with your username and password. The My projects page will be displayed, showing all projects you have previously created and those that have been shared with you by other data curators, if any. If you are using the application for the first time and no project has been shared with you, the project list will be empty.

| Type | Title | Owner | Last modified | Modified | Actions |
|-----------|--|----------|---------------|------------|----------------|
| Microdata | Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population) 4333f56f-6614-4470-b4c3-d84c699495ce | John Doe | John Doe | 2025-02-13 | ⋮ |
| Table | Popstan Synthetic Household Survey 2023 d1cff314-d3f0-4ea7-bdd6-b39966721646 | John Doe | John Doe | 2025-02-12 | ⋮ |
| Image | The Analysis of Household Surveys 8e05b200-1753-4c41-8fcfa-5ee50197694b | John Doe | John Doe | 2025-02-12 | ⋮ |

Click on **CREATE NEW PROJECT** and select *Image* when prompted to indicate the type of resource you will be documenting.

Create new project

 Microdata

 Timeseries

 Timeseries database

 Document

 Table

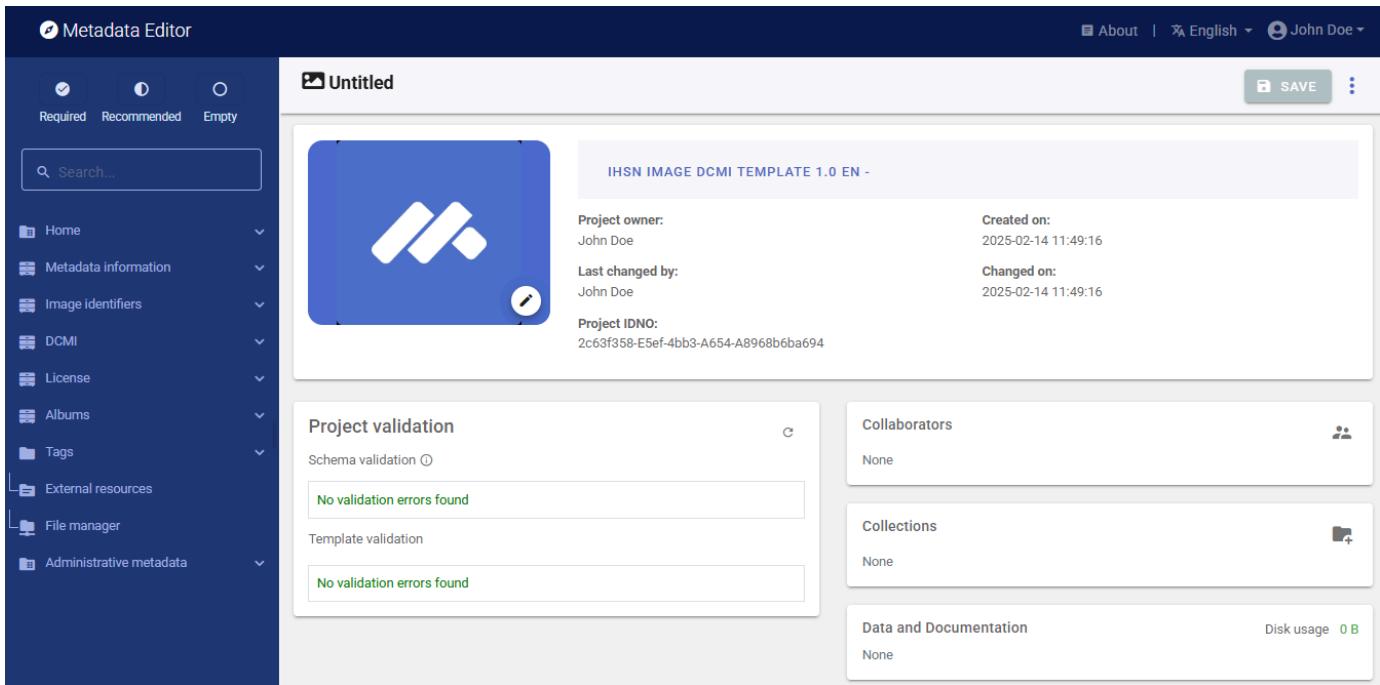
 Image

 Script

 Video

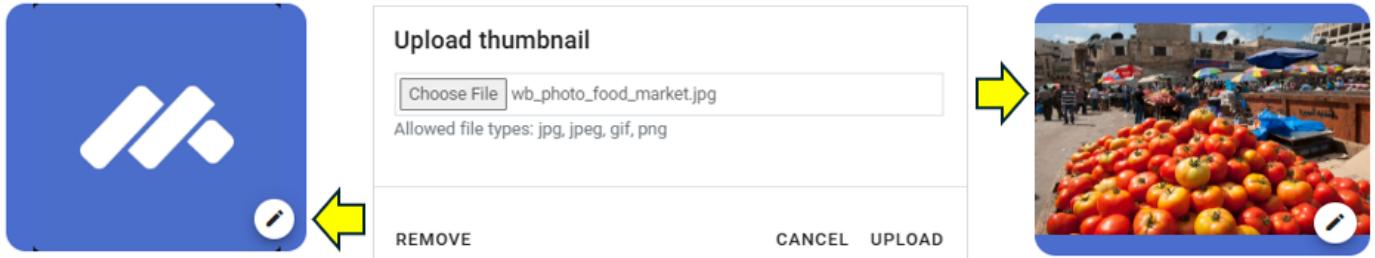
 Geospatial

A new project page will open in a new tab.



The screenshot shows the Metadata Editor interface. On the left is a sidebar with navigation links: Home, Metadata information, Image identifiers, DCMI, License, Albums, Tags, External resources, File manager, and Administrative metadata. Above the sidebar are three buttons: Required (with a checkmark), Recommended, and Empty. A search bar is also present. The main area is titled 'Untitled' and contains a thumbnail placeholder for the project. To the right of the thumbnail, the title 'IHSN IMAGE DCMI TEMPLATE 1.0 EN -' is displayed. Below the title, project details are listed: Project owner: John Doe, Created on: 2025-02-14 11:49:16; Last changed by: John Doe, Changed on: 2025-02-14 11:49:16; Project IDNO: 2c63f358-E5ef-4bb3-A654-A8968b6ba694. The main area is divided into several sections: Project validation, Collaborators, Collections, and Data and Documentation. The Project validation section shows 'No validation errors found' for both Schema validation and Template validation. The Collaborators section shows 'None'. The Collections section shows 'None'. The Data and Documentation section shows 'Disk usage 0 B'.

We will use the image itself as a thumbnail. Providing a thumbnail is not required, but recommended. The thumbnail will be displayed in the Metadata Editor project list, and in the NADA catalog if the metadata is published in NADA. Click on the  icon in the screenshot image, and select the image file when prompted.



The Metadata Editor offers two options to document images, corresponding to two different metadata standards: the Dublin Core (DCMI), or the IPTC. The decision to use one or the other option is made by selecting a DCMI-based or IPTC-based template. An image should not be documented using both options. In this Quick-start example, we will use the default DCMI metadata template, so there is no need to switch template (if the template shown in the *Template selection* frame is not the DCMI template, click on the template name and select a DCMI template from the list).

Step 2: Enter metadata

In the navigation tree, select *Metadata information / Information on metadata* to enter information on who documented the image and when. All information in this section is optional. Enter your name (as **Metadata producer**) and the **date** of the day in ISO format YYYY-MM-DD. This is the date when the metadata, not the image, was produced. Then click on **SAVE**.

| Name | Abbreviation | Affiliation | Role |
|----------|--------------|-------------|------|
| John Doe | | | |

You can now start entering the metadata related to the image itself. In the navigation tree, first select *Image identifiers* and enter the required **Primary ID** (if you want to publish the image in a NADA catalog, make sure that the identifier is not used by another user or for another project). Also enter the (optional) **Other identifiers** for the image. In this example, we have one other identifier, the one provided in the World Bank Flickr album: *Hoel_121012_DSC_3684*.

The screenshot shows the 'Identifiers' section of the Metadata Editor. The left sidebar has 'Image identifiers' selected under 'Identifiers'. The main area shows a table for 'Other identifiers' with one row added: Type 'World Bank Flickr' and Identifier 'Hoel_121012_DSC_3684'. A 'SAVE' button is at the top right.

Then proceed with the other sections in the navigation tree and fill out the following metadata elements using the following information provided in the World Bank Flickr album:

- **Title:** Market near Ramallah's main mosque
- **ID:** Hoel_121012_DSC_3684
- **Description:** Tomato stand in market near Ramallah's main mosque
- **Photographer:** Arne Hoel / World Bank
- **Taken on:** October 12, 2012 (2012-10-12 in ISO format)
- **Tags:** Middle East; Private Sector Development; West Bank & Gaza; market; Food; Tomato
- **Resource type:** Digital photo
- **Format:** JPG
- **License:** CC BY-NC-ND 2.0 (URL: <https://creativecommons.org/licenses/by-nc-nd/2.0/>)

This information can be entered in the Metadata Editor as follows:

| From World Bank | In the metadata template |
|----------------------------|--|
| Resource type | DCMI / Image description / Resource type |
| Taken on | DCMI / Image description / Date |
| Title | DCMI / Image description / Title |
| Description | DCMI / Image description / Caption |
| Format | DCMI / Image description / Format |
| Tags | DCMI / Image description / Keywords (keyword) |
| (derived from title) | DCMI / Country |
| Photographer (name) | DCMI / Authors and rights / Creator |
| Photographer (affiliation) | DCMI / Authors and rights / Publisher |
| License | License / License (name and URL) |

The screenshot shows the 'Image description' section of the Metadata Editor. The title of the resource is 'Market near Ramallah's main mosque'. The 'Resource type' is set to 'Digital image'. The 'Date' is '2012-10-12'. The 'Title' is 'Market near Ramallah's main mosque'. The 'Caption' is 'Tomato stand in market near Ramallah's main mosque'. There is an empty 'Description' field. Under 'Keywords', there are three entries: 'Middle East', 'Private Sector Development', and 'West Bank & Gaza'. A 'SAVE' button is located at the top right of the form.

Step 3: Add external resources

Once you have entered the descriptive metadata, you can finalize the documentation of the image by documenting and attaching *external resources*. External resources include all materials you want to make accessible to users when you publish the image in a catalog. In this example, we will only add one external resource: the link to the Flickr album.

To create external resources, click on *External resources* in the navigation tree and then click on **Create resource**. Select the resource type ("Web Site" in this case), give it a short title (*World Bank Flickr Album*), and enter the URL (<https://www.flickr.com/photos/worldbank/14131666634/in/album-72157626025379650>). Then click **SAVE**. You will now have two external resources listed.

Note: instead of providing a link to the Flickr album, you could have uploaded the image (in its highest available resolution) as an external resource of type *Photo / image*, and uploaded it on the web server.

Market near Ramallah's main mosque

Edit resource

Resource description

Resource type Web Site

Title World Bank Flickr Album

Author

Date 2025-02-14

Country

Step 4: Export and publish metadata

In the project page, a menu of options will be available to you.

Market near Ramallah's main mosque

Project

- Export package (ZIP)
- Export JSON
- Publish to NADA
- PDF documentation
- Change log

Metadata

- Apply default values from template
- Import project metadata
- Import external resources
- Export RDF/XML
- Export RDF/JSON

- Export package (ZIP)

This option allows you to generate a ZIP file containing all metadata and resources related to the project. This package can be shared with others, who can then import it in their own Metadata Editor.

- **Export JSON**

Export metadata to JSON will generate a JSON file containing the metadata. The option is provided to include all elements or only the non-private ones. The JSON file will look like this:

Pretty-print

```
{
  "type": "image",
  "idno": "2c63f358-e5ef-4bb3-a654-a8968b6ba694",
  "changed": "1739559976",
  "changed_utc": "2025-02-14T19:06:16+00:00",
  "created": "1739551756",
  "created_utc": "2025-02-14T16:49:16+00:00",
  "created_by": "11",
  "changed_by": "11",
  "metadata_information": {
    "producers": [
      {
        "name": "John Doe"
      }
    ],
    "production_date": "2025-02-14"
  },
  "image_description": {
    "idno": "JD_IMG_001",
    "identifiers": [
      {
        "identifier": "Hoel_121012_DSC_3684",
        "type": "World Bank Flickr"
      }
    ],
    "license": [
      {
        "name": "CC BY-NC-ND 2.0",
        "uri": "https://creativecommons.org/licenses/by-nc-nd/2.0/"
      }
    ],
    "dcmi": {
      "type": "Digital image",
      "date": "2012-10-12",
      "title": "Market near Ramallah's main mosque",
      "caption": "Tomato stand in market near Ramallah's main mosque",
      "creator": "Arne Hoel",
      "publisher": "World Bank",
      "country": [
        {
          "name": "West Bank",
          "code": "WBC"
        }
      ]
    }
  }
}
```

- **Export RDF/XML and Export RDF/XML**

These options allow you to export the metadata related to external resources in JSON or XML format.

- **PDF documentation**

A PDF version of the metadata can be automatically created. Select PDF documentation then click on **GENERATE PDF**. When the PDF is generated, click on **DOWNLOAD PDF**. You will obtain a bookmarked PDF file with all entered metadata. Note that such a PDF document is more relevant for data types other than images.

- **Publish to NADA**

If you have a NADA catalog and the credentials to publish content in it, you can also [Publish to NADA](#). Select a configured NADA catalog, select the options as shown in the screenshot below, and click [PUBLISH](#).

The screenshot shows the 'Metadata Editor' interface. On the left is a sidebar with a search bar and a tree view of metadata categories: Home, Preview, Metadata information, Image identifiers, DCMI, License, Albums, Tags, External resources (with 'World Bank Flickr Album' expanded), File manager, and Administrative metadata. The main content area is titled 'Market near Ramallah's main mosque'. It shows a 'Catalog' section with a configuration for 'NADA demo - https://nada-demo.ihsn.org' (id: 16, title: 'NADA demo', url: 'https://nada-demo.ihsn.org', user_id: 11). Below this is a 'Project options' section with fields for Overwrite if already exists (Yes), Publish (Publish), Data access (Direct access - [direct]), and Collection (N/A). Under 'External resources', it says 'Select external resources to publish' and shows 'Overwrite resources' selected. A table lists one resource: 'Title' (World Bank Flickr Album, Type: Web Site [web], URL: https://www.flickr.com/photos/worldbank/14131666634/in/album-72157626025379650). There are three 'Options' radio buttons: Publish project (selected), Publish thumbnail, and External resources (1). At the bottom is a large blue 'PUBLISH' button.

The image will now be listed and made discoverable in the NADA catalog, with a link to the Flickr Album.

The screenshot shows the 'Demo NADA Catalog' Data Catalog page. At the top, there are links for Home, Catalog, Collections, Citations, How to?, and Login. Below that, the breadcrumb navigation shows Home / Central Data Catalog / JD_IMG_001. The main content area features a thumbnail image of a market scene, the title 'Market near Ramallah's main mosque' (West Bank, 2012), and metadata fields: Reference ID (JD_IMG_001), Producer(s) (Arne Hoel), and Metadata (JSON). To the right, a sidebar displays creation and modification dates (both Feb 14, 2025) and a page view count (1). Below the main content are tabs for IMAGE DESCRIPTION, DOWNLOADS, and Identification, with Identification being the active tab. The Identification tab contains sections for TITLE (Market near Ramallah's main mosque), DOWNLOAD (link), and IDNO (JD_IMG_001). The Identification tab also has a sub-section for 'Image description'.

Quick start: Video

In this Quick start example, we will document a video titled *Reshaping Food Security Analytics: An In-Depth Look at Emerging Solutions for Rapid Insights* published on the World Bank Youtube channel (<https://www.youtube.com/watch?v=px1EeqpKDUI&list=PLopq6yGfmFAu3tscprzTPpoPrP1qOE9XZ>). The only file you need to reproduce this Quick-Start example is the image file .../video/video_food_security.jpg, a screenshot of the video introduction.

This Quick Start section does not include detailed guidance on documenting videos. For comprehensive instructions, see the chapter **Documenting Data – Video**.

Step 1: Create a new project and add a thumbnail

To begin, open the Metadata Editor in your web browser (the URL is determined by where you installed the application), and log in with your username and password. The *My projects* page will be displayed, showing all projects you have previously created and those that have been shared with you by other data curators, if any. If you are using the application for the first time and no project has been shared with you, the project list will be empty.

| Showing 1 - 5 of 5 projects | | | | | |
|-----------------------------|--|----------|---------------|------------|----------------------------------|
| | Title | Owner | Last modified | Modified | Actions |
| <input type="checkbox"/> | Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population) 4333f56f-6614-4470-b4c3-d84c699495ce | John Doe | John Doe | 2025-02-20 | <input type="button" value="⋮"/> |
| <input type="checkbox"/> | Popstan Synthetic Household Survey 2023 | John | John Doe | 2025-02-19 | <input type="button" value="⋮"/> |

Click on **CREATE NEW PROJECT** and select *Video* when prompted to indicate the type of resource you will be documenting.

Create new project

 Microdata

 Timeseries

 Timeseries database

 Document

 Table

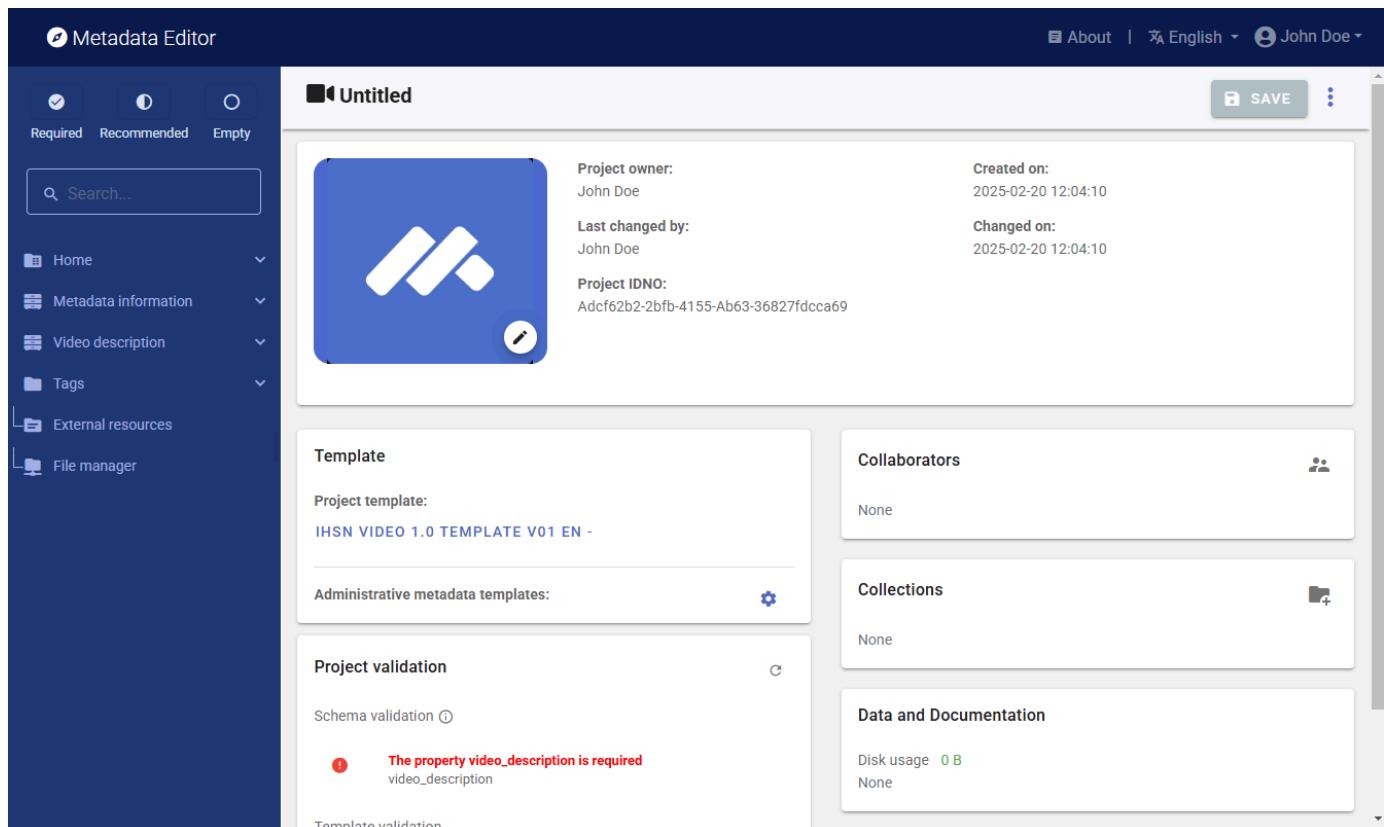
 Image

 Script

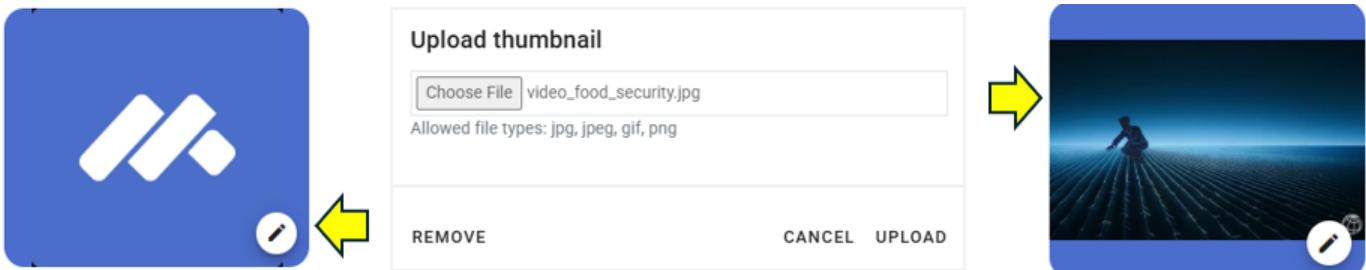
 Video

 Geospatial

A new project page will open in a new tab.



We will use the screenshot image *video_food_security.jpg* as a thumbnail (but feel free so select another JPG or PNG image file). Note that providing a thumbnail is not required, but recommended. The thumbnail will be displayed in the Metadata Editor project list, and in the NADA catalog if the metadata is published in NADA. Click on the  icon in the screenshot image, and select the image file when prompted.



Step 2: Enter metadata

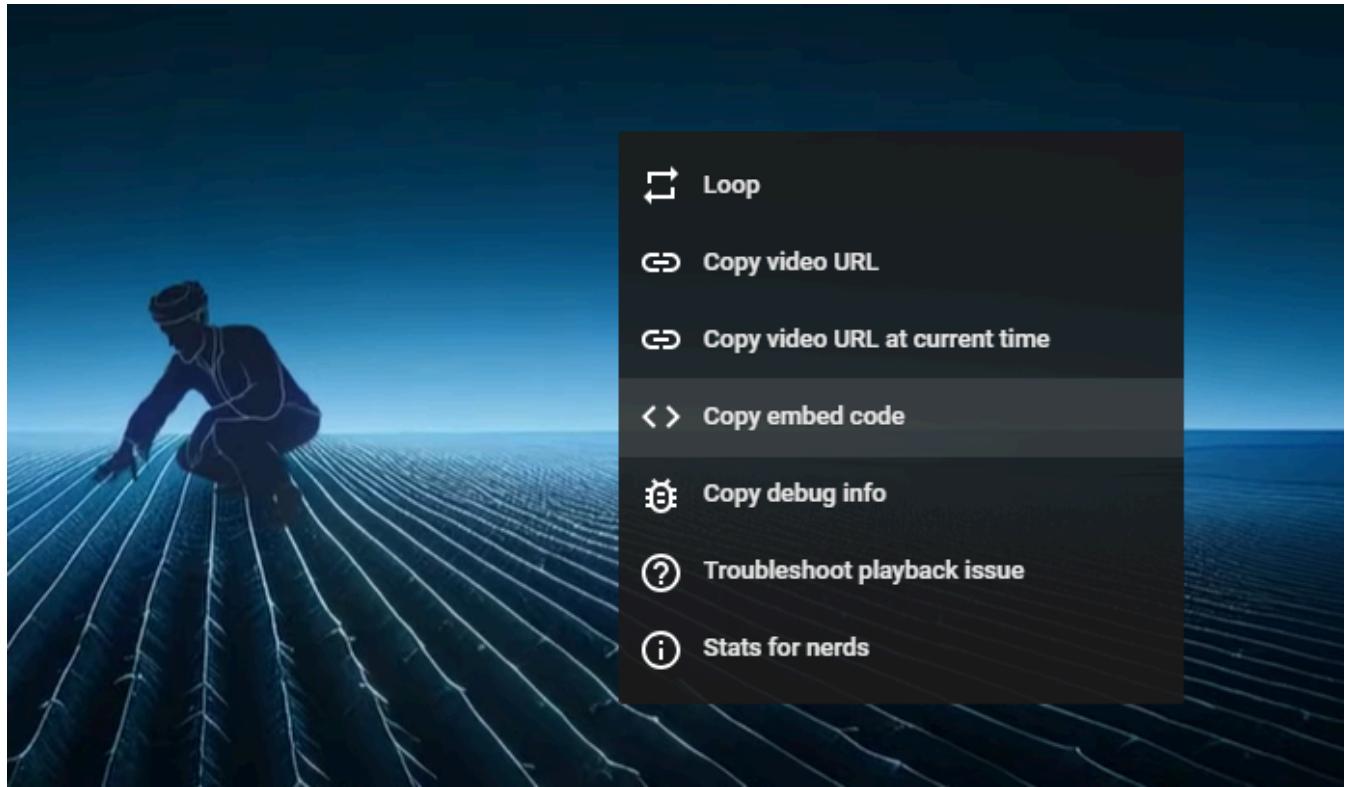
In the navigation tree, select *Metadata information / Information on metadata* to enter information on who documented the publication and when. All information in this section is optional. Enter your name (as *Metadata producer*) and the *date* of the day in ISO format YYYY-MM-DD. This is the date when the metadata, not the video, was produced. Then click on *SAVE*.

You can now start entering the metadata related to the video itself. In the navigation tree, first select *Title statement* under *Video description*, and enter the required *Primary ID*, a unique identifier of your choice, e.g., JD_VDO_001 (if you want to publish the document in a NADA catalog, make sure that the identifier is not used by another user or for another project). Also, enter the *title* of the video: *Reshaping Food Security Analytics: An In-Depth Look at Emerging Solutions for Rapid Insights*.

The screenshot shows the 'Title statement' section of the Metadata Editor. The primary ID is set to JD_VDO_001. The title is Reshaping Food Security Analytics: An In-Depth Look at Emerging Solutions for Rapid Insights. There is a table for 'Other identifiers' with one row added, and an empty 'Alternate title' field.

Then proceed with the other sections in the navigation tree and fill out the following elements using the information below provided in the World Bank Youtube channel:

- **Author:** World Bank
- **Description:** New methods are paving the way for faster data and insight. Get an in-depth look at how AI and machine learning are reshaping food security analytics to deliver actionable information for humanitarian aid, policy-making, and crisis response. From conflict-affected regions to global inflation trends, discover how data-driven solutions—such as the Joint Monitoring Report (JMR), Real-Time Prices (RTP), and the World Food Security Outlook (WFSO)—generate crucial information to provide timely assistance where it matters most.
- **Genre:** Documentary
- **Keywords:** Extract keywords you find relevant from the transcript or from the description of the video provided in the Youtube channel. Some suggestions: "Food Security", "Famine", "Nutrition data", "Food crisis", "Global Alliance for Food security", "Dry corridor", "El Nino", "Hunger".
- **Language:** English (code EN)
- **Date published:** 2025-01-23 (in ISO format YYYY-MM-DD)
- **Geographic coverage:** Not specific, so we will enter "World" (code WLD)
- **Duration:** 3 minutes and 25 seconds. The duration must be entered in ISO8601 format: PT3M25S.
- **Video URL:** <https://youtu.be/px1EeqpKDUI?list=PLopq6yGfmFAu3tscprzTPpoPrP1q0E9XZ>
- **Embed URL:** <https://www.youtube.com/embed/px1EeqpKDUI?list=PLopq6yGfmFAu3tscprzTPpoPrP1q0E9XZ> (Note: This information can be obtained by right-clicking on the video in Youtube. It will open a menu, in which you will find an option to "Copy embed code". Select only the URL (src) part of it.)



- **Transcript:** (*This was copy/pasted from the Youtube channel, and edited for formatting*): "In 2022 as 250 million people faced food crises, the G7 presidency and the Worldbank Group came together to launch the Global Alliance for Food Security, mobilising a swift, coordinated response to the growing global hunger crisis. Working alongside international partners, we transformed food and nutrition security data systems by introducing advanced country level assessment and prediction tools. The World Food Security Outlook was developed to enhance understanding of global food and nutrition security. Using machine learning to analyse data from various sources and project food security up to six years in advance. Instead of waiting for crises to emerge, we can now utilise these insights to proactively address and mitigate potential impacts. Additionally, we've created tools to monitor prices of household items and unofficial exchange rates in real time across many locations. This aids in understanding rapid changes in food and fuel affordability, and identifying areas of urgent need. Furthermore, second layer monitoring tools like the Joint Monitoring Report build on these new data sources to provide more in-depth insights into complex conflict affected areas. These innovations aim to provide precise, timely, and actionable food and nutrition security data to guide Worldbank programming and partner actions. For instance, in the Dry Corridor, the World Food Security Outlook has brought together government, humanitarian and development leaders to assess risks, scale up early action, strengthen safety nets, and adjust agricultural planning for El Nino impacts. In the Horn of Africa, real time prices track water and staple food prices during droughts. Enabling the identification of operational and funding gaps in real time. The shift towards machine learning and real time data is already transforming food security analysis. Through the Joint Monitoring report, humanitarian and development partners now follow the same data driven approach to collectively recognise emerging crises early and enhance crisis preparedness. These first of their kind reports are speeding up response times and strengthening evidence based decision making, demonstrating how data driven tools deliver transparent, robust and high frequency analysis at a fraction of the cost of traditional food and nutrition security updates. Join the fight against hunger. Explore real time crises and the world food security outlook at microdata.worldbank.org. For more insights and our progress on Sustainable Development Goal 2, visit worldbank.org."



The screenshot shows a video player interface. On the left is a dark video frame showing a person working in a field of crops. On the right is a transcript panel with a red box highlighting the word "Transcript". The transcript contains several lines of text with timestamps and corresponding subtitles.

Transcript

Search in video

Global Alliance for Food Security

0:01 In 2022 as 250 million people faced food crises,
 0:06 the G7 presidency and the Worldbank Group came together
 0:10 to launch the Global Alliance for Food Security,
 0:14 mobilising a swift, coordinated response to the growing global hunger crisis.
 0:21 Working alongside international partners,
 0:23 we transformed food and nutrition security data systems
 0:28 by introducing advanced country level assessment
 0:31 and prediction tools.
 0:34 The World Food Security Outlook was developed

World Food Security Outlook

0:37 to enhance understanding of global food and nutrition

This information can be entered in the Metadata Editor template in the following elements:

| Information | In the metadata template |
|---------------------|--|
| Author | Video description / Authors and contributors / Creator |
| Description | Video description / Content / Description |
| Genre | Video description / Content / Genre |
| Keywords | Video description / Content / Keywords |
| Language | Video description / Content / Languages |
| (derived) | Video description / Status ("Published") |
| Date published | Video description / Dates and version / Date published |
| Geographic coverage | Video description / Geographic and time coverage / Country |
| Video URL | Video description / Access and rights / Video URL |
| Embed URL | Video description / Access and rights / Embed URL |
| Duration | Video description / Technical information / Duration |

The screenshot shows the 'Metadata Editor' interface. On the left, a sidebar lists categories: Home, Metadata information, Video description, Title statement, Authors and contributors, Dates and version, Content (which is selected), Description, Genre, Audience, Keywords, Topics, Persons, Main entity, and Transcript. The main area is titled 'Reshaping Food Security Analytics: An In-Depth Look at Emerging Solutions for ...'. It contains sections for 'Content', 'Description' (with a rich text editor containing placeholder text about AI and machine learning), 'Genre' (set to 'Documentary'), 'Audience' (empty), and 'Keywords' (a table with rows for Food Security, Famine, Nutrition data, Food crisis, Global Alliance for Food security, and Dry corridor). A 'SAVE' button is visible in the top right.

After entering all available information, click on **SAVE**. Click on *Preview* in the navigation tree to view all information you have entered so far.

Step 3: Add information on related resources

Once you have entered the metadata, you can finalize the documentation of the video by documenting and attaching external resources. External resources include all materials (files and links) that you want to make accessible to users when you publish the video in a catalog. In this example, we will add one external resource: a link to the World Bank YouTube channel.

To create an external resource, click on *External resources* in the navigation tree and then click on **Create resource**. Select the resource type ("Web Site"), give it a short **title** (*Video in the World Bank YouTube channel*), and enter the **URL** in **Resource attachment** (<https://www.youtube.com/worldbank>).

The screenshot shows the 'Edit resource' form for a video. The 'Resource type' is set to 'Web Site'. The 'Title' field contains 'World Bank YouTube channel'. The 'Author' field is empty. At the top right, there are 'SAVE' and 'CANCEL' buttons.

Then click **SAVE**. The video will now be listed as an external resource.

Step 4: Export and publish metadata

In the *Project* page, a menu of options is available to you.

The Project page provides several options for managing metadata:

- Project** section:
 - Export package (ZIP)
 - Export JSON
 - Publish to NADA
 - PDF documentation
 - Change log
- Metadata** section:
 - Apply default values from template
 - Import project metadata
 - Import external resources
- External resources** section:
 - Export RDF/XML
 - Export RDF/JSON

- **Export package (ZIP)**

This option allows you to generate a ZIP file containing all metadata and resources related to the project. This package can be shared with others, who can then import it in their own Metadata Editor.

- **Export JSON-**

Export metadata to JSON will generate a JSON file containing the metadata. The option is provided to include all elements or only the non-private ones. The JSON file will look like this:

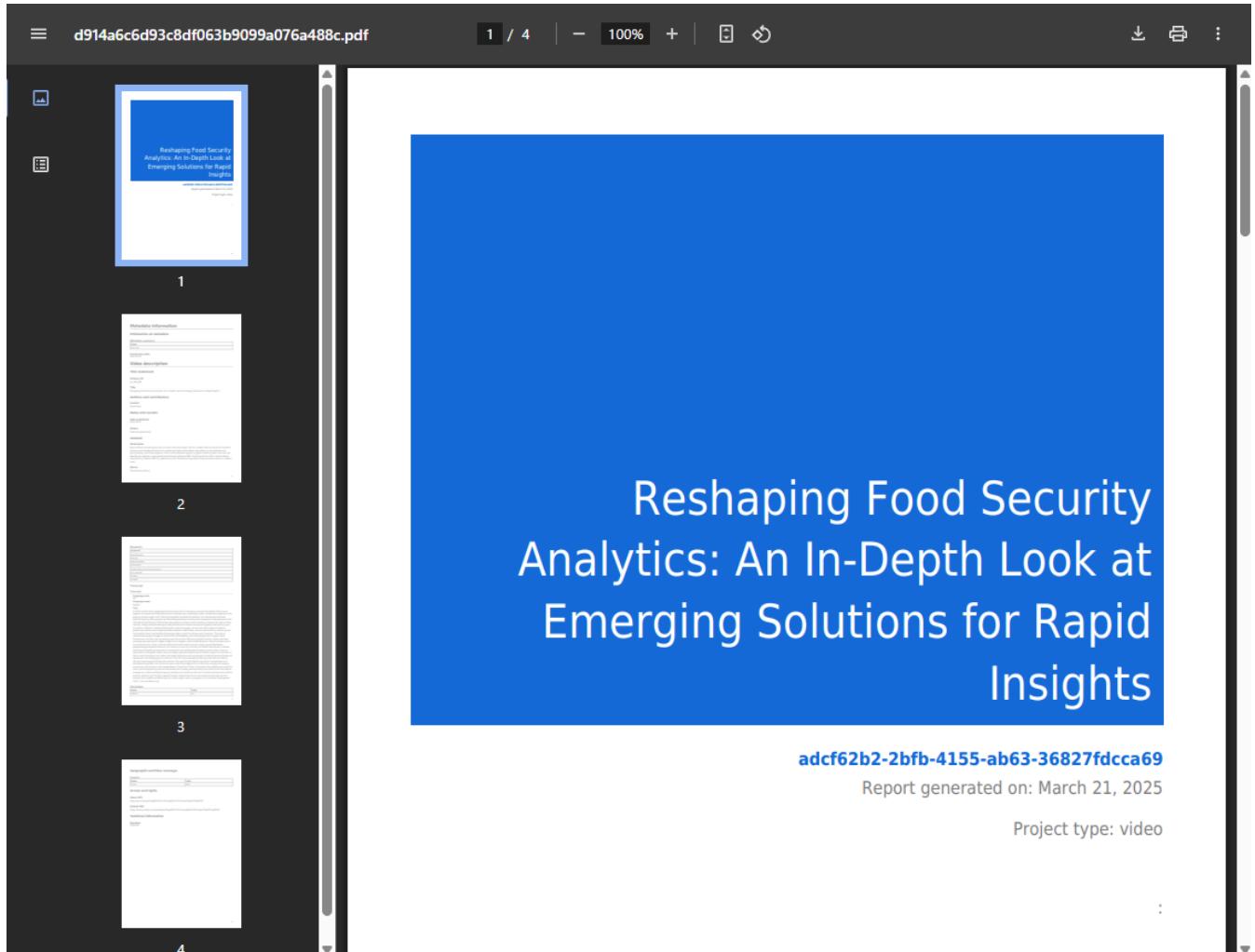
```
pretty-print 
{
  "type": "video",
  "idno": "adcf62b2-2bfb-4155-ab63-36827fdcca69",
  "changed": "1740086586",
  "changed_utc": "2025-02-20T21:23:06+00:00",
  "created": "1740071050",
  "created_utc": "2025-02-20T17:04:10+00:00",
  "created_by": "11",
  "changed_by": "11",
  "metadata_information": {
    "producers": [
      {
        "name": "John Doe"
      }
    ],
    "production_date": "2025-02-20"
  },
  "video_description": {
    "idno": "JD_VDO_001",
    "title": "Reshaping Food Security Analytics: An In-Depth Look at Emerging Solutions for Rapid Insights",
    "description": "New methods are paving the way for faster data and insight. Get an in-depth look at how AI and machine learning are conflict-affected regions to global inflation trends, discover how data-driven solutions—such as the Joint Monitoring Report (JMR), Real matters most.",
    "genre": "Documentary [docu]",
    "keywords": [
      {
        "name": "Food Security"
      },
      {
        "name": "Famine"
      },
      {
        "name": "Nutrition data"
      },
      {
        "name": "Food crisis"
      },
      {
        "name": "Global Alliance for Food security"
      },
      {
        "name": "Dry corridor"
      },
      {
        "name": "El Nino"
      },
      {
        "name": "Hunger"
      }
    ],
    "country": [
      {
        "name": "World",
        "code": "WLD"
      }
    ]
  },
  "date_published": "2025-01-23".
}
```

- **Export RDF/XML and Export RDF/XML**

These options allow you to export the metadata related to external resources in JSON or XML format.

- **PDF documentation**

A PDF version of the metadata can be automatically created. Select PDF documentation then click on [GENERATE PDF](#). When the PDF is generated, click on [DOWNLOAD PDF](#). You will obtain a bookmarked PDF file with all entered metadata.



- **Publish to NADA**

If you have a NADA catalog and the credentials to publish content in it, you can also *Publish to NADA*. Select a configured NADA catalog, select the options as shown in the screenshot below, and click **PUBLISH**.

The screenshot shows the 'Metadata Editor' application interface. On the left is a sidebar with navigation links: Home, Preview, Metadata information, Video description, Tags, External resources (selected), and File manager. The main area is titled 'Reshaping Food Security Analytics: An In-Depth Look at Emerging Solutions for Rapid Insights'. A blue banner at the top says 'Publish to NADA'. Below it, a 'Catalog' section shows 'NADA demo - https://nada-demo.ihsn.org' with an ID of 16. The 'Project options' section includes fields for Overwrite if already exists? (Yes), Publish (Publish), Data access (Direct access - [direct]), and Collection (N/A). The 'External resources' section shows one resource selected: 'World Bank YouTube channel' (https://www.youtube.com/worldbank) under the 'Title' category. Under 'Options', 'Publish project' is selected. At the bottom is a large blue 'PUBLISH' button.

The video will now be listed and made discoverable in the NADA catalog, with a link to the YouTube channel. The video can be viewed from within the NADA page.



Reshaping Food Security Analytics: An In-Depth Look at Emerging Solutions for Rapid Insights

World, 2025

Reference ID [JD_VDO_001](#)

Metadata [JSON](#)

CREATED ON
Feb 20, 2025
LAST MODIFIED
Feb 20, 2025
PAGE VIEWS
104

[VIDEO DESCRIPTION](#)

[DOWNLOADS](#)



TITLE

Reshaping Food Security Analytics: An In-Depth Look at Emerging Solutions for Rapid Insights

DESCRIPTION

New methods are paving the way for faster data and insight. Get an in-depth look at how AI and machine learning are reshaping food security analytics to deliver actionable information for humanitarian aid, policy-making, and crisis response. From conflict-affected regions to global inflation trends, discover how data-driven solutions—such as the Joint Monitoring Report (JMR), Real-Time Prices (RTP), and the World Food

Purpose of templates

Simplify and tailor the use of metadata standards

Metadata standards include more metadata elements than needed by any user for documenting a specific dataset. The reason is that standards are developed to address many use cases and meet the needs of diverse organizations. To simplify their usage, **metadata templates** are created and used. Metadata templates are **tailored subsets** of the metadata standard's elements, designed for particular purposes or user groups while maintaining compliance with the standard. Templates will automatically generate the metadata entry pages that the data curators will see when they document a dataset.

The screenshot shows the Metadata Editor interface with a sidebar on the left containing navigation links like Home, Metadata information, and Indicator description. Under Indicator description, 'Title statement' is selected, which is expanded to show 'Primary ID', 'Other identifiers', 'Name', 'Aliases', 'Database ID', 'Sources, concepts, and r...', 'Standards and framework...', and 'Quality'. A yellow callout points to the 'Title statement' header with the text: 'Define what metadata elements should be required, recommended, or optional'.

The main content area shows a form for a 'Title statement'. It includes fields for 'Primary ID' (set to 'JD_IND_001') and 'Other identifiers'. A yellow callout points to the 'Other identifiers' field with the text: 'Customize the description/instructions for each metadata element, which will be displayed in metadata entry screens.' Below these fields is a detailed description of the 'Other identifiers' element.

Further down, there is a section titled 'Select the list of metadata elements to be used, customize their label, group them in a meaningful way.' A yellow callout points to this section with the text: 'Select the list of metadata elements to be used, customize their label, group them in a meaningful way.'

On the right side of the screen, there is a table with columns 'Database', 'URL', and 'Notes'. One row in the table is labeled 'World Development Indica'. A yellow callout points to this table with the text: 'Also:' followed by '- Provide code lists that will be used as drop-downs for selected elements.' and '- Provide default values for some elements.'

Metadata templates allow for tailoring metadata elements in several ways:

- **Label:** The label of a metadata element provided in the metadata standard can be replaced with a label that conforms to an organization's lexicon or stipulations.
- **Status:** Some metadata elements may be declared as *required* in the metadata standard itself. This status cannot be changed. But metadata elements that are not required in a standard can be designated as *required* or *recommended* in a template. Declaring a metadata element as required will not prevent the metadata from being *validated*. It thus serves as a useful quality control. It is however advised against making too many elements required, and to set important but not crucial elements as recommended rather than required.
- **Description/instructions:** A tailored description and instructions can be provided in a template for each metadata element. These descriptions serve as guidelines for data curators.
- **Controlled vocabularies:** A controlled vocabulary can be specified for a metadata element, when applicable.
- **Default values:** A default value can be set for a metadata element. Default values in the Metadata editor will not be automatically applied; an option is provided to the data curator to apply default values that have been entered in the

template.

- **Validation rules:** Customized validation rules can be set for each metadata element, using regular expressions, by setting a valid range for values, or other, to guarantee metadata coherence and uniformity.
- **Adding elements:** Elements that are not part of a metadata standard can be added in a template, as *additional* elements.

These various customization options help tailor metadata standards to specific use cases in an organization. They also enable the creation of templates in different languages.

The Metadata Editor provides a *Template Manager* to generate and edit metadata templates.

Templates offer flexibility and convenience, but coherence across an organization should be maintained. Ideally, each data type would use a unique template. However, multiple templates per data type are allowable for specialized cases, though this number should be minimized.

Embed controlled vocabularies in the metadata standards

A **controlled vocabulary**, also known as a **code list**, is a predefined and structured set of terms (with corresponding codes) that are consistently used to populate specific metadata elements. Controlled vocabularies are applied to a limited number of elements within a metadata standard to ensure uniformity and precision.

Utilizing controlled vocabularies helps ensure that the same concept is consistently represented by the same term across various records, thereby reducing ambiguity and enhancing searchability and interoperability. Examples of controlled vocabularies include widely recognized national and international classifications, such as ISO country codes and names or the International Standard Industrial Classification (ISIC). Additionally, they may comprise organization-specific vocabularies, such as keyword taxonomies or tailored code lists designed for specific domains or purposes.

Controlled vocabularies play a critical role in data discovery. By applying standardized terms to metadata elements, data catalogs can offer these terms as filtering options (facets), allowing users to narrow their search based on specific attributes or categories. This approach significantly enhances the discoverability of datasets, facilitating a more efficient search experience.

Define schemas for administrative metadata

Administrative metadata is the metadata needed for the administration of specific data management and dissemination systems. As the requirements of such systems vary widely, no pre-defined standard or schema is provided. Templates are used in the Metadata Editor to define the content of administrative metadata schemas, fully tailored to the needs of the systems in place in the organization that uses the Metadata Editor.

Instructions related to administrative metadata and their templates are provided in the chapter **Documenting data: administrative metadata**.

Creating and editing templates

This section is about designing templates for the various data types. For the design of administrative metadata templates, see chapter **Administrative metadata**.

Data curators will rarely make use of all elements available in a metadata standard. Many of the available elements may not be relevant in the context of a specific organization. Metadata templates provide a solution to define and tailor subsets of metadata elements available in a standard.

Metadata templates are created and/or edited using the *Template Manager* tool in the Metadata Editor.

Not all users of the Metadata Editor will have the credentials to create or edit templates; this is a specific role that must be assigned to selected users in an organization. Templates should be created and maintained by one or multiple managers(s) of the data curation process. The metadata templates they create can then be made available to all data curators.

More than one template can be developed for each metadata standard (i.e., for each main data type). It is highly recommended to keep the number of metadata templates small. This will foster consistency in the metadata being produced, facilitate the work of data curators, and reduce the burden of maintaining a collection of templates.

Pre-designed templates and template list

You access the **Template Manager** by clicking on **TEMPLATES** in the main menu of the Metadata editor. This will open a page showing all available templates by type of data. The list can be filtered by selecting a data *Type* in the left frame.

| Type | Default | Title | Language | Version | Owner | Last updated by | Updated on |
|--------|----------------------------------|------------------------------|----------|----------|----------|-----------------|------------|
| core | <input type="radio"/> | Microdata DDI 2.5 EN | en | | | | |
| core | <input type="radio"/> | IHSN DDI 2.5 Modèle v01 FR | fr | | | | |
| core | <input checked="" type="radio"/> | IHSN DDI 2.5 Template v01 EN | en | | | | |
| custom | <input type="radio"/> | DDI 2.5 Administrative | en | Cathrine | Cathrine | 02/03/2025 | : |

For each main data type (i.e., for each metadata standard), the Metadata Editor provides one or multiple **core** templates which are non-editable. One of the core templates contains **all elements** of the corresponding metadata standard, with their default parameters (label, description, etc.). Other core templates are provided as suggested templates, which only contain what is considered as the most important metadata elements for a general use case. Typically, new templates will be created by generating a copy of a core template, then editing it.

Actions on templates

A set of options and actions is available for each template in the *Template list* page (accessed by clicking on the three-dots icon):

Microdata

| Type | Default | Title | Language | Version | Owner | Last updated by | Updated on | |
|------|-----------------------|----------------------|----------|---------|-------|-----------------|------------|--|
| core | <input type="radio"/> | Microdata DDI 2.5 EN | en | | | | | |

- DUPLICATE
- EXPORT
- PREVIEW
- TABLE
- PDF
- REVISIONS
- UUID

- **DUPLICATE**

Generate an editable copy of the selected template. After duplicating a template, click on its title in Templates list to open the new template page for editing.

Metadata Editor

About | English John Doe

Template manager

IMPORT TEMPLATE

Types

- All
- Microdata**
- Timeseries
- Timeseries DB
- Script
- Geospatial
- Document

TEMPLATES

| Type | Default | Title | Language | Version | Owner | Last updated by | Updated on | |
|--------|----------------------------------|------------------------------|----------|----------|----------|-----------------|------------|--|
| core | <input type="radio"/> | Microdata DDI 2.5 EN | en | | | | | |
| core | <input type="radio"/> | IHSN DDI 2.5 Modèle v01 FR | fr | | | | | |
| core | <input checked="" type="radio"/> | IHSN DDI 2.5 Template v01 EN | en | | | | | |
| custom | <input type="radio"/> | Microdata DDI 2.5 EN - copy | en | John Doe | John Doe | 02/13/2025 | | |

The duplicated template can be customized, and saved under a new name (edit the **Name** field in the *Template description* page, and click on **SAVE**).

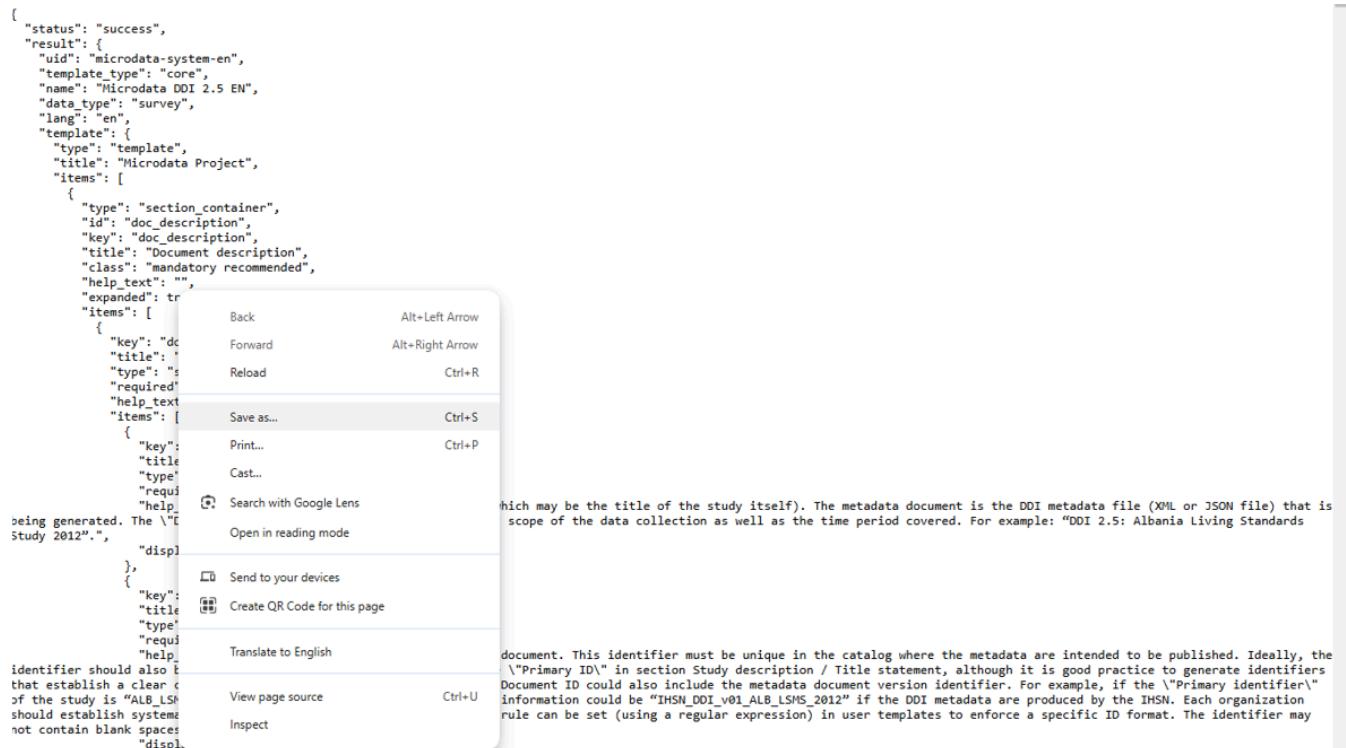
The screenshot shows the 'Template manager' interface. On the left, a sidebar lists categories: Description, Document description, Study description, Data files, Variables, Variable groups, Tags, and Additional Fields. The 'Description' category is selected and highlighted in orange. On the right, the 'Microdata DDI 2.5 EN - copy' template is displayed with the following fields:

- Type:** survey
- Language:** en
- Name:** Microdata DDI 2.5 EN - copy
- Version:** (empty)
- Organisation:** (empty)
- Author:** (empty)

At the top right, there are 'Save *' and 'Close' buttons.

- **EXPORT** (and **IMPORT**)

Create a JSON copy of a template, which can then be saved as a file with [.json] extension (to save the JSON file, position the cursor in the text, right-click, and select **Save as**). Exporting templates allows sharing them with other organizations, who can **IMPORT** templates in their own instance of the Metadata Editor.



Exported templates can be imported in the Metadata Editor by clicking on **IMPORT TEMPLATE**.

The screenshot shows the 'Template manager' section of the Metadata Editor. On the left, there's a sidebar with 'Types' filters: All, Microdata, Timeseries, and Timeseries DR. Below these are two buttons: 'DELETE' and 'IMPORT TEMPLATE'. The 'IMPORT TEMPLATE' button is highlighted with a red box. The main area displays a table for the 'Microdata' template, with columns for Type, Default, Title, Language, Version, Owner, Last updated by, and Updated on. One row is shown: core, false, Microdata DDI 2.5 EN, en.

- DELETE**

Delete the selected template (only available for custom templates, not for core templates).

- PREVIEW**

Generate an HTML version of the template, which will open in a web browser.

The screenshot shows the generated HTML preview for the 'Microdata DDI 2.5 EN' template. On the left, a tree view shows the structure of the template, including sections like 'Information' and 'study_desc'. On the right, the preview itself is displayed. It starts with the title 'Microdata DDI 2.5 EN'. Below it, under 'Information on metadata', are fields for 'ID' (microdata-system-en), 'Language' (en), and 'Data type' (microdata). The main content area is titled 'Document description' and contains a section for 'Information on metadata'. This section includes a 'Document title' field, which is described as 'The title of the metadata document (which may be the title of the study itself). The metadata document is the DDI metadata file (XML or JSON file) that is being generated. The "Document title" should mention the geographic scope of the data collection as well as the time period covered. For example: "DDI 2.5: Albania Living Standards Study 2012."'. There are also sections for 'Document ID' and 'Version'.

- TABLE**

Generate a tabular description of the template, which can for example be copy-pasted in MS-Excel if needed.

| Parent | Field | Type | Title | Description |
|--|---|--------|----------------------------|---|
| | doc_desc | object | doc_desc | NOT AVAILABLE |
| doc_desc | doc_desc/title | string | Document title | The title of the metadata document (which may be the title of the study itself). The metadata document is the DDI metadata file (XML or JSON file) that is being generated. The "Document title" should mention the geographic scope of the data collection as well as the time period covered. For example: "DDI 2.5: Albania Living Standards Study 2012". |
| doc_desc | doc_desc/idno | string | Document ID | A unique identifier for the metadata document. This identifier must be unique in the catalog where the metadata are intended to be published. Ideally, the identifier should also be unique globally. This is different from the "Primary ID" in section Document ID. The Document ID could also include the metadata document version identifier. For example, if the "Primary identifier" of the study is "ALB LSMS_2012", the "Document ID" in the Metadata information could be "IHSN_DDI_v01_ALB_LSMS_2012" if the DDI metadata are produced by the IHSN. Each organization should establish systematic rules to generate such IDs. A validation rule can be set (using a regular expression) in user templates to enforce a specific ID format. The identifier may not contain blank spaces. |
| doc_desc | doc_desc/producers | array | Metadata producers | The metadata producer is the person or organization with the financial and/or administrative responsibility for the processes whereby the metadata document was created. This identifier should also be unique in the catalog where the metadata are intended to be published. Ideally, the identifier should also be unique globally. This is different from the "Primary ID" in section Document ID. The Document ID could also include the metadata document version identifier. For example, if the "Primary identifier" of the study is "ALB LSMS_2012", the "Document ID" in the Metadata information could be "IHSN_DDI_v01_ALB_LSMS_2012" if the DDI metadata are produced by the IHSN. Each organization should establish systematic rules to generate such IDs. A validation rule can be set (using a regular expression) in user templates to enforce a specific ID format. The identifier may not contain blank spaces. |
| doc_desc/producers | doc_desc/producers/name | string | Name | The name of the person or organization in charge of the production of the DDI metadata. If the name of individuals cannot be provided due to an organization's data protection rules, the title of the person, or an anonymized identifier, can be provided (or this field can be left blank if no other option is available). |
| doc_desc/producers | doc_desc/producers/abbr | string | Abbreviation | The initials of the person, or the abbreviation of the organization's name mentioned in "Name". |
| doc_desc/producers | doc_desc/producers/affiliation | string | Affiliation | The affiliation of the person or organization mentioned in "Name". |
| doc_desc/producers | doc_desc/producers/role | string | Role | The specific role of the person or organization mentioned in "Name" in the production of the DDI metadata. |
| doc_desc | doc_desc/prod_date | string | Production date | The date the DDI metadata document was produced (not the date it was distributed or archived), preferably entered in ISO 8601 format (YYYY-MM-DD or YYYY MM). A validation rule can be set in user templates to enforce a date format. This is a "Recommended" element, as information on the producer and on the date of metadata production is useful for catalog administration purposes. |
| | doc_desc/version_statement | object | doc_desc/version_statement | NOT AVAILABLE |
| doc_desc/version_statement | doc_desc/version_statement/version | string | Version | Documenting a dataset is not a trivial exercise. It may happen that, having identified errors or gaps in a DDI document, or after receiving suggestions for improvement, one adds new fields to the DDI document. The "Version" element and the elements "Version date", "Version responsibility", and "Version notes" describe the version of the metadata document. The "Version" element is used to enter the label of the version of the metadata document, also known as release or edition. For example, "Version 1.1" or "v1.1". |
| doc_desc/version_statement | doc_desc/version_statement/version_date | string | Version date | The date when this version of the metadata document (DDI file) was produced, preferably identifying an exact date. This will usually correspond to the "Production date" element. It is recommended to enter the date in the ISO 8601 date format (YYYY-MM-DD or YYYY-MM or YYYY). A validation rule can be entered in customized templates to ensure that dates are entered in the appropriate format. |
| doc_desc/version_statement | doc_desc/version_statement/version_resp | string | Version responsibility | The organization or person responsible for this version of the metadata document. |
| doc_desc/version_statement | doc_desc/version_statement/version_notes | string | Version notes | This element can be used to clarify information/annotation regarding this version of the metadata document, for example to indicate what is new or specific in this version comparing with a previous version. |
| | study_desc/title_statement | object | study_desc/title_statement | NOT AVAILABLE |
| study_desc/title_statement | study_desc/title_statement/idno | string | Primary ID | The "Primary ID" (also referred to as IDNO) is a unique identification number used to identify the study (survey, census or other). A unique identifier is required for cataloging purpose, so this element is declared as "Required". The identifier will allow users to cite the dataset properly. The identifier must be unique within the catalog. Ideally, it should also be globally unique; the recommended option is to obtain a Digital Object Identifier (DOI) for the study. Alternatively, the "Primary ID" can be constructed by an organization using a common schema and scheme. For example for "catalog country+study year+version", where country is the 3 letter ISO country code, prefix is the abbreviation of the program, year is the year the survey was conducted, and version is the version year (the year the study started); version is a version number. Using the schema, the Uganda 2005 Demographic and Health Survey for example would have the following ID (where "MDA" stand for "My Data Archive"): MDA_UGA DHS 2005 v01. Note that the schema allows you to provide more than one identifier for a same study (in element "Other identifiers"); a catalog specific identifier is thus not incompatible with a globally unique identifier like a DOI. A validation rule can be set (using a regular expression) in user templates to enforce a specific ID format. The identifier may not contain blank spaces. |
| study_desc/title_statement | study_desc/title_statement/identifiers | array | Other identifiers | This repeatable element is used to enter identifiers (IDs) other than the "Primary ID" (IDNO). It can for example be a Digital Object identifier (DOI). The "Primary ID" can be repeated here (the "Primary ID" does not provide a "Type" parameter, so if a DOI or other standard ID type is used as main identifier, it is recommended to repeat it here with the identification of the type). |
| study_desc/title_statement/identifiers | study_desc/title_statement/identifiers/type | string | Type | The type of identifier. For example: "DOI". |
| study_desc/title_statement/identifiers | study_desc/title_statement/identifiers/identifier | string | Identifier | The identifier itself. |

PDF

Generate a PDF version of the template. If the template contains detailed descriptions of the metadata elements, and good examples of content, the PDF file can serve as a useful instruction guide for data curators.

Document description

section_container doc_description

Information on metadata

section doc_desc.title_statement

Document title

string doc_desc.title

The title of the metadata document (which may be the title of the study itself). The metadata document is the DDI metadata file (XML or JSON file) that is being generated. The "Document title" should mention the geographic scope of the data collection as well as the time period covered. For example: "DDI 2.5: Albania Living Standards Study 2012".

Document ID

string doc_desc.idno

A unique identifier for the metadata document. This identifier must be unique in the catalog where the metadata are intended to be published. Ideally, the identifier should also be unique globally. This is different from the "Primary ID" in section Study description / Title statement, although it is good practice to generate identifiers that establish a clear connection between these two identifiers. The Document ID could also include the metadata document version identifier. For example, if the "Primary identifier" of the study is "ALB LSMS_2012", the "Document ID" in the Metadata information could be "IHSN_DDI_v01_ALB_LSMS_2012" if the DDI metadata are produced by the IHSN. Each organization should establish systematic rules to generate such IDs. A validation rule can be set (using a regular expression) in user templates to enforce a specific ID format. The identifier may not contain blank spaces.

REVISIONS

Provide a history of changes to the template since its creation.

- **UUID**

Allows template administrators to edit the unique identifier of the template. By default, a system identifier is created. This identifier can be changed to a more readable one. This will typically be done for administrative metadata templates.

Editing a template

Description page

You access the *Description* page of a template (other than custom templates) by clicking on the template *Title* in the list of templates. The *Description* page is where you provide the main identification information of the template. It includes the following elements:

- **Type:** The type of data to which the template applies (microdata, indicator, database, geographic dataset, document, etc.)
- **Language:** The language of the template
- **Name:** The name of the template.
- **Version:** The version of the template.
- **Organization:** The organization that developed the template or for whom the template was developed.
- **Author:** The author(s) of the template.
- **Description:** A brief description of the template.
- **Instructions:** A set of overall instructions related to the template (not the instructions related to each specific metadata element, which will be added in the template itself). The content of this element can be plain text or formatted text (using markdown syntax; see <https://www.markdownguide.org/basic-syntax/> for a guide on formatting text using markdown).

T Template manager
IHSN DDI 2.5 Template v01 EN
Save **Close**

Type: survey

Language: en

Name: IHSN DDI 2.5 Template v01 EN

Version: 1.0

Organisation: IHSN / World Bank Development Data Group

Author: MA TB OD MW

Description: A DDI 2.5 (DDI Codebook) template recommended by the International Household Survey Network (IHSN) for the documentation of survey and census datasets.

Instructions: Markdown

Navigation tree

The Template Manager navigation tree shows the structure and content of the template. This structure and content will define the structure and content of the metadata entry pages that data curators will see when they document datasets using the template.

T Template manager

- Description**
- > **Metadata information**
- > **Indicator description**
 - > **Title statement**
 - Primary ID**
 - Other identifiers**
 - Name**
 - Aliases**
 - Database ID**
 - > **Sources, concepts, and methods**
 - > **Standards and frameworks**
 - > **Quality**
 - > **Geographic and time coverage**
 - > **Description**
 - > **Access and use**
 - > **Contacts**
 - > **Tags**



Metadata Editor

Poverty headcount ratio at \$2.15 a day (2017 PPP) (%)

Title statement

Primary ID JD_IND_001

Other identifiers

| Type | Identifier |
|------|------------|
| WDI | SI.POVDAY |

Name Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)

Aliases

| Alias |
|-------|
| WDI |

Database ID WDI

Page 7 of 14 | Chapter: Creating and editing templates | Metadata Editor

The navigation tree in the Template Manager indicates the type of element using the following icons:

 **Group of elements**, hard-coded in the metadata standard (cannot be deleted)

 **User-defined group** (or subgroup) of metadata elements

 **Metadata element – Simple text**

Example: [Name](#) 

Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)

 **Metadata element – Array**

Example: [Other identifiers](#) 

| Type | Identifier | Database | URL | Notes |
|------|-------------|------------------|-----|-------|
| WDI | SI.POV.DDAY | World Developmen | | |

[+ ADD ROW](#)

 **Metadata element – Nested array**

Example: [Contact](#) 

1 - Contact [Clear](#) 

| |
|---|
| Individual name  |
| Organisation name  |

The navigation frame contains a toolbar used to edit the structure and content of the template. The available tools are the following:

-  Add the element selected in the pool of “available elements” to the group
 -  Remove the selected element from the group and put it back in the pool of “available elements”
 -  Add a group or subgroup of elements. An untitled empty group is created.
 -  Delete a group or subgroup of elements. The group and its content will be deleted.
 -  Move a group or element up in the navigation tree.
 -  Move a group or element down in the navigation tree.
-
-  Copy the selected element.
 -  Paste the selected element.
-
-  Add a custom text field (not selected from the list of available elements).
 -  Add a custom array field (not selected from the list of available elements).
 -  Add a custom nested array field (not selected from the list of available elements).

Customizing a template

Groupings

Metadata standards organize metadata elements into some main *container*. These groupings are not editable; they are "hardcoded" in the respective metadata standards. Within containers, elements can be organized by *group* and *sub-group* ("sections"). Groups and sub-groups are user-defined. The purpose of the groupings is to organize the metadata elements in a way that will be user-friendly and intuitive to data curators. All metadata elements in a template must be placed in a user-defined group (they cannot be placed directly under a container).

The screenshot shows the 'Template manager' interface. On the left, a navigation tree lists various metadata elements like 'Description', 'Document description', 'Study description', and 'My elements'. 'My elements' is selected and highlighted in orange. On the right, the 'DDI 2.5 Administrative Records v01 EN - copy' template is being edited. The template details include:

- Key:** study_desc1739478912452
- Label:** My elements
- Type:** section
- Description:** This is a new section containing demo metadata elements
- Original description:** N/A
- Available items:**
 - Show all elements (button)
 - Production
 - Other producers (selected, highlighted in orange)
 - Funding agencies
 - Other contributors
 - Production date
 - Production place
 - Sampling
 - Sampling procedure
 - Deviations from sample design

A detailed description table for 'Other producers' is shown on the right:

| Description | | | |
|-------------|---|--|--|
| Field | study_desc.production_statement.producers | | |
| Type | array | | |
| Title | Other producers | | |
| Description | This field is used to list other parties and persons that have played a significant but not the leading technical role in implementing and producing the data (which will be listed in "Authoring entity"). Do not include here the financial sponsors, who should be listed in "Funding agencies". | | |

Below this is a table for 'Array properties':

| Key | Title | Type | Description |
|------|--------------|--------|---|
| name | Name | string | The name of the person or organization. |
| abbr | Abbreviation | string | The official abbreviation of the |

Groups and subgroups are created by clicking on the [+] button in the navigation tree (the button is only active when a container or group is activated in the navigation tree). When a new group is created, replace the *Untitled* label with a (short) label of your choice, and provide a brief description of it (optional). You can move the group up or down the navigation tree by using the *Up* and *Down* arrows.

You can delete a group by clicking on the [-] button. The group and all its elements will then be removed from the navigation tree. All elements that were included in the tree will also be removed. The elements that belong to the standard will be put back to the list of available elements (see below), but their customization will be lost. The *additional* elements that may have been included in the group (elements that do not belong to the standard) will be lost.

Adding metadata elements from the standard

Templates are intended to be a customized organization of metadata elements from a metadata standard (with possible addition of *additional* elements not found in the standard - see next section). A template is therefore created mainly by selecting elements from the list of elements available in the metadata standard, and placing them in the structure shown in the navigation bar. The list of available elements (which contains all elements from the metadata standard that have not yet been selected, i.e. not found in the navigation bar) is shown in the right frame of the Template Manager, **when a group is selected in the navigation bar**. An option is provided to **show all elements**. The list of available elements is the pool of metadata elements from the metadata standard that can be added to the template.

The screenshot shows the 'Template manager' and 'Test' panels of the Metadata Editor. The 'Template manager' panel on the left lists various metadata groups and elements. The 'Test' panel on the right shows a selected element with its properties: Key (study_desc.title_statement), Label (Identification), Type (section), and Description (empty). Below these, the 'Available items' section is highlighted with a red border, showing categories like Production, Data collection and processing, and Data processing, each with sub-elements.

Metadata elements must be placed within groups or sub-groups (not directly under a container).

- **To add an element from the standard:** In the navigation tree, select the group in which you want to add the element. Then select the element from the list of available elements by clicking on the + button next to the element. The element will now be listed in the group. You can move the element up or down the list within the group. You can remove the element from the navigation tree by selecting it and clicking on the [>] button in the toolbar. The element is sent back to the list of available elements, with its default description (i.e. customizations will be lost). You can also copy/paste elements to move them from one group to another (within the same container). To do this, select the element(s) to be copied, and click on the **Copy** button in the toolbar (the elements included in the clipboard will be marked in the navigation tree). Select the group where the elements have to be pasted, and click on the **Paste** button.
- **To edit metadata elements:** When you select a metadata element in the navigation tree, all information about the element is displayed in the right frame. Some of this information can be edited. The information includes the following:
 - **Key:** The key is the unique identifier of the element in the metadata standard. This information is not editable, except for additional elements created by the user.
 - **Label:** The label of the element can be edited. It should be short and informative.
 - **Type:** The type of element. A metadata element can be a text field, an array, a nested array, or a simple array.
 - **Status:** Each element can be categorized as:
 - **Required:** Required means that metadata for any dataset must contain information for this element. Metadata that fail to include content for a required element will not be validated (validation errors will be displayed).

- **Recommended:** This status is mainly used to facilitate metadata entry by data curators and for quality assurance.
- **Private:** Some metadata may be useful to the organization who generates the metadata, but not be part of the metadata to be published. Metadata elements marked as private may be excluded from the metadata files exported from the Metadata Editor.
- **Read-only.**
- **Description:** The description of the metadata element should provide a clear indication of what data curators are expected to enter in the field. The instructions will be displayed as "help" in the metadata entry pages. By default, the instructions are those that are provided in the metadata standard description. They can be customized.
- **Field properties:** This information only applies to elements of type "array" and "nested_array". Arrays contain multiple elements. The *Field properties* is where the content of the array is selected and edited.
- **DISPLAY:** This tab contains information that only applies to elements of type "text". The following information can be edited:
 - **Data type:** This indicates the type of content expected in the element: string (text), number, integer, or boolean.
 - **Display:** This indicates the *Data Type*, and *Display options*: how the field will appear in the metadata entry pages, with the following options: "text" (one-line text field), "text area" (multi-line text field), "date" (date in ISO format; the metadata entry page will show a calendar from which the data curator can select a date); "dropdown" (one-line text field with a drop down from which the data curator must select an entry, with no option to enter free text); and "dropdown-custom" (one-line text field with a drop down list, but allowing data curators to enter content other than what the dropdown suggests). The content of the dropdown lists is defined by the *controlled vocabulary* for the element (see below). For selected text fields, the *DISPLAY* tab will also contain information on *Input format* which indicates whether formatted text can be entered for the element. By default, only non-formatted text is allowed. But exceptions can be made to allow Markdown, LaTex, or HTML content to be entered by the data curator. LaTex allows capturing formulas. See section *Documenting data - General instructions* for an example.

| DISPLAY | CONTROLLED VOCABULARY | DEFAULT | VALIDATION RULES | JSON |
|--|-----------------------|---------|------------------|------|
| Data type: string | | | | |
| Display: textarea | | | | |
| Input format: Text, Markdown, LaTex, HTML. Default is Text | | | | |

- **CONTROLLED VOCABULARY:** A controlled vocabulary (or code list) A controlled vocabulary, or *code list*, is a predefined and structured set of terms (with corresponding codes) that are consistently used to populate specific metadata elements. Controlled vocabularies may be applied to some metadata elements to ensure uniformity and precision. Utilizing controlled vocabularies helps ensure that the same concept is consistently represented by the same term across various records, thereby reducing ambiguity and enhancing searchability and interoperability. A controlled vocabulary will typically provide, for each term, a code and a label. The Template Manager allows defining which column should be used as value: the *Code*, the *Label*, or both (*Label with code*). When a controlled vocabulary is entered, a green dot appears next to the title of the tab.

DISPLAY • **CONTROLLED VOCABULARY** **DEFAULT** **VALIDATION RULES** **JSON**

Controlled vocabulary:

Select column to use as value:
Label with code

| Code | Label |
|------|--------------|
| 1 | Urban |
| 2 | Rural |
| 9 | Not declared |

- **DEFAULT:** A default value can be provided for an element. This will rarely be used. Default values will not be automatically entered in the metadata; instead, the data curator will have the option to "Add default values" when documenting a dataset.

DISPLAY • **CONTROLLED VOCABULARY** **DEFAULT** **VALIDATION RULES** **JSON**

Default:

- **VALIDATION RULES:** Validation rules can be set for a metadata element, to control quality. The content entered for the element by the data curator will be validated against this set of rules, and Validation errors will be shown in the project home page. Validation rules can be of different types: regex (regular expression), min or max (minimum or maximum value, for numeric files), max_length (maximum number of characters in the entry), alpha (only letters accepted), alpha_num (only alphanumeric characters allowed), numeric (numeric value must be entered), is_uri (entry must be a URI). When one or multiple validation rules are entered, a green dot appears next to the title of the tab.

DISPLAY • **CONTROLLED VOCABULARY** **DEFAULT** • **VALIDATION RULES** **JSON**

Validation rules:

Select rule **Add**

| Rule | Value |
|-------|---|
| regex | <input type="text"/> Regular expression - |

- **JSON:** The JSON version of the element description. This is not an editable content.

DISPLAY **CONTROLLED VOCABULARY** **DEFAULT** **VALIDATION RULES** **JSON**

JSON:

```
{
  "key": "study_desc.title_statement.title",
  "title": "Title",
  "type": "string",
  "required": true,
  "help_text": "This element is \"Required\". Provide here the full authoritative title for the study. Make sure to use a unique name for each distinct study.",
  "rules": [],
  "is_required": true,
  "display_type": "text"
}
```

Creating additional fields

Metadata elements that are not provided by a metadata standard can be added as "additional fields". Such metadata elements are created and managed the same way as other metadata elements, except that a unique Key has to be provided, which will be the identifier of the newly created element.

Setting a template as default

The Template Manager allows the administrator of the system to select, for each data type, the template to be used by default (one default template per type, indicated by the radio button).

Administrative metadata

Purpose

Administrative metadata refers to the metadata required for managing and operating data management and dissemination systems. Administrative metadata will only be used by large data cataloguing systems, and by data systems that require automation of processes. Not all users of the Metadata Editor will need administrative metadata.

Unlike metadata intended for data users, administrative metadata is primarily used internally and is not shared externally. It contains essential instructions that guide software applications, such as data catalogs, in handling data storage, display, and accessibility parameters.

For example:

- An organization may maintain two versions of a data catalog: one for internal use, the other one accessible to external users. While the descriptive and structural metadata to be displayed is the same for both versions of the catalog, the data access conditions may differ (for example, data may be openly accessible to internal users, but disseminated externally under different conditions). Administrative metadata may in that case be used to store information on the access policy to be applied to each version of the data catalog.
- An international organization that publishes time series of indicators in an on-line platform wants to include some visualizations for each indicator. But not all types of data visualization apply to all indicators (for example, it would not make sense to show a choropleth world map of GDP per capita in local currency, as the estimates are not comparable across countries). In that case, administrative metadata can be used to indicate, for each indicator, what visualizations should be displayed in the platform.

The content of administrative metadata will be specific to each organization and IT system. For that reason, no metadata standard is provided for administrative metadata. Instead, administrative metadata templates (or *schemas*) are entirely created "from scratch" using the Template Manager tool in the Metadata Editor.

The creation of administrative metadata follows a structured approach similar to that of descriptive metadata. A set of metadata elements forms a structured template. These elements are defined by IT specialists to ensure alignment with the organization's system functionalities and operational needs.

Metadata schemas created by an organization can then be used by the organization to capture administrative metadata for any project. Each project can be assigned one or more administrative metadata templates.

Administrative metadata will be stored within the Metadata Editor and can be exported (as JSON files). It will be accessible via API to allow data management and dissemination systems to retrieve and utilize it as needed. Administrative metadata is not included in metadata exported for public use.

Creating administrative metadata templates

Although administrative metadata schemas are specific to each organization and IT system, the Metadata Editor provides a starter template named **Core administrative metadata**. This core template is not editable. But it can be duplicated, and

the copy can then be edited for creating custom templates.

The screenshot shows the 'Template manager' interface in the Metadata Editor. On the left, a sidebar lists various 'Types' such as All, Microdata, Timeseries, etc., with 'Administrative Metadata' highlighted by a red box. The main area displays a table for the 'Administrative metadata' section. The columns are Type, Default, Title, Language, Version, Owner, Last updated by, and Updated on. Three rows are shown: a core template titled 'Core administrative metadata template' in English, a custom template titled 'OD test chart type' in English, and another custom template titled 'NADA' in English, version 1.0.0, owned by 'admin admin'.

| Type | Default | Title | Language | Version | Owner | Last updated by | Updated on |
|--------|-----------------------|---------------------------------------|----------|---------|-------------|-----------------|------------|
| core | <input type="radio"/> | Core administrative metadata template | EN | | | | |
| custom | <input type="radio"/> | OD test chart type | en | | John Doe | John Doe | 02/13/2025 |
| custom | <input type="radio"/> | NADA | EN | 1.0.0 | admin admin | admin admin | 02/13/2025 |

To create a new administrative template, select **DUPLICATE** in the list of options available for the template (triple-dot icon next to the template title). In the *Description* page, provide a name (at least), and other information describing the new template being created. Then **SAVE** the template.

You may now start customizing the template by adding your own metadata elements. Start by removing all fields under the *Metadata* section in the navigation tree (but do not delete the section/folder). To remove a field, select the field in the navigation tree and click on the right blue arrow **>**. After removing all fields, you will obtain an empty template, ready for customization.

Do NOT remove the *Metadata* section. You cannot add elements directly under the container, so this folder is necessary.

Template manager

- Description
- Administrative metadata
 - Metadata
 - Options
 - Additional Fields

Core administrative metadata template - copy

Key:
options
options

Label:
Options
Original label: Options Name: options Type: array

Type:
array

Required Recommended Private Read-only

Description:
Options

Start adding the metadata elements you need. Select the *Metadata* folder in the navigation tree, then add an element by clicking on one of the three possible types of elements. The template supports 3 types of elements:

-  **Text field**
-  **Array/tables**
-  **Nested Array**

Once created, add the following information on the metadata element:

- **Key:** The Key is the name under which the element will be stored in the JSON template file (the unique identifier of the metadata element). The key can only contain alphanumeric values, and must be unique to each metadata element within the template.
- **Label:** Give a label to the new element (replace the *untitled* label).
- For all other components, see section [Designing templates](#).

The screenshot shows the 'Template manager' interface. On the left, a sidebar lists categories: 'Description', 'Administrative metadata' (selected), 'Metadata' (expanded), 'Access policy', 'Overwrite if exists', 'Publish', and 'Embargo' (selected). Below these are 'Additional Fields' and a set of icons. The main panel is titled 'Demo administrative metadata template'. It contains the following fields:

- Key:** options.1739994271269
options.1739994271269
- Label:** Embargo
- Type:** string
- Description:** Date when project has to be converted from DRAFT to PUBLISH
- Original description:** N/A

Below these are tabs: DISPLAY (selected), CONTROLLED VOCABULARY, DEFAULT, VALIDATION RULES, and JSON. Under DISPLAY, there is a 'Data type:' section with 'string' selected.

Defining who can enter administrative metadata for a project

Administrative templates are designed, usually by IT/system experts, for the needs of specific data management or dissemination systems. The information to be entered in an administrative template when a project is documented will usually not be entered by the data curator (who will document the data), but by a system expert. This means that a different permission system will apply to the management of the content of administrative metadata.

The information on who has permission to enter administrative metadata is not set at the project level, but at the template level. The list of contributors authorized to enter content in a metadata template is defined in the ACL tab of the metadata sharing screen.

To access the screen, click on SHARE in the template menu (accessed by clicking on the triple-dot icon in the template list).

| Type | Default | Title | Language | Version | Owner | Last updated by | Updated on |
|--------|-----------------------|--|----------|---------|----------------|-----------------|------------|
| core | <input type="radio"/> | Core administrative metadata template | EN | | | | |
| custom | <input type="radio"/> | Core administrative metadata template - copy | en | | John Doe | John Doe | |
| custom | <input type="radio"/> | OD test chart type | en | | John Doe | John Doe | |
| custom | <input type="radio"/> | NADA | EN | 1.0.0 | admin admin | admin admin | |

A context menu is open on the last row, with the following options:

- SHARE (highlighted with a red box)
- DUPLICATE
- EXPORT
- DELETE

A popup menu will open, with the option to share the template itself (tab SHARE), and to enter the list of collaborators authorized to enter content when the template is used in a project (in tab ACL). Add the collaborators, making sure to give them Edit permission.

SHARE ACL

Administrative metadata access control

Select users who can use this template for filling out Administrative Metadata

| Search users to select | View | SHARE |
|--|------|---|
| Username | Role | |
| John Doe johndoe@ihsn.org | Edit | Delete |
| CLOSE | | |

Adding administrative templates to a project

Administrative metadata templates are added to a project by selecting one or multiple administrative templates in the *Templates* frame of the project home page.

The screenshot shows the 'Metadata Editor' interface for a project titled 'Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)'. On the left, there's a sidebar with filters for 'Required', 'Recommended', and 'Empty' templates, and a search bar. Below that is a navigation tree with 'Home' (selected), 'Preview', 'Metadata information', and 'Indicator description'. The main content area displays the project details. In the center, under the 'Template' section, it says 'Project template: IHSN INDICATOR 1.0 TEMPLATE V01 EN -'. Below that, the 'Administrative metadata templates:' section is highlighted with a red box. It contains two buttons: 'Demo administrative metadata template' and 'NADA'. To the right, there are sections for 'Collaborators' (None) and 'Collections' (None).

The added templates will be displayed in the navigation tree, under section *Administrative templates*.

To remove a template from a project, select the template in the navigation tree and click on DELETE.

The screenshot shows the 'Edit - Demo administrative metadata template' page in the Metadata Editor. The left sidebar has a tree structure with 'Administrative metadata' selected, and 'Demo administrative met' is highlighted. The main area shows a form with sections for 'Metadata', 'Overwrite if exists', 'Publish', and 'Embargo'. On the right, there is a 'Metadata' panel with fields for 'UID', 'Name', 'Created', and a 'PREVIEW' tab. A red box highlights the top right corner with 'DELETE', 'SAVE', and 'CANCEL' buttons, and another red box highlights the 'Metadata' section on the right.

Content can then be filled out by an authorized person.

The content entered in the template can be exported as JSON, and is accessible via API (see chapter [Metadata Editor API](#)).

See also section ***Administrative metadata*** in chapter **Documenting data - General instructions**

Managing projects

A **project** in the Metadata Editor represents a "dataset" of any type. This may include:

- A **microdataset** obtained from a survey, census, sensor, or administrative data recording system.
- An **indicator** or **database** of indicators.
- A vector or raster **geographic dataset** (or a geographic service).
- A **document** of any type.
- A research project with its associated **scripts**.
- An **image** or **video**.

The *My Projects* Page

The *MY PROJECTS* page is the default page of the Metadata Editor, where all projects available to a user are listed. Other main pages include *COLLECTIONS* and *TEMPLATES*, which are accessible only to users with credentials to manage these sections.

All users, regardless of their roles and permissions, can access the *My Projects* page. The content displayed in the page is determined by the user's credentials. Users will only see the projects they have permission to view or edit (i.e., the project they created or the projects that were shared with them).

| Showing 1 - 8 of 8 projects | | | | | |
|-----------------------------|---|----------|------------------|---------------|---------|
| | Title | Owner | Last modified by | Last modified | Actions |
| <input type="checkbox"/> | Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers | John Doe | John Doe | 2025-03-14 | |
| <input type="checkbox"/> | Outline of camps of Rohingya refugees in Cox's Bazar, Bangladesh <small>(My new collection)</small> | John Doe | John Doe | 2025-03-14 | |
| <input type="checkbox"/> | Popstan Synthetic Household Survey 2023 <small>Popstan, 2023</small> | John Doe | John Doe | 2025-03-13 | |

The list of projects can be filtered by:

- Type
- Collection (if applicable)
- Ownership (which distinguishes projects you own from projects shared with you by another user); you own a project if you created it, or if its ownership was transferred to you by another user.

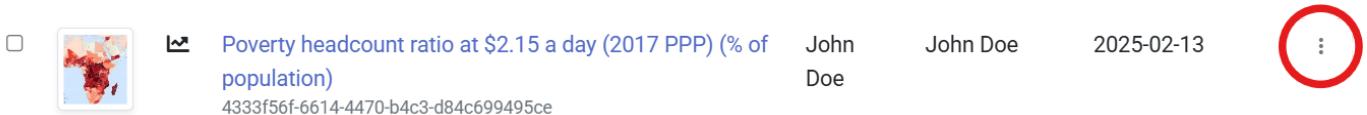
You can also **search** for projects using keywords. Note that the search only looks for keywords in the project title, not in the full project metadata.

Actions on a project

From the *My Projects* page, you can perform the following actions based on your role and permissions:

- **Create a new project.** Click **CREATE NEW PROJECT** and select a data type from the list. A new project page will open with an *Untitled* project. You can start entering content immediately, and the new project will be added to your *My Projects* list. You will be the *owner* of the project you create.
- **Import an existing project.** Click **IMPORT** and specify the project type. Upload a ZIP package generated by the Metadata Editor (a ZIP file that contains all materials related to the project, including the metadata generated using the Metadata Editor). This tool allows you to archive projects, and to share projects with other organizations (sharing projects with other users of the same instance of the Metadata Editor is done using the **SHARE** option, not by exporting/importing packages - see below).
- **Open an existing project.** Click on the project title in the list to open the project page and view or edit its content.

From a specific *Project* page, several actions can be performed by clicking the **Options** button.

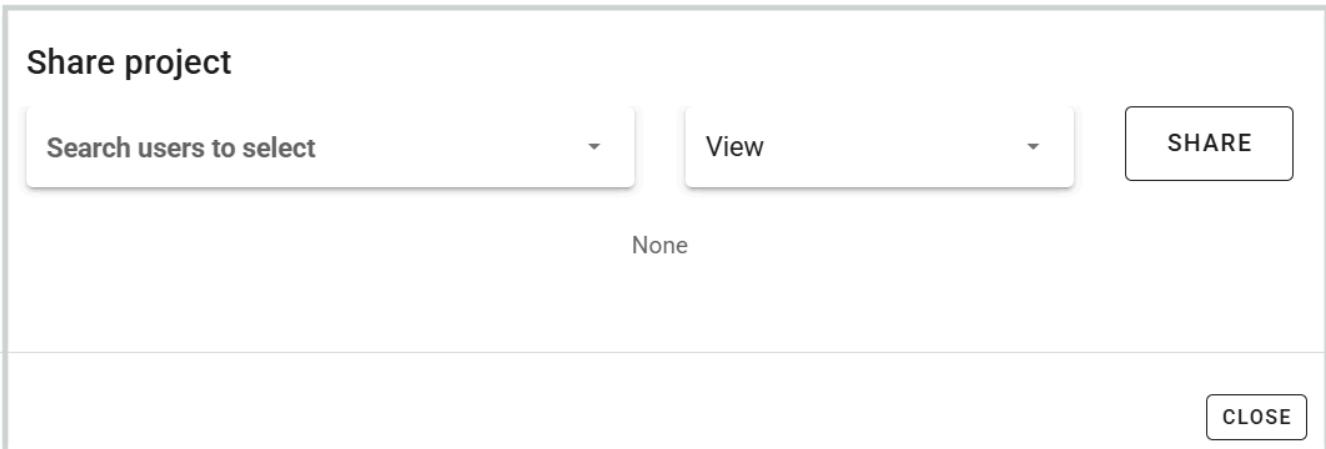


A screenshot of a project card. The card includes a checkbox, a small thumbnail image of a map, a title, a progress bar, names, a date, and a three-dot menu icon. A red circle highlights the three-dot menu icon.

| | | | | | | | |
|--------------------------------------|--|--|----------------------------------|----------|----------|------------|--|
| <input type="checkbox"/> |  | Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population) | <div style="width: 50%;"> </div> | John Doe | John Doe | 2025-02-13 |  |
| 4333f56f-6614-4470-b4c3-d84c699495ce | | | | | | | |

The available actions include:

- **SHARE** Share the project with one or more registered users. You can assign different levels of access: *View*, *Edit*, or *Admin*. Shared projects will then appear on the recipients' *My Projects* page.



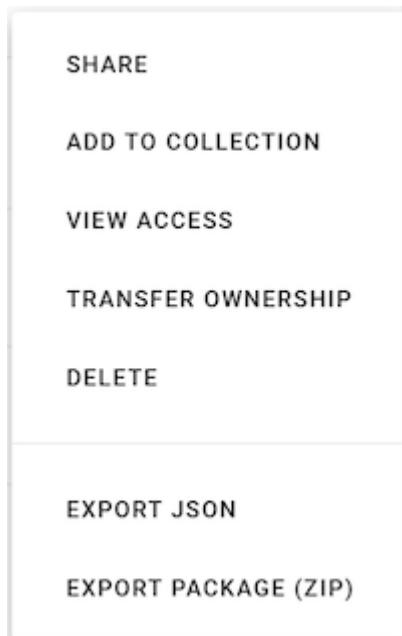
A screenshot of a 'Share project' dialog box. It features a title, two dropdown menus ('Search users to select' and 'View'), a 'SHARE' button, a note ('None'), and a 'CLOSE' button.

Share project

Search users to select

View

None



- **ADD TO COLLECTION** Add the project to one or more collections on which you have access permissions. A popup will prompt you to select the collection(s). See the *Managing Collections* section for more details.
- **VIEW ACCESS** Display information on project permissions, listing users with View, Edit, or Admin rights, as well as the collections the project belongs to. Permissions cannot be modified from this page. To change permissions for a project you own, open the project and use the *Collaborators* section on the *Project home page*.
- **TRANSFER OWNERSHIP** Transfer the ownership of a project you own to another registered user. You will be given the option to retain some level of access after the transfer.
- **DELETE** Delete a project where you have admin rights.
- **EXPORT JSON** Generate a JSON file containing the project metadata. You can exclude metadata elements marked as *Private* in the metadata template, and choose whether to include administrative metadata and metadata on external resources.
- **EXPORT PACKAGE (ZIP)** Generate a ZIP package with all project materials (data, metadata, and related resources). Use this option for archiving or sharing projects with another organization that can import the package into its own instance of the Metadata Editor.
- **LOCK & VERSION** Lock the project to prevent further edits and to assign a version identifier to the metadata. This option will be used when an organization's metadata governance requires a review and approval process. Upon locking, a reviewer assigns a version number to the metadata (following semantic versioning rules) and records relevant version details. Once locked, the metadata cannot be edited, but an editable copy remains available for future modifications and versioning. All locked versions are preserved and listed in the Metadata Editor. A checksum is generated for each locked version of the metadata and stored in the Metadata Editor database.

Batch actions on projects

You can select multiple projects using the checkboxes next to the project titles and apply batch actions. In the current version of the Metadata Editor, the only available batch action is to add the selected projects to a collection.

Showing 1 - 5 of 5 projects

Title

Add to collection

One of camps of Rohingya refugees in Cox's Bazar, Bangladesh

753c9a91-82c4-44f4-86c3-2bd0f7714af8

Market near Ramallah's main mosque

2c63f358-e5ef-4bb3-a654-a8968b6ba694

The project home page

Clicking on a project title in the *My Projects* page, or creating a new project, opens the project's *Home* page. If you are already working on a project, you can access the home page by clicking *Home* in the navigation tree.

Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)

IHSN INDICATOR 1.0 TEMPLATE V01 EN -

Project owner: John Doe
Created on: 2025-02-12 16:18:33

Last changed by: John Doe
Changed on: 2025-02-12 17:30:55

Project IDNO: 4333f56f-6614-4470-B4c3-D84c699495ce

Project validation

Schema validation: No validation errors found

Template validation: No validation errors found

Collaborators
None

Collections
None

Data and Documentation
Disk usage: 154.65 KB

| Title | Type |
|---------------------------------------|----------------|
| Poverty and Inequality Platform (PIP) | Web Site [web] |

The project *Home* page contains the following components:

- **Navigation tree.** Displays the metadata structure defined by the selected template for documenting the dataset.
- **Header.** Shows core information, including data type, creation and modification dates, project owner, project identifier (user-defined and system-generated), version, and status.
- **Template selection.** Allows selection of a metadata template and the option to add administrative metadata.
- **Project validation.** Lists violations, if any, of standard requirements (schema validation) and user-defined template rules (template validation).
- **Collaborators.** Lists registered users with access to the project, along with their assigned roles (*View, Edit, Admin*).
- **Collections.** Shows the collections to which the project belongs, if any.
- **File Management.** Displays files associated with the project and their storage details.

Special attention is required to the file management for microdata to ensure compliance with security and access rules. Data files uploaded to the server allow automatic generation of data dictionaries and summary statistics. Sensitive data should be deleted from the server to prevent unauthorized access or storage on unaccredited servers.

For more information about the available options in the project *Home* page and other project pages, refer to the section *Documenting Data – General Instructions*.

Managing collections

When a project is created, it is initially visible only to its owner (creator). However, projects can be shared with other registered users of the Metadata Editor, granting different levels of permission (View, Edit, or Admin). This enables project owners to invite specific individuals to collaborate. For further details, refer to the section *Managing Projects*.

Another way to share projects and foster collaboration is by publishing them in *Collections*. Collections (and sub-collections) serve as virtual containers, allowing projects to be grouped based on themes, teams, or other organizational criteria. A project can belong to multiple collections simultaneously.

Collections are particularly useful for organizing projects when the Metadata Editor contains a large number of them. On the *My Projects* page, users can filter project lists by collection. Additionally, collections enable metadata administrators to generate summary reports for specific project groups (refer to the section *Administrator Tools*).

The primary and most significant role of collections is to facilitate permission management. Organizations often prefer to assign roles to teams rather than individual users. Permissions can be granted at both the project level (using the **SHARE** option, see section *Managing Projects*) and the collection level. Each collection has a defined list of collaborators, and all users with access to a collection automatically gain access to all projects published within it.

To access the collection management page in the Metadata Editor, click on **COLLECTIONS** in the main menu.

The screenshot shows the Metadata Editor's navigation bar with three main tabs: 'PROJECTS', 'COLLECTIONS' (which is highlighted with a red box), and 'TEMPLATES'. Below the navigation bar, the title 'Collections' is displayed, followed by a 'CREATE NEW COLLECTION' button. A table header row includes columns for '#', 'Collection', 'Users', 'Projects', and a three-dot menu icon. The main content area below the table is currently empty, indicating no collections have been created yet.

Creating, editing, or deleting a collection

Creating collections requires specific credentials. The system administrator has full authority to create collections, including those at the root level. Other authorized users can create sub-collections under a designated root-level collection. Up to two sub-levels of collections can be created.

To create a new root-level collection, click on **CREATE NEW COLLECTION**. Enter a short title and an optional description for the collection.

Edit collection

Title

My new collection

Description

Collection created for demo purpose

SAVE

CLOSE

To create a sub-collection, open the *Options* menu for the parent collection and select **ADD SUB-COLLECTION**. Enter a name and description.

My new collection

1

0

⋮

EDIT

ADD SUB-COLLECTION

MANAGE ACCESS

DELETE

The **EDIT** button allows modification of the name and description of an existing collection.

The **DELETE** button removes a collection. Since collections serve as virtual containers, deleting a collection does not affect the projects within it.

Setting permissions for a collection

To manage access, use the **MANAGE ACCESS** option to add or modify users and their roles. This feature allows you to grant or revoke access and adjust user roles. All users in the access list can view, edit, or delete collection entries, depending on their assigned roles.

The screenshot shows the 'My new collection' page in the Metadata Editor. At the top, there's a header bar with 'Metadata Editor', 'About', 'English', and 'John Doe'. Below the header, a 'Return to collections' link is visible. The main section is titled 'My new collection' and contains a 'Add user' form with fields for 'Search users to select' and 'View'. A table lists a single user: 'John Doe' (email: johndoe@ihsn.org) with a role of 'ADMIN'. An 'Actions' column contains a red trash icon. A blue 'ADD' button is located at the bottom right of the 'Add user' form.

Adding projects to a collection

A project can be added to a collection from the *My Projects* page or the *Project* home page by selecting **ADD TO COLLECTION**. For additional details, refer to the sections *Managing Projects* and *Documenting Data*.

The screenshot shows the 'My Projects' page. It features a 'Collections' section with a list containing 'My new collection' (with an 'X' icon) and a large red circle highlighting a 'Folder +' icon, which is used to add projects to a collection.

The collections to which a project belongs are displayed on the *My Projects* page. To remove a project from a collection, click the "X" next to the collection name.

My projects

[PROJECTS](#)[COLLECTIONS](#)[TEMPLATES](#)[CREATE NEW PROJECT](#)[IMPORT](#) Search...

Recent ↑

Showing 1 - 5 of 5 projects

 [1](#)

| <input type="checkbox"/> | | Title | Owner | Last modified | Modified | Actions |
|--------------------------|--|--|----------|---------------|------------|---------|
| <input type="checkbox"/> | | Outline of camps of Rohingya refugees in Cox's Bazar, Bangladesh 733c9a91-82c4-44f4-8cc3-2bd0f7714af8 | John Doe | John Doe | 2025-02-15 | |

My new collection

Managing users and roles

The Metadata Editor operates as a centralized metadata production and management system. It allows the collaborative creation, editing, and deletion of projects. In this context, it is essential to clearly define what each user's role and permissions are. This is defined by (i) controlling who can access the Metadata Editor (who are the "members"), and (ii) defining the role of each member. The Metadata editor provides tools for defining roles, and for assigning them to members. The users and roles management system is accessible only to system administrators. It is accessed from the *Site administration* menu.

Global permission settings and account activation

Member

The Metadata Editor is a web-based application. It is accessed by entering a URL in a web browser. Everyone with access to this URL can open the URL. Typically, the Metadata Editor will be installed on an intranet, not internet. If the person is not registered as a "Member" of the Metadata Editor, opening this link will not provide any access to any information. A person becomes a member when the system administrator register that person, either by activating an organization's authentication system, or by manually registering members in the system.

A registered member can login to the application. By default, members have no access to any information until their account is activated by the system administrator. When a member logs-in, if her/his account has not been activated, a page will be shown allowing her/him to request activation.

When an account is activated, the system administrator gives the member a role as "Viewer" (the member will not have any active role in the Metadata Editor; s/he will only be able to see projects that have been shared with her/him) or as "Contributor" (the member will be able to create and edit projects; what s/he can do will be determined by her/his status for each project or collection).

Collection manager

The system administrator can assign the role of "Collection manager" to some members. Collection managers will be allowed to create and edit collections, and to manage permissions associated with collections.

Roles and permissions at project level

A set of pre-defined roles is provided by default in the Metadata Editor. Each role come with a specific set of permissions. The roles can be edited, and new ones can be created. When creating new roles, be aware that only one role can be assigned to a member.

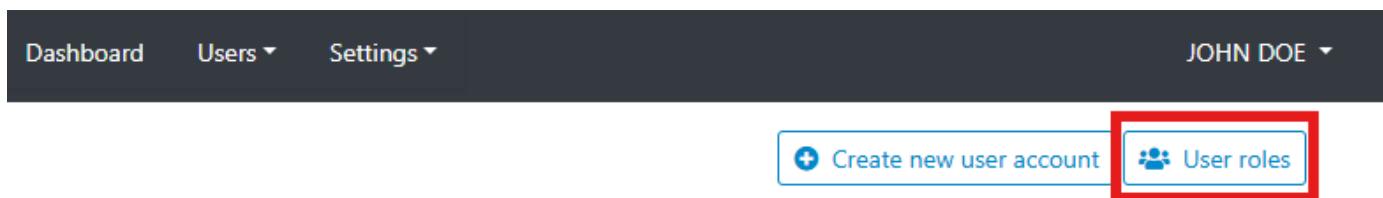
The following roles are pre-defined, which apply to specific **Projects**:

- **Member.** A *member* does not have any permission. This is the default role that is assigned when a user registers on the Metadata Editor (or is automatically registered when an organization's authentication system is configured to provide access to all authenticated staff members). Members will only obtain a set of permissions when a system administrator grants them specific roles.
- **Viewer.** The member is authorized to view the project (and export metadata), but NOT to export data or modify the project in any way.
- **Editor.** This role allows users to have full access to create and manage projects their own projects or any projects shared with them. The role allows view only access to templates and collections. The users can view, print (PDF, HTML preview), export to JSON all available templates but cannot make any changes to the templates or create new templates. For projects, users can view all available templates and change the template used by the project.
- **Owner.** The owner of a project is by default the member who created the project. The owner has full permissions on the project (edit, delete, share). Ownership can be transferred.
- **Owner with locking.** The owner of a project is by default the member who created the project. The owner has full permissions on the project (edit, delete, share). Ownership can be transferred.
- **Co-owner.** A co-owner of a project has the same permissions as the owner.
- **Co-owner with locking.** A co-owner of a project has the same permissions as the owner.
- **Editor and Reviewer.** An Editor and Reviewer has all permissions that an Editor has, plus te permission to lock and version projects.
- **Reviewer.** A Reviewer can view a project and lock/version it, but does not have Editor permissions.

For collections: The roles allow the user to manage collections such as create, edit, delete and add users to a collection. For adding projects to a collection, a user must have both admin/owner access to the project and admin/edit role for the collection.

Defining a new role

The editor allows creating new custom roles. To create a new role, Administrators can go to users under [site administration](#) and then click on the navigation link for [User Roles](#).



Roles and permissions for administrative data:

For administrative metadata, user access is managed via the template manager.

Publishing to NADA

For publishing projects to NADA data catalogues, user is required to have admin access to the NADA.

Permission levels in API

API key(s) inherit the same permissions as the user's interface access. Any operation a user is authorized to perform through the Metadata Editor UI can also be performed programmatically using their API key. Conversely, actions restricted in the UI remain inaccessible via the API as well.

General instructions

This section provides instructions related to the aspects of data documentation that apply to all data types. Instructions specific to each data type is provided in the subsequent sections.

The "My projects" page

Every dataset documented in the Metadata Editor is a **project**. A **project** therefore corresponds to a dataset or digital resource that needs to be documented. What a project includes depends on the type of data.

- For microdata, a project is one or a collection of data files, with all related resources.
- For indicators, a project is an indicator (or time series).
- For database, a project corresponds to the "container" of indicators. It includes a list of indicators, but not the indicators themselves.
- For geographic datasets, a project corresponds to one or multiple raster files, or to one or multiple vector data files.
- For research and scripts, a project corresponds to one or multiple scripts and outputs.
- For images, a project corresponds to one image (which may be available in multiple formats and resolutions)
- For videos, a project corresponds to one single video.

The "My projects" page is accessed by clicking on **PROJECTS** in the top menu.

See the **Managing projects** section for a detailed description of this page.

Creating a new project

For all data types, the documentation process for a new project consists of:

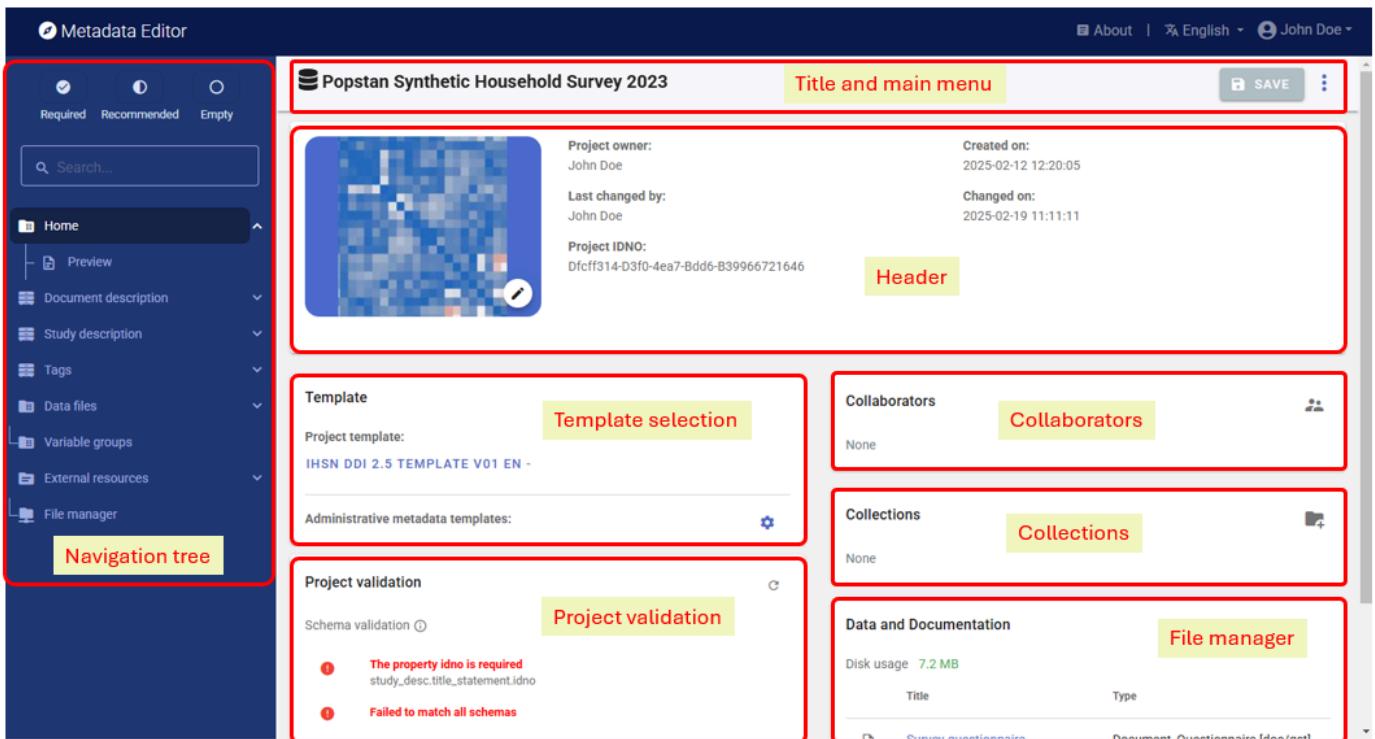
1. Creating a project (by clicking on **CREATE NEW PROJECT** or **IMPORT** in the My Projects page. Refer to the Quick Start examples).
2. Selecting a template for the documentation of the project (optional, as a default template is automatically selected).
3. Populating the metadata fields
4. Adding external resources
5. Adding administrative metadata (optional)
6. Locking and versioning the metadata (optional)
7. Exporting or publishing the metadata

When a project has already been created, its content can be edited by clicking on the project name in the *My projects* page.

This work can be made collaborative, by sharing the project with other data curators or publishing the project in one or multiple collection(s); see section **Managing projects** and **Managing collections**. A project will be edited by one or

multiple data curators. It may be reviewed, locked and versioned. A system expert may then add administrative metadata (this is usually not done by the data curators; it is a specific role).

The project "Home page"



Home and preview

Home opens the *Project home* page with its multiple frames, allowing project owners to select templates, manage files, and obtain summary information on the project.

Preview shows all metadata entered for the project in one single page, in a non-editable page.

Title and menu bar

The title and menu bar will display the project title (*untitled* until you added a project title and saved the metadata), the **SAVE** button, and a button (triple dots) that opens the project main menu. The icon shown in front of the project title indicates the data type of the project.

Header

The header frame in the *Project home* page provides information on the project identifiers (the editable user-defined primary identifier, and the read-only system-generated unique identifier), on the project owner (typically the project creator, or the user to whom the ownership was transferred), and on the dates the project was created and last modified. The Header also provides an option to select an image (JPG or PNG file) to be used as thumbnail. The thumbnail will be used in the Metadata Editor, and in NADA catalogs if the metadata are published in NADA.

Template selection

Template

Project template:

IHSN DDI 2.5 TEMPLATE V01 EN -

Administrative metadata templates:



- **Project template**

Datasets are documented based on templates. Templates are customized subsets of the metadata elements available from the metadata standard, possibly complemented by user-defined additional elements. The template will determine what data curators see in the navigation tree of the *Project* pages, and the metadata entry pages. A default template is automatically selected when a new project is created (the template to be used as default is selected by the system administrator; see section *Designing templates*).

The data curator can select another template available in the Metadata Editor. The list of available templates is also selected by the system administrator.

Changing the template used for an existing project will NOT impact the content of the metadata that has already been entered. No information is lost when a less comprehensive template is selected; all metadata already captured is preserved, even if it is not displayed when the new template is used.

- **Administrative metadata templates**

Select one or multiple administrative metadata templates (optional). See *Administrative metadata* below, and *Administrative metadata templates* in *Designing templates* for more information on the purpose and use of administrative metadata.

Navigation tree

The navigation tree shown in the *Project* page reflects the content of the selected templates. In a template, metadata elements can be tagged as *required* or *recommended*. The navigation frame provides an option to filter elements, to display only required fields or recommended fields. It also provides an option to only display *empty fields*, i.e. metadata elements for which no content has been provided.

A search box is also provided, allowing users to search a metadata element based on keywords found in the element label.

Project validation

The *Project home* page contains a frame titled *Project validation*, which will indicate whether the metadata that has been entered and saved violates some of the requirements of the standard itself or of the validation rules defined in the template used to document the dataset.

- **Schema validation** lists the violations of requirements of the metadata standard.
- **Template validation** lists the violations of custom validation rules defined in the metadata template.

Clicking on a validation error will take you to the element that needs to be edited.

Collaborators

The frame *Collaborators* in the *Project home* page will show the list of collaborators who have access to the project, with information on their permission level (View, Edit, Admin, Owner). If you are the owner or administrator of the project, you may edit this list (adding or removing collaborators) from the list by clicking on the icon on top of the frame.

| Username | Role |
|------------|------|
| [Redacted] | view |
| [Redacted] | view |

Collections

The frame *Collections* in the *Project home* page will show the list of collections to which the project belongs. If you are the owner or administrator of the project, you may edit this list (adding or removing collections) from the list by clicking on the icon on top of the frame.

File manager

| Data and Documentation | | Disk usage: 7.2 MB |
|---|-------|-----------------------------------|
| DOCUMENTATION | FILES | |
| | Title | Type |
|  Survey questionnaire | | Document, Questionnaire [doc/qst] |
|  Survey information | | Document, Technical [doc/tec] |
|  Full dataset in Stata 17 format | | Microdata File [dat/micro] |

| Data and Documentation | | | | Disk usage: 7.2 MB |
|------------------------|---|-----------|------------|--------------------|
| DOCUMENTATION | FILES | | | |
| | Name | Size | Created | |
| | dfcff314-d3f0-4ea7-bdd6-b39966721646.json | 131.87 KB | 2025-02-12 | |
| | thumbnail-4709.jpg | 13.99 KB | 2025-02-12 | |
| | WLD_2023_SYNTH-SVY-HLD-EN_v01_M.csv | 1.57 MB | 2025-02-12 | |
| | WLD_2023_SYNTH-SVY-IND-EN_v01_M.csv | 2.42 MB | 2025-02-12 | |
| | dfcff314-d3f0-4ea7-bdd6-b39966721646.rdf | 1.07 KB | 2025-02-12 | |
| | dfcff314-d3f0-4ea7-bdd6-b39966721646.rdf.json | 1.66 KB | 2025-02-12 | |

Common metadata sections

A few groups of metadata elements are common to all metadata standards, and will be found in the navigation tree.

Information on metadata

All metadata standards and schemas supported by the Metadata Editor include a set of elements intended to document the metadata itself. This set of elements is found in section *Document description* in the DDI (microdata), and *Metadata information* for other types of data. Although these elements are all optional, it is good practice to enter content in this section, if only on the author of the metadata and on the date it was generated. This information is not useful to the data users, but will be useful to catalog administrators to ensure traceability of the information stored in the metadata.

Tags

All metadata standards and schemas supported by the Metadata Editor include a **Tags** element (this element is not part of all standards; it has been added to standards that did not include it). This element enables the implementation of filters (facets) in data cataloguing applications, in a flexible manner. The tags metadata element is repeatable (meaning that more than one tag can be attached to a dataset) and contains two sub-elements to capture a tag (word or phrase), and the tag_group (if any) it belongs to.

To illustrate the use of tags, let's assume that you want to indicate whether a dataset is available free of charge or for a fee, and another tag that indicates whether the dataset meets differential privacy or not. None of the metadata schemas contains an element specifically designed to indicate the "free" or "for a fee" nature of the dataset, or "differentially private" or not. But this information can be captured in a tag "Free" or "For a fee" within a tag group that could be named "free_or_fee", and "Differentially private" or "Not differentially private" in a tag group that could be named "differential_privacy". This information becomes part of the metadata, and can be used by catalog administrators to create customized filtering options (facets) in their user interfaces.

| Tag | Tag group |
|----------------------------|----------------------|
| Free | free_or_fee |
| Not differentially private | differential_privacy |

External resources

External resources are not a specific type of data. They are resources of any type (data, document, web page, or any other type of resource that can be provided as an electronic file or a web link) that can be attached as a "related resource" to a catalog entry. A metadata schema that is intentionally kept very simple, based on the Dublin Core standard, is used to describe these resources.

The table below shows some examples of the kind of external resources that may be attached to the metadata of different data types.

| Data type | Examples of resources that may be documented and published as external resources |
|--------------------|--|
| Document | MS-Excel version of tables included in a publication ; PDF/DOC version of the publication ; visualizations files (scripts and image) for visualizations included in the publication ; link to electronic annexes |
| Microdata | survey questionnaire ; survey report ; technical documentation (sampling, etc.) ; data entry application ; survey budget in Excel ; microdata files in different formats ; link to an external website |
| Geographic dataset | link to an interactive web application ; technical documentation in PDF ; data analysis scripts ; publicly accessible data files |
| Time series | link to a database query interface ; technical documents ; link to external websites ; visualization scripts |
| Tables | link to an organization website ; tabulation scripts |
| Images | image files in different formats and resolutions ; link to a photo album application ; link to a photographer website |
| Audio recordings | audio file in MP3 or other format ; transcript in PDF |
| Videos | video file in WAV or other format ; transcript in PDF |
| Scripts | publication ; link to a package/library web page ; link to datasets |

To add an external resource, click on *External resources* in the navigation tree. This will open the **External resource** page which lists all external resources already added. The click on CREATE RESOURCE. This will open a new page, where information on the resource can be added. Enter at least the title, type, and upload a file or provide a URL. Then SAVE.

The *Resource type* element is very important; it will determine how the resource is published in a NADA catalog. Particular attention must be paid to resources of type *Microdata*. When publishing the resource in a NADA catalog, resources of type *Microdata* will not automatically be made available to users of the catalog; the access policy selected when publishing the project in NADA will apply. This could be "Open data" or "Direct access", which will make the data downloadable without restriction, but it could be another access policy such as "Licensed access" which would require that users request access to the data.

DataCite

A Digital Object Identifier (DOI) is a unique, persistent identifier assigned to a digital object, such as a research article, dataset, report, or other scholarly content. It provides a permanent link to the object, ensuring that it can always be reliably located, even if the URL or hosting platform changes. DOIs facilitate accurate citation, improve discoverability, and promote long-term access to digital resources, making them essential for maintaining the integrity and traceability of academic and scientific work.

A DOI is issued by a DOI Registration Agency (RA), which is a member of the International DOI Foundation (IDF). When a publisher, data repository, or other content provider wants to assign a DOI to a digital object, they register the object with

an RA, such as Crossref or DataCite. The content provider submits metadata about the object, including its title, authors, publication date, and a URL where the object can be accessed. The RA assigns a unique DOI, which is permanently linked to the metadata and the URL. This ensures that even if the object's location changes, the DOI remains a persistent identifier that redirects users to the correct location.

DataCite is a service that offers Fabrica as a DOI and metadata management service allowing organizations to register and manage DOIs for their data products (see <https://datacite.org/create-dois/>). With Fabrica, organizations can assign DOIs, maintain accurate and FAIR metadata, and ensure persistent links for long-term accessibility and citation of their valuable research outputs. Generating a DOI requires that a core set of metadata be provided to the DOI registration service. This section of the navigation tree contains the elements that are needed for that purpose. This section is only used when you plan to issue a DOI for the dataset.

The screenshot shows the DataCite section of the Metadata Editor. On the left, a sidebar lists categories: DataCite, DOI, Prefix, Suffix, Creators, Titles, Publisher, Publication year, Resource type, Resource type genera, URL, and Language. The main area is titled 'DataCite' and contains the following fields:

- DOI**: A text input field.
- Prefix**: A text input field.
- Suffix**: A text input field.
- Creators**: A table with columns: Name, Name type, Given name, Family name. It has an 'ADD ROW' button at the bottom right.
- Titles**: A table with columns: Title, Title type, Language. It has a minus sign at the bottom right.

Provenance

Projects can easily be shared across organizations. They can be shared by transferring ZIP packages or via API. When a project is published in a catalog like NADA, a project can be imported into the Metadata Editor from the NADA catalog using the API. This means that projects found in an instance of the Metadata Editor are not always created in the same instance of the application. When projects are obtained from an external source, it is important to keep track of (i) from where the project originated, and (ii) from where the project was imported, which may be different from the originating repository. To maintain traceability, all metadata standards supported by the Metadata Editor include a common *Provenance* section with two main components: (i) original repository, and (ii) source repository.

```

- "provenance": {
    - "original_repository": {
        "repository_name": "string",
        "url": "string",
        "dataset_identifier": "string",
        "doi": "string",
        "dataset_title": "string",
        "date_published": "string",
        "notes": "string"
    },
    - "source_repository": [
        - {
            "repository_name": "string",
            "url": "string",
            "dataset_identifier": "string",
            "dataset_title": "string",
            "date_acquired": "string",
            "acquisition_mode": "string",
            "notes": "string"
        }
    ]
},

```

The goal is to preserve information on at least the name, URL, and identifier of the original catalog, and the name, URL, identifier of the repository from which the project was imported, and its identifier and date/time when it was harvested.

Administrative metadata

Administrative metadata will only appear in the navigation tree if administrative metadata templates have been selected in the **Template selection**. Administrative metadata consists of information needed by systems or systems administrators to determine how a dataset should be published in a particular data dissemination platform. This information is not intended to be used or visible by the data users; they only serve an internal purpose for the organization that creates the data. For that reason, when a project metadata is exported, administrative metadata is by default not included.

The content of the administrative metadata is determined by the administrative metadata template(s) selected for the project. If the project (meta)data is intended to be published in multiple platforms, multiple templates will be used for the project.

The content of the administrative metadata section is not intended to be entered by data curators. They will typically be entered by system administrators. Entering administrative metadata therefore requires a specific role/permission (see *Setting roles and permissions**).

Project-level tools and options

Some elements in the menu apply to all data types, some are specific to the data type.

| Project | Metadata |
|--|--|
| <input checked="" type="checkbox"/> Export package (ZIP) | <input checked="" type="checkbox"/> Apply default values from template |
| <input checked="" type="checkbox"/> Export DDI Codebook | <input checked="" type="checkbox"/> Import project metadata |
| <input checked="" type="checkbox"/> Export JSON | <input checked="" type="checkbox"/> Import external resources |
| <input checked="" type="checkbox"/> Publish to NADA | External resources |
| <input checked="" type="checkbox"/> PDF documentation | <input checked="" type="checkbox"/> Export RDF/XML |
| <input checked="" type="checkbox"/> Change log | <input checked="" type="checkbox"/> Export RDF/JSON |

Saving and exporting project and metadata

- **Export package** A project contains the metadata you enter, the data files you may have imported (for microdata), the project thumbnail, and possibly external resources of different types. The Metadata Editor stores this information in a database and on the webserver that hosts the Editor. You may export a *package* that contains all materials related to the project. This package will consist of a ZIP file containing all files (including the metadata you entered, even if you did not export them). The ZIP file can be archived, or shared. A package can be imported in the Metadata Editor.
- **Export DDI Codebook** This option only applies to project of type *microdata*. It will generate a DDI Codebook file (XML format) containing the metadata, **not including** the metadata related to external resources.
- **Export JSON** This option applies to project of all types. It will generate a metadata file in JSON format. The file will contain the core metadata and other components based on the options you will select: including external resources, including metadata elements marked as *private* in the metadata template, and including the administrative metadata.

Export project metadata as JSON

- Export all fields
 Exclude fields marked as 'private' in the template
 Include external resources
 Include administrative metadata

EXPORT CLOSE

- **Export MSD (SDMX/XML 3.0):** (applies to project of type *indicator* only)
- **Export Metadataset (SDMX/JSON):** (applies to project of type *indicator* only)
- **Export RDF/XML:** This option will export the metadata related to the external resources as an Resource Description Framework (RDF) / XML file.

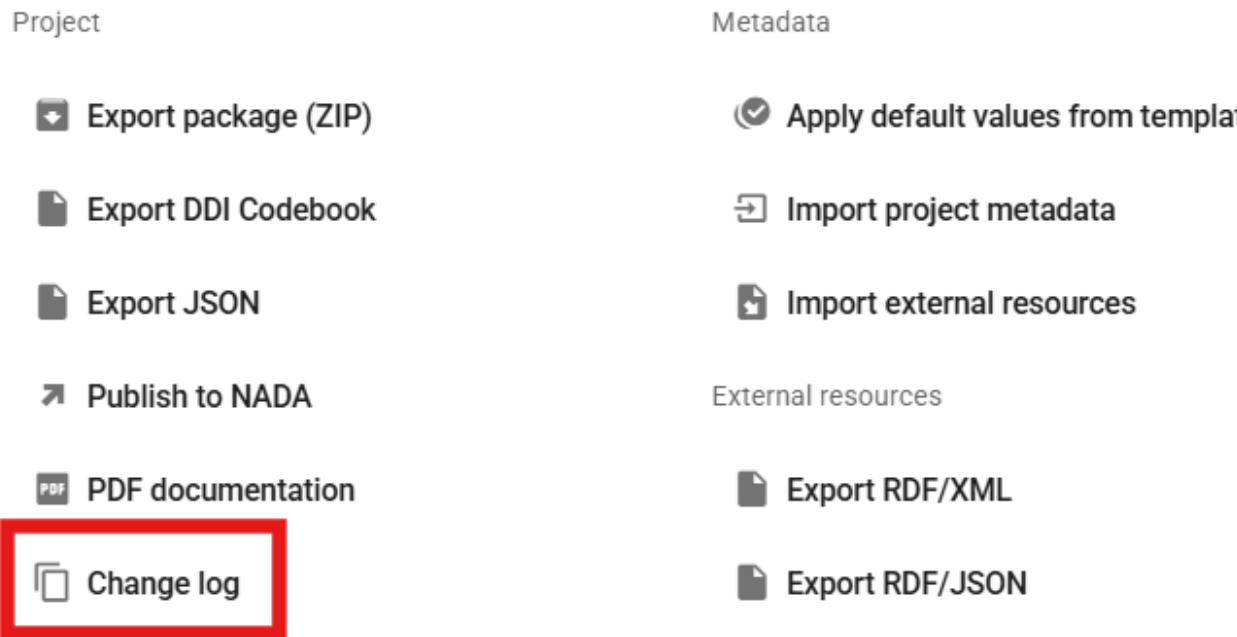
- **Export RDF/JSON:** This option will export the metadata related to the external resources as an Resource Description Framework (RDF) / JSON file.

Locking and versioning

- *Lock & version:

Publishing metadata

- **Publish to NADA:** This option allows you to publish your metadata (and related materials, optionally including data) to a NADA catalog. See chapter **Publish to NADA** for more information.
- **PDF documentation:** This option will generate a formatted PDF document containing the project metadata, including metadata on external resources.
- **Change log:** This option will open a page that shows all actions taken on the project, with identification of who took the action (change log). This option can be used to undo some actions.



| Date | User | Change Log (JSON) |
|---------------------|----------|---|
| 2025/02/12 06:12 PM | John Doe | [{"op": "add", "path": "/doc_desc", "value": {"producers": [...], "prod_date": "2025-02-12T05:00:00.000Z"}}, {...}, {...}, {...}] |
| 2025/02/12 05:20 PM | John Doe | null |

Importing metadata

- **Applying default values from template:** If the template used for documenting the dataset contains default values, this option will allow you to apply them to your project. Default values are not imputed in a project unless this option is selected. When you apply default value, you will be offered to apply default values to all metadata elements for which a default value exists, or only to metadata elements that do not contain any information (this option will protect information you may have entered against overwriting).

Apply default values from template

Apply defaults values defined in the template to the metadata

Update only empty fields
 Update all fields (overwrite existing values)

APPLY CLOSE

- **Import project metadata:** This option will allow you to select a metadata file (JSON) and import its content in your project. For microdata, you will be provided with an option to select the components of the metadata you want to import.

Import project metadata

Choose DDI/XML or a JSON file

 No file chosen

Options

Document description
 Study description
 File description
 Variable information
 Variable documentation
 Variable categories
 Variable questions
 Variable weights
 Variable groups

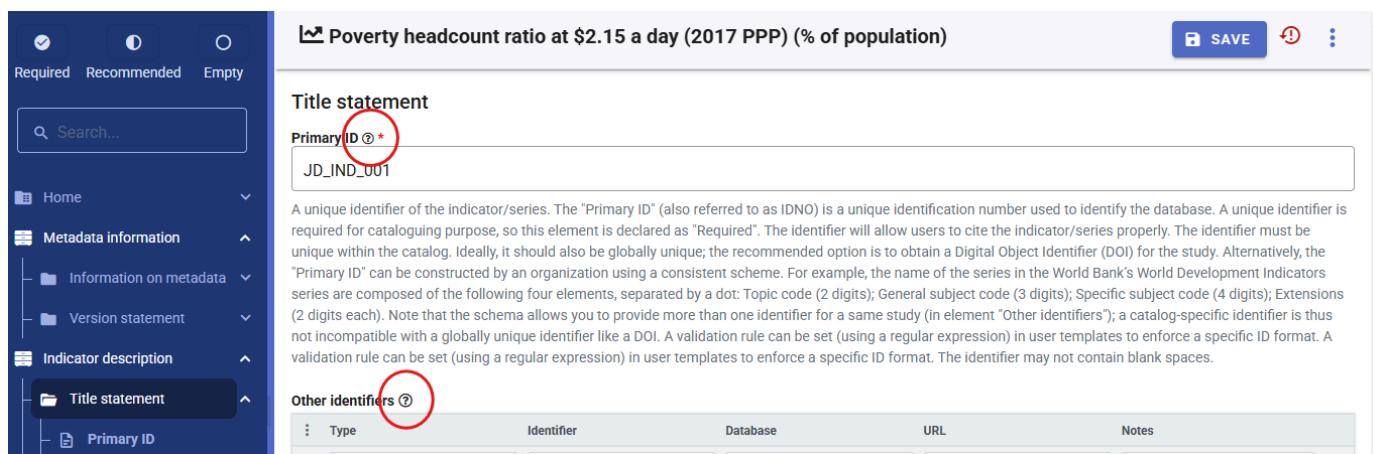
- **Import external resources:** This option allows you to import an external resource that has been exported from another project as a RDF/JSON file (see below). Select the Resource Description Framework (RDF) / JSON file and click IMPORT FILE.

Other tools and options

- **Diagnostic:** This option will open an HTML page that provides a diagnostic of the metadata for the project. The diagnostic will cover both the issues of required or recommended information that may be missing, errors in schema and template validation (also shown in the project Home page), and diagnostic based on other criteria based on good practice.

Help

The metadata entry pages will show a [?] icon next to the title of all metadata elements. Clicking on this icon will show the instructions for the element, extracted from the selected metadata template.



Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)

Title statement

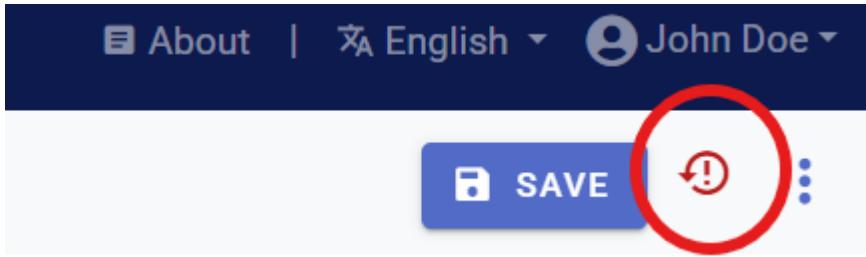
Primary ID (Required) *
JD_IND_001

A unique identifier of the indicator/series. The "Primary ID" (also referred to as IDNO) is a unique identification number used to identify the database. A unique identifier is required for cataloguing purpose, so this element is declared as "Required". The identifier will allow users to cite the indicator/series properly. The identifier must be unique within the catalog. Ideally, it should also be globally unique; the recommended option is to obtain a Digital Object Identifier (DOI) for the study. Alternatively, the "Primary ID" can be constructed by an organization using a consistent scheme. For example, the name of the series in the World Bank's World Development Indicators series are composed of the following four elements, separated by a dot: Topic code (2 digits); General subject code (3 digits); Specific subject code (4 digits); Extensions (2 digits each). Note that the schema allows you to provide more than one identifier for a same study (in element "Other identifiers"); a catalog-specific identifier is thus not incompatible with a globally unique identifier like a DOI. A validation rule can be set (using a regular expression) in user templates to enforce a specific ID format. A validation rule can be set (using a regular expression) in user templates to enforce a specific ID format. The identifier may not contain blank spaces.

Other identifiers (Optional)

| Type | Identifier | Database | URL | Notes |
|------|------------|----------|-----|-------|
| | | | | |

Canceling changes (Undo)



Formatted metadata

Metadata is intended to be used in many applications. It is therefore essential to make the file format as open and interoperable as possible. For that reason, the preferred format for saving metadata is JSON. The JSON format is non-proprietary plain text format. It is important to also ensure that the content of the file is as interoperable as possible. For that reason, including formatting in the metadata is not recommended, unless absolutely necessary. The metadata template may however mark some metadata elements to allow using markdown, HTML, or LaTex formulas (see *Designing templates*).

For example:

- ***Plain text (no formatting):***

The Metadata Editor is compatible with Version 2.5 of the DDI Codebook metadata standard.

- ***Formatted text entered as markdown:***

The Metadata Editor is compatible with **Version 2.5** of the *DDI Codebook* metadata standard.

- ***Formatted text entered as HTML:***

<p>The Metadata Editor is compatible with Version 2.5 of the DDI Codebook me

- ***Formatted text, as rendered in a web browser:***

The Metadata Editor is compatible with **Version 2.5** of the *DDI Codebook* metadata standard.

Formatting metadata elements in the Metadata Editor may also consist of entering LaTex formulas, which is a way of capturing complex formulas in plain text format, leaving it to other applications like web browsers to render the LaTex content into readable formulas. For example:

- ***Variance formula as entered in the Metadata Editor:***

```
\text{Variance: } \sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}
```

latex

- ***Variance formula as rendered in a web browser:***

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

Documenting microdata

Defining microdata

When surveys or censuses are conducted, or when administrative data are recorded, information is collected on each unit of observation. The unit of observation can be a person, a household, a firm, an agricultural holding, a facility, or other. **Microdata** are the data files resulting from these data collection activities. They contain the **unit-level** information (as opposed to aggregated data in the form of counts, means, or other). Each row in a microdata file is referred to as an **observation**. Information on each unit of observation is stored in **variables**, which can be of different types (e.g. numeric or character variables, with discrete or continuous values). These variables may contain data reported by the respondent (e.g., the marital status of a person), obtained by observation or measurement (e.g., the GPS location of a dwelling or other sensor), or generated by calculation, recoding or derivation (e.g., the sample weight in a survey).

For efficiency reasons, categorical variables are usually stored in numeric format (i.e. coded values). For example, the sex of a respondent may be stored in a variable named 'Q_01', and include values 1, 2, and 9 where 1 represents "Male", 2 represents "Female", and 9 represents "Unreported". Microdata must therefore be provided with a *data dictionary* (i.e., structural metadata) containing the variables and value labels and, for the derived variables (if any), some information of the derivation process.

Many other features (descriptive and reference metadata) of a micro-dataset should also be described such as the objectives and the methodology of data collection, a description of the sampling design for sample surveys, the period of data collection, the identification of the primary investigator and other contributors, the scope and geographic coverage of the data, and much more. This information is essential to make the microdata usable and discoverable.

Metadata standard: the DDI Codebook

The **Data Documentation Initiative (DDI)** metadata standard provides a structured and comprehensive list of metadata elements and attributes for the documentation of microdata. The DDI originated in the Inter-university Consortium for Political and Social Research (ICPSR), a membership-based organization with more than 500 member colleges and universities worldwide. The DDI is now the project of an alliance of North American and European institutions. Member institutions comprise many of the largest data producers and data archives in the world. The DDI standard is published under the terms of the [GNU General Public License]((<http://www.gnu.org/licenses>)) (version 3 or later).

The DDI standard is used by a large community of data archivists, including data libraries from the academia and research centers, national statistical agencies and other official data producing agencies, and international organizations. The DDI standard has two branches: the *DDI-Codebook* (version 2.x) and the *DDI LifeCycle* (version 3.x). These two branches serve different purposes and audiences.

The Metadata Editor implements the DDI-Codebook. Internally, it uses a slightly simplified version of the DDI Codebook 2.5, to which a few elements are added. The DDI Alliance publishes the DDI-Codebook as an XML schema. The Metadata Editor uses a JSON implementation of the schema; the Metadata Editor however exports fully-compliant DDI Codebook 2.5 metadata in XML format (among other options).

The DDI Alliance developed the DDI-Codebook for organizing the content, presentation, transfer, and preservation of metadata in the social and behavioral sciences. It enables documenting microdata files in a simultaneously flexible and rigorous way. The DDI-Codebook aims to provide a straightforward means of recording and communicating all the salient characteristics of a micro-dataset. It is designed to encompass the kinds of data resulting from surveys, censuses, administrative records, experiments, direct observation and other systematic methodology for generating empirical measurements. The unit of observation can be individual persons, households, families, business establishments, transactions, countries or other subjects of scientific interest.

The technical description of the JSON schema used for the documentation of microdata is available at <https://worldbank.github.io/metadata-schemas/#tag/Microdata>.

Before you start

Before you start documenting a micro-dataset, it is highly recommended to carefully prepare the data and the related materials.

- **Prepare your data files**
 - **Variable and value labels.** Ensure that all variables and values are labeled in the data files (if the data are stored in Stata, SPSS, or another application that allows documentation of variables).
 - **Direct identifiers and confidential information.** Drop the direct identifiers from the dataset (names, phone number of respondents, addresses, social security numbers, etc) and other confidential information if you plan to share the data.
 - **Unique identifiers and relationships.** Check that all observations in each data file has a unique identifier, in the form of a specific variable or a combination of variables. The unique identifiers can vary across data files. Ensure that there are no duplicated identifiers in any data file. If your dataset is composed of multiple related data files, check that the files can be merged without any issue. For example, if you have distinct data files at the household and individual levels (i.e., if you have a hierarchical data structure), use a statistical package to verify that all households have at least one corresponding individual, and that each individual belongs to one and only one household.
 - **Missing values.** It is preferable (but not required) to use system missing values (instead of values like '999') for indicating missing values. If missing values are indicated by values other than *system missing*, make sure you are aware of these values (which will have to be marked as representing *missing values* when documenting the data in the Metadata Editor).
 - **Temporary variables.** Drop all temporary variables (variables that were created for testing or other purpose, but that do not need to be kept in the dataset) and other unnecessary variables from the data files.
 - **Weighting.** For sample survey datasets, it is recommended to include the relevant sampling weight variables in all data files where they apply (for the convenience of data users).
 - **File names.** It is recommended to name your data files (and all other files you want to share) using a consistent naming convention, and in a way that will make it easier for users to understand the content of the file.
 - **Data file formats.** The Metadata Editor provides an option to read data files to automatically extract the metadata available in them. If necessary, export your data files to a format supported by the Metadata Editor.
- **Prepare the related materials to be included as external resources** External resources are all the electronic files (documents, data files, scripts, or other) that you want to preserve or disseminate with the data. All these digital resources should be gathered, and saved preferably in open or standard format under user-friendly names. When documenting a survey or census dataset for example, ensure that you:
 - Have a copy of the questionnaire(s) in electronic format. Include the file in both the original format and in a PDF version. If the survey was conducted using computer-assisted interviews using a software like Survey Solutions or CsPro, generate a PDF copy of the electronic form (Survey Solutions provide an option to generate such a file).

- Collect an electronic copy of all other relevant documents, such interviewer manuals, technical documentation on sampling, survey technical and analytical reports, presentations of results, press releases, etc. For documents available in non-standard format, generate a PDF copy.
 - Collect all other digital materials related to the dataset, such as data scripts, photos, videos, and others.
-

Create a new project

The first step in documenting a dataset is to create a new project. You do that by clicking on **CREATE NEW PROJECT** in the *My projects* page. Select *Microdata* as data type. This will open a new, untitled project *Home* page.

In the **Templates** frame, select the template you want to use to document the dataset. A default template is proposed; no action is needed if you want to use that template. Otherwise, switch to another template by clicking on the template name. Note that you can at any time change the template used for the documentation of a project. The selected template will determine what you see in the navigation tree and in the metadata entry pages.

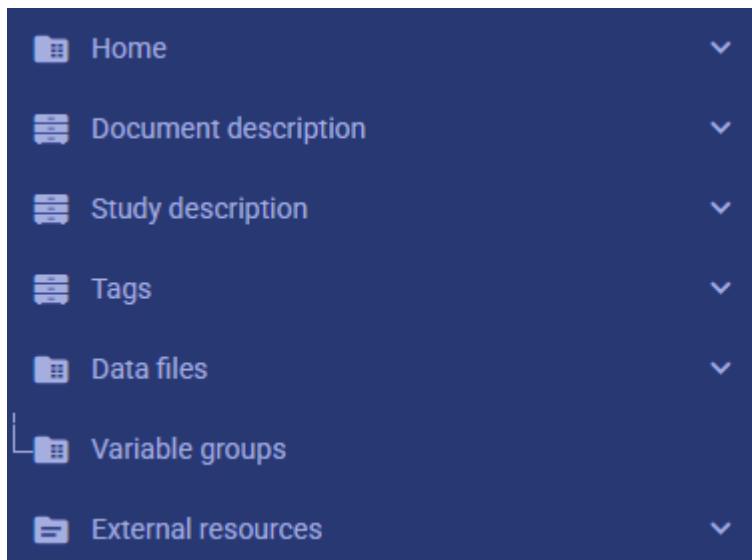
Switching from one template to another will not impact the metadata that has already been entered; no information will be deleted from the metadata.

Once a project has been created, you can import the data files (if available) and start documenting the dataset.

Import and document the dataset

The DDI Codebook metadata standard used to document microdata has the following sections (*containers*), which will be found in any template designed based on that standard:

- **Document description.** This section contains metadata on the metadata (to capture information on who documented the dataset, when, etc.)
- **Study description.** This section contains cataloguing and referential metadata that describe the study (survey, census, administrative recording system), with information like the title, geographic and temporal coverage, producers and sponsors, and many more.
- **File description.** This section provides a brief description of each data file.
- **Variable description.** The section contains the documentation of each variable, with elements like variable name and label, value labels, literal questions and interviewer instructions, universe, summary statistics, and more.
- **Variable groups.** This optional section provides a way to organize variables into groups other than the data files, e.g., grouping of variables by theme. Variable grouping is intended to increase the convenience to users, and possibly the online discoverability of data (if useful group descriptions are provided).



In addition to these sections specific to the DDI Codebook metadata standard, the Metadata Editor includes sections common to all data types: *External resources*, *Tags*, *DataCite*, *Provenance*, and (optional) *Administrative metadata*.

The DDI-specific sections are described below. For instructions on sections common to all data types, refer to the chapter **General instructions**.

The description of metadata elements provided below corresponds to the metadata elements included in the default template XXXX provided with the Metadata Editor. The list, the groupings, and the label of the elements may differ if you use a different template.

Document description

The **Document description** section of the navigation tree contains elements intended to document the metadata being generated ("metadata about the metadata"). In the DDI jargon, the *document* is the DDI-compliant metadata file (XML or JSON) that contains the structured metadata related to the dataset. This section corresponds to the section **Information on metadata** in other standards.

All content in this section is optional; it is however good and recommended practice to document the metadata as precisely as possible. This information will not be useful to data users, but it will be to catalog administrators. When metadata is shared across catalogs (or automatically harvested from DDI-compliant data catalogs), the information entered in the **Document description** provides transparency and clarity on the origin of the metadata.

Study description

The **Study description** section of the DDI Codebook metadata standard contains the **cataloguing and referential metadata**. Two metadata elements are required in this container: the *primary unique identifier* of the dataset, and its *title*. Other elements are optional, unless they have been set as *required* in a custom-defined template.

The metadata entered for this section should be as comprehensive as possible. You enter metadata by selecting a section in the navigation tree, and entering the information in the metadata entry page.

Note that all dates should preferably be entered in ISO format (YYYY-MM-DD or YYYY-MM or YYYY).

Refer to chapter **General instructions** for information on importing metadata from an existing project, and on copy/pasting content in fields that allow it.

By default, all content entered in the metadata entry page will consist of plain text (which will be saved in XML or JSON format). A few fields may allow entering formatted content (markdown, HTML, or LaTex formulas); which fields allow for such formating is determined in the template design.

We provide below a description of the metadata elements contained in the **Study description** section of the default metadata template provided in the Metadata Editor. Other templates may show a different selection, different labels, or present the elements in a different sequence. When you document a dataset, it is not expected that all these elements will be filled. Fill all required elements, all recommended elements when content can be made available, and fill as many as the other elements (the required and recommended elements will be those marked as *required* or *recommended* in the metadata standard or template).

In the list of metadata elements below, the *key* of each element in the metadata standard is provided between brackets next to the corresponding element's label in the template.

IDENTIFICATION

- **Title** (*title*) This element is "Required". Provide here the full authoritative title for the study. Make sure to use a unique name for each distinct study. The title should indicate the time period covered. For example, in a country conducting monthly labor force surveys, the title of a study would be like "Labor Force Survey, December 2020". When a survey spans two years (for example, a household income and expenditure survey conducted over a period of 12 months from June 2020 to June 2021), the range of years can be provided in the title, for example "Household Income and Expenditure Survey 2020-2021". The title of a survey should be its official name as stated on the survey questionnaire or in other study documents (report, etc.). Including the country name in the title is optional (another metadata element is used to identify the reference countries). Pay attention to the consistent use of capitalization in the title.
- **Subtitle** (*sub_title*) The **sub-title** is a secondary title used to amplify or state certain limitations on the main title, for example to add information usually associated with a sequential qualifier for a survey. For example, we may have "[country] Universal Primary Education Project, Impact Evaluation Survey 2007" as **title**, and "Baseline dataset" as **sub-title**. Note that this information could also be entered as a Title with no Subtitle: "[country] Universal Primary Education Project, Impact Evaluation Survey 2007 - Baseline dataset".
- **Alternate title** (*alternate_title*) The **alternate_title** will typically be used to capture the abbreviation of the survey title. Many surveys are known and referred to by their acronym. The survey reference year(s) may be included. For example, the "Demographic and Health Survey 2012" would be abbreviated as "DHS 2012", or the "Living Standards Measurement Study 2020-2021" as "LSMS 2020-2021".
- **Translated title** (*translated_title*) In countries with more than one official language, a translation of the title may be provided here. Likewise, the translated title may simply be a translation into English from a country's own language. Special characters should be properly displayed, such as accents and other stress marks or different alphabets.
- **Primary ID** (*idno*) **idno** is the primary identifier of the dataset. It is a unique identification number used to identify the study (survey, census or other). A unique identifier is required for cataloguing purpose, so this element is declared as "Required". The identifier will allow users to cite the dataset properly. The identifier must be unique within the catalog. Ideally, it should also be globally unique; the recommended option is to obtain a Digital Object Identifier (DOI) for the study. Alternatively, the **idno** can be constructed by an organization using a consistent scheme. The scheme could for example be "catalog-country-study-year-version", where country is the 3-letter ISO country code, producer is the abbreviation of the producing agency, study is the study acronym, year is the reference year (or the year the study started), version is a version number. Using that scheme, the Uganda 2005 Demographic and Health Survey for example would have the following **idno** (where "MDA" stand for "My Data Archive"): MDA_UGA_DHS_2005_v01. Note that the schema allows you to provide more than one identifier for a same study (in element **identifiers**); a catalog-specific identifier is thus not incompatible with a globally unique identifier like a DOI. The identifier should not contain blank spaces.
- **Other identifiers** (*identifiers*) This repeatable element is used to enter identifiers (IDs) other than the **idno** entered in the Title statement. It can for example be a Digital Object Identifier (DOI). The **idno** can be repeated here (the **idno** element does not provide a **type** parameter; if a DOI or other standard reference ID is used as **idno**, it is recommended to repeat it here with the identification of its **type**).

- **Type** (type) The type of unique ID, e.g. "DOI".
- **Identifier** (identifier) The identifier itself.
- **Study type** (series_name) The name of the series to which the study belongs. For example, "Living Standards Measurement Study (LSMS)" or "Demographic and Health Survey (DHS)" or "Multiple Indicator Cluster Survey VII (MICS7)". A description of the series can be provided in the element "series_info".
- **Series information** (series_info) A study may be repeated at regular intervals (such as an annual labor force survey), or be part of an international survey program (such as the MICS, DHS, LSMS and others). The series statement provides information on the series. The element is a brief description of the characteristics of the series, including when it started, how many rounds were already implemented, and who is in charge would be provided here.

VERSION

- **Version name** (version) The version number, also known as release or edition.
- **Version date** (version_date) The ISO 8601 standard for dates (YYYY-MM-DD) is recommended for use with the "date" attribute.
- **Version responsibility** (version_resp) The person(s) or organization(s) responsible for this version of the study.
- **Notes on version** (version_notes) Version notes should provide a brief report on the changes made through the versioning process. The note should indicate how this version differs from other versions of the same dataset.

OVERVIEW

- **Abstract** (abstract) An un-formatted summary describing the purpose, nature, and scope of the data collection, special characteristics of its contents, major subject areas covered, and what questions the primary investigator(s) attempted to answer when they conducted the study. The summary should ideally be between 50 and 5000 characters long. The abstract should provide a clear summary of the purposes, objectives and content of the survey. It should be written by a researcher or survey statistician aware of the study. Inclusion of this element is strongly recommended.
- **Kind of data** (data_kind) This field describes the main type of microdata generated by the study: 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. A controlled vocabulary should be used as this information may be used to build facets (filters) in a catalog user interface.
- **Unit of analysis** (analysis_unit) A study can have multiple units of analysis. This field will list the various units that can be analyzed. For example, a Living Standard Measurement Study (LSMS) may have collected data on households and their members (individuals), on dwelling characteristics, on prices in local markets, on household enterprises, on agricultural plots, and on characteristics of health and education facilities in the sample areas.

SCOPE

- **Keywords** (keywords) Keywords are words or phrases that describe salient aspects of a data collection's content. The addition of keywords can significantly improve the discoverability of data. Keywords can summarize and improve the description of the content or subject matter of a study. For example, keywords "poverty", "inequality", "welfare", and "prosperity" could be attached to a household income survey used to generate poverty and inequality indicators (for which these keywords may not appear anywhere else in the metadata). A controlled vocabulary can be employed. Keywords can be selected from a standard thesaurus, preferably an international, multilingual thesaurus.
 - **Keyword** (keyword) A keyword (or phrase).
 - **Vocabulary** (vocab) The controlled vocabulary from which the keyword is extracted, if any.
 - **URL** (uri) The URI of the controlled vocabulary used, if any.
- **Topics** (topics) The **topics** field indicates the broad substantive topic(s) that the study covers. A topic classification facilitates referencing and searches in on-line data catalogs.

- **Topic** (**) The label of the topic. Topics should be selected from a standard controlled vocabulary such as the [Council of European Social Science Data Archives \(CESSDA\) Topic Classification](#).
- **Vocabulary** (vocab) The specification (name including the version) of the controlled vocabulary in use.
- **URL** (uri) A link (URL) to the controlled vocabulary website.

UNIVERSE AND GEOGRAPHIC COVERAGE

- **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. 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.
- **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.
 - **Name** (name) The country name, even in cases where the study does not cover the entire country.
 - **Code** (abbreviation) The **abbreviation** will contain a country code, preferably the 3-letter [ISO 3166-1 country code](#).
- **Geographic coverage** (geog_coverage) Information on the geographic coverage of the study. This includes the total geographic scope of the data, and any additional levels of geographic coding provided in the variables. Typical entries will be "National coverage", "Urban areas", "Rural areas", "State of ...", "Capital city", etc. This does not describe where the data were collected; it describes which area the data are representative of. This means for example that a sample survey could be declared as having a national coverage even if some districts of the country were not included in the sample, as long as the sample is nationally representative.
- **Geographic coverage notes** (geog_coverage_notes) Additional information on the geographic coverage of the study entered as a free text field.
- **Geographic unit** (geog_unit) Describes the levels of geographic aggregation covered by the data. Particular attention must be paid to include information on the lowest geographic area for which data are representative.
- **Bounding box** (bbox) This element is used to define one or multiple bounding box(es), which are the rectangular fundamental geometric description of the geographic coverage of the data. A bounding box is defined by west and east longitudes and north and south latitudes, and includes the largest geographic extent of the dataset's geographic coverage. The bounding box provides the geographic coordinates of the top left (north/west) and bottom-right (south/east) corners of a rectangular area. This element can be used in catalogs as the first pass of a coordinate-based search. This element is optional, but if the **bound_poly** element (see below) is used, then the **bbox** element must be included.
 - **West** (west) West longitude of the bounding box.
 - **East** (east) East longitude of the bounding box.
 - **South** (south) South latitude of the bounding box.
 - **North** (north) North latitude of the bounding box.
- **Bounding polygon** (bound_poly) The **bbox** metadata element (see above) describes a rectangular area representing the entire geographic coverage of a dataset. The element **bound_poly** allows for a more detailed description of the geographic coverage, by allowing multiple and non-rectangular polygons (areas) to be described. This is done by providing list(s) of latitude and longitude coordinates that define the area(s). It should only be used to define the outer boundaries of the covered areas. This field is intended to enable a refined coordinate-based search, not to actually map an area. Note that if the **bound_poly** element is used, then the element **bbox** MUST be present as well, and all points enclosed by the **bound_poly** MUST be contained within the bounding box defined in **bbox**.

- **Latitude** (*lat*) The latitude of the coordinate.
- **Longitude** (*lon*) The longitude of the coordinate.

PRODUCERS AND SPONSORS

- **Primary producer/investigator** (*authoring_entity*) The name and affiliation of the person, corporate body, or agency responsible for the study's substantive and intellectual content (the "authoring entity" or "primary investigator"). Generally, in a survey, the authoring entity will be the institution implementing the survey. Repeat the element for each authoring entity, and enter the **Affiliation** when relevant. If various institutions have been equally involved as main investigators, then should all be listed. This only includes the agencies responsible for the implementation of the study, not sponsoring agencies or entities providing technical assistance (for which other metadata elements are available). The order in which authoring entities are listed is discretionary. It can be alphabetic or by significance of contribution. Individual persons can also be mentioned, if not prohibited by privacy protection rules.
 - **Name** (*name*) The name of the person, corporate body, or agency responsible for the work's substantive and intellectual content. The primary investigator will in most cases be an institution, but could also be an individual in the case of small-scale academic surveys. If persons are mentioned, use the appropriate format of *Surname, First name*.
 - **Affiliation** (*affiliation*) The affiliation of the person, corporate body, or agency mentioned in **Name**.
- **Other producers** (*producers*) This field is provided to list other interested parties and persons that have played a significant but not the leading technical role in implementing and producing the data (which will be listed in **authoring_entity**), and not the financial sponsors (which will be listed in **funding_agencies**).
 - **Name** (*name*) The name of the person or organization.
 - **Abbreviation** (*abbr*) The official abbreviation of the organization mentioned in **Name**.
 - **Affiliation** (*affiliation*) The affiliation of the person or organization mentioned in **Name**.
 - **Role** (*role*) A succinct description of the specific contribution by the person or organization in the production of the data.
- **Funding agencies** (*funding_agencies*) The source(s) of funds for the production of the study. If different funding agencies sponsored different stages of the production process, use the **role** attribute to distinguish them.
 - **Name** (*name*) The name of the funding agency.
 - **Abbreviation** (*abbr*) The abbreviation (acronym) of the funding agency mentioned in **Name**.
 - **Grant number** (*grant*) The grant number. If an agency has provided more than one grant, list them all separated with a ";".
 - **Role** (*role*) The specific contribution of the funding agency mentioned in **Name**. This element is used when multiple funding agencies are listed to distinguish their specific contributions.
- **Budget** (*study_budget*) This is a free-text field, not a structured element. The budget of a study will ideally be described by budget line. The currency used to describe the budget should be specified. This element can also be used to document issues related to the budget (e.g., documenting possible under-run and over-run).
- **Other contributors** (*oth_id*) This element is used to acknowledge any other people and organizations that have in some form contributed to the study. This does not include other producers which should be listed in **producers**, and financial sponsors which should be listed in the element **funding_agencies**.
 - **Name** (*name*) The name of the person or organization.
 - **Affiliation** (*affiliation*) The affiliation of the person or organization mentioned in **Name**.
 - **Role** (*role*) A brief description of the specific role of the person or organization mentioned in **Name**.

STUDY AUTHORIZATION

Provides structured information on the agency that authorized the study, the date of authorization, and an authorization statement. This element will be used when a special legislation is required to conduct the data collection (for example a Census Act) or when the approval of an Ethics Board or other body is required to collect the data.

- **Authorization date** (date) The date, preferably entered in ISO 8601 format (YYYY-MM-DD), when the authorization to conduct the study was granted.
- **Authorizing agency** (agency) Identification of the agency that authorized the study.
 - **Name** (name) Name of the agent or agency that authorized the study.
 - **Affiliation** (affiliation) The institutional affiliation of the authorizing agent or agency mentioned in **name**.
 - **Abbreviation** (abbr) The abbreviation of the authorizing agent's or agency's name.
- **Authorization statement** (authorization_statement) The text of the authorization (or a description and link to a document or other resource containing the authorization statement).

SAMPLING

- **Sample frame** A description of the sample frame used for identifying the population from which the sample was taken. For example, a telephone book may be a sample frame for a phone survey. Or the listing of enumeration areas (EAs) of a population census can provide a sample frame for a household survey. In addition to the name, label and text describing the sample frame, this structure lists who maintains the sample frame, the period for which it is valid, a use statement, the universe covered, the type of unit contained in the frame as well as the number of units available, the reference period of the frame and procedures used to update the frame.
 - **Name** (name) The name (title) of the sample frame.
 - **Valid periods** (valid_period) Defines a time period for the validity of the sampling frame, using a list of events and dates.
 - **Event** (event) The event can for example be **start** or **end**.
 - **Date** (date) The date corresponding to the event, entered in ISO 8601 format: YYYY-MM-DD.
 - **Custodian** (custodian) Custodian identifies the agency or individual responsible for creating and/or maintaining the sample frame.
 - **Universe** (universe) A description of the universe of population covered by the sample frame. 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, etc. The universe may consist of elements other than persons, such as housing units, court cases, deaths, countries, 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 included in the sample frame.
 - **Unit type** (unit_type) The type of the sampling frame unit (for example "household", or "dwelling").
 - **Units are primary** (is_primary) This boolean attribute (true/false) indicates whether the unit is primary or not.
 - **Number of units** (num_of_units) The number of units in the sample frame, possibly with information on its distribution (e.g. by urban/rural, province, or other).
 - **Update procedure** (update_procedure) This element is used to describe how and with what frequency the sample frame is updated. For example: "The lists and boundaries of enumeration areas are updated every ten years at the occasion of the population census cartography work. Listing of households in enumeration areas are updated as and when needed, based on their selection in survey samples."
 - **Reference periods** (reference_period) Indicates the period of time in which the sampling frame was actually used for the study in question. Use ISO 8601 date format to enter the relevant date(s).
 - **Event** (event) Indicates the type of event that the date corresponds to, e.g., "start", "end", "single".
 - **Date** (date) The relevant date in ISO 8601 date/time format.
 - **Sampling procedure** (sampling_procedure) This field only applies to sample surveys. It describes the type of sample and sample design used to select the survey respondents to represent the population. This section should include summary information that includes (but is not limited to): sample size (expected and actual) and how the

sample size was decided; level of representation of the sample; sample frame used, and listing exercise conducted to update it; sample selection process (e.g., probability proportional to size or over sampling); stratification (implicit and explicit); design omissions in the sample; strategy for absent respondents/not found/refusals (replacement or not). Detailed information on the sample design is critical to allow users to adequately calculate sampling errors and confidence intervals for their estimates. To do that, they will need to be able to clearly identify the variables in the dataset that represent the different levels of stratification and the primary sampling unit (PSU).

In publications and reports, the description of sampling design often contains complex formulas and symbols. As the XML and JSON formats used to store the metadata are plain text files, they cannot contain these complex representations. You may however provide references (title/author/date) to documents where such detailed descriptions are provided, and make sure that the documents (or links to the documents) are provided in the catalog where the survey metadata are published.

- **Deviations from sample design** (*sampling_deviation*) 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.
- **Response rates** (*response_rate*) The response rate is the percentage of sample units that participated in the survey based on the original sample size. Omissions may occur due to refusal to participate, impossibility to locate the respondent, or other reason. This element is used to provide a narrative description of the response rate, possibly by stratum or other criteria, and if possible with an identification of possible causes. If information is available on the causes of non-response (refusal/not found/other), it can be reported here. This field can also be used to describe non-responses in population censuses.
- **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.

SURVEY INSTRUMENT

- **Questionnaires** (*research_instrument*) The research instrument refers to the questionnaire or form used for collecting data. The following should be mentioned:
 - List of questionnaires and short description of each (all questionnaires must be provided as External Resources)
 - In what language(s) was/were the questionnaire(s) available?
 - Information on the questionnaire design process (based on a previous questionnaire, based on a standard model questionnaire, review by stakeholders). If a document was compiled that contains the comments provided by the stakeholders on the draft questionnaire, or a report prepared on the questionnaire testing, a reference to these documents can be provided here.
- **Instrument development** (*instru_development*) Describe any development work on the data collection instrument. This may include a description of the review process, standards followed, and a list of agencies/people consulted.
- **Notes on methodology** (*method_notes*) This element is provided to capture any additional relevant information on the data collection methodology, which could not fit in the previous metadata elements.

DATA COLLECTION

- **Dates of data collection** (*coll_dates*) Contains the date(s) when the data were collected, which may be different from the date the data refer to (see **time_periods** above). For example, data may be collected over a period of 2 weeks (**coll_dates**) about household expenditures during a reference week (**time_periods**) preceding the beginning of data collection. Use the event attribute to specify the "start" and "end" for each period entered.
 - **Start** (*start*) Date the data collection started (for the specified cycle, if any). Enter the date in ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY).

- **End** (end) Date the data collection ended (for the specified cycle, if any). Enter the date in ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY).
- **Cycle** (cycle) Identification of the cycle of data collection. The `cycle` attribute permits specification of the relevant cycle, wave, or round of data. For example, a household consumption survey could visit households in four phases (one per quarter). Each quarter would be a cycle, and the specific dates of data collection for each quarter would be entered.
- **Time method** (time_method) The time method or time dimension of the data collection. A controlled vocabulary can be used. The entries for this element may include "panel survey", "cross-section", "trend study", or "time-series".
- **Frequency** (frequency) For data collected at more than one point in time, the frequency with which the data were collected.
- **Time periods** (time_periods) This refers to the time period (also known as span) covered by the data, not the dates of data collection.
 - **Start** (start) The start date for the cycle being described. Enter the date in ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY).
 - **End** (end**) The end date for the cycle being described. Enter the date in ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY). Indicate open-ended dates with two decimal points (..)
 - **Cycle** (cycle) The `cycle` attribute permits specification of the relevant cycle, wave, or round of data.
- **Data collection sources** (sources) A description of sources used for developing the methodology of the data collection.
 - **Name** (name) The name and other information on the source. For example, "United States Internal Revenue Service Quarterly Payroll File"
 - **Origin** (origin) For historical materials, information about the origin(s) of the sources and the rules followed in establishing the sources should be specified. This may not be relevant to survey data.
 - **Characteristics** (characteristics) Assessment of characteristics and quality of source material. This may not be relevant to survey data.
- **Mode of data collection** (coll_mode) The mode of data collection is the manner in which the interview was conducted or information was gathered. Ideally, a controlled vocabulary will be used to constrain the entries in this field (which could include items like "telephone interview", "face-to-face paper and pen interview", "face-to-face computer-assisted interviews (CAPI)", "mail questionnaire", "computer-aided telephone interviews (CATI)", "self-administered web forms", "measurement by sensor", and others).

This is a repeatable field, as some data collection activities implement multi-mode data collection (for example, a population census can offer respondents the options to submit information via web forms, telephone interviews, mailed forms, or face-to-face interviews. Note that in the API description (see screenshot above), the element is described as having type "null", not {}). This is due to the fact that the element can be entered either as a list (repeatable element) or as a string.
- **Data collectors** (data_collectors) The entity (individual, agency, or institution) responsible for administering the questionnaire or interview or compiling the data.
 - **Name** (name) In most cases, we will record here the name of the agency, not the name of interviewers. Only in the case of very small-scale surveys, with a very limited number of interviewers, the name of persons will be included as well.
 - **Affiliation** (affiliation) The affiliation of the data collector mentioned in `name`.
 - **Abbreviation** (abbr) The abbreviation given to the agency mentioned in `name`.
 - **Role** (role) The specific role of the person or agency mentioned in `name`.
 - **Collector training** (collector_training) Describes the training provided to data collectors including interviewer training, process testing, compliance with standards etc. This set of elements is repeatable, to capture different

aspects of the training process.

- **Type** (type) The type of training being described. For example, "Training of interviewers", "Training of controllers", "Training of cartographers", "Training on the use of tablets for data collection", etc.
- **Training** (training) A brief description of the training. This may include information on the dates and duration, audience, location, content, trainers, issues, etc.
- **Control operations** (control_operations) This element will provide information on the oversight of the data collection, i.e. on methods implemented to facilitate data control performed by the primary investigator or by the data archive.
- **Supervision** (act_min) A summary of actions taken to minimize data loss. This includes information on actions such as follow-up visits, supervisory checks, historical matching, estimation, etc. Note that this element does not have to include detailed information on response rates, as a specific metadata element is provided for that purpose in section **analysis_info / response_rate** (see below).
- **Notes on data collection** (coll_situation) A description of noteworthy aspects of the data collection situation. This element is provided to document any specific situations, observations, or events that occurred during data collection. Consider stating such items like:
 - Was a training of enumerators held? (elaborate)
 - Was a pilot survey conducted?
 - Did any events have a bearing on the data quality? (elaborate)
 - How long did an interview take on average?
 - In what language(s) were the interviews conducted?
 - Were there any corrective actions taken by management when problems occurred in the field?

DATA PROCESSING

- **Data processing** (data_processing) This element is used to describe how data were electronically captured (e.g., entered in the field, in a centralized manner by data entry clerks, captured electronically using tablets and a CAPI application, via web forms, etc.). Information on devices and software used for data capture can also be provided here. Other data processing procedures not captured elsewhere in the documentation can be described here (tabulation, etc.)
 - **Type** (type) The type attribute supports better classification of this activity, including the optional use of a controlled vocabulary. The vocabulary could include options like "data capture", "data validation", "variable derivation", "tabulation", "data visualizations", anonymization", "documentation", etc.
 - **Description** (description) A description of a data processing task.
- **Cleaning operations** (cleaning_operations) A description of the methods used to clean or edit the data, e.g., consistency checking, wild code checking, etc. The data editing should contain information on how the data was treated or controlled for in terms of consistency and coherence. This item does not concern the data entry phase but only the editing of data whether manual or automatic. It should provide answers to questions like: Was a hot deck or a cold deck technique used to edit the data? Were corrections made automatically (by program), or by visual control of the questionnaire? What software was used? If materials are available (specifications for data editing, report on data editing, programs used for data editing), they should be listed here and provided as external resources in data catalogs (the best documentation of data editing consists of well-documented reproducible scripts).

STUDY ACTIVITIES

This section is used to describe the process that led to the production of the final output of the study, from its inception/design to the dissemination of the final output.

Each activity will be documented separately. The [Generic Statistical Business Process Model \(GSBPM\)](#) provides a useful decomposition of such a process, which can be used to list the activities to be described. This is a repeatable set of

metadata elements; each activity should be documented separately.

- **Type** (*activity_type*) The type of activity. A controlled vocabulary can be used, possibly comprising the main components of the GSBPM: {[Needs specification](#), [Design](#), [Build](#), [Collect](#), [Process](#), [Analyze](#), [Disseminate](#), [Evaluate](#)} .
- **Description** (*activity_description*) A brief description of the activity.
- **Participants** (*participants*) A list of participants (persons or organizations) in the activity. This is a repeatable set of elements; each participant can be documented separately.
 - **Name** (*name*) Name of the participating person or organization.
 - **Affiliation** (*affiliation*) Affiliation of the person or organization mentioned in [name](#) .
 - **Role** (*role*) Specific role (participation) of the person or organization mentioned in [name](#) .
- **Resources** (*resources*) A description of the data sources and other resources used to implement the activity.
 - **Name** (*name*) The name of the resource.
 - **Origin** (*origin*) The origin of the resource mentioned in [name](#) .
 - **Characteristics** (*characteristics*) The characteristics of the resource mentioned in [name](#) .
- **Activity outcome** (*outcome*) Description of the main outcome of the activity.

QUALITY STANDARDS

This section lists the specific standards complied with during the execution of this study, and provides the option to formulate a general statement on the quality of the data. Any known quality issue should be reported here. Such issues are better reported by the data producer or curator, not left to the secondary analysts to discover. Transparency in reporting quality issues will increase credibility and reputation of the data provider.

- **Standard compliance** (*compliance_description*) A statement on compliance with standard quality assessment procedures. The list of these standards can be documented in the next element, [standards](#) .
- **Quality standards** (*standards*) An itemized list of quality standards complied with during the execution of the study.
 - **Standard** (*name*) The name of the quality standard, if such a standard was used. Include the date when the standard was published, and the version of the standard with which the study is compliant, and the "URI" attribute includes .
 - **Producer** (*producer*) The producer of the quality standard mentioned in [name](#) .
- **Other quality statement** (*other_quality_statement*) Any additional statement on the quality of the data, entered as free text. This can be independent of any particular quality standard.

DATA APPRAISAL

- **Sampling errors** (*sampling_error_estimates*) Sampling errors are intended to measure how precisely one can estimate a population value from a given sample. For sampling surveys, it is good practice to calculate and publish sampling error. This field is used to provide information on these calculations (not to provide the sampling errors themselves, which should be made available in publications or reports). Information can be provided on which ratios/indicators have been subjected to the calculation of sampling errors, and on the software used for computing the sampling error. Reference to a report or other document where the results can be found can also be provided.
- **Ex-post evaluation** Ex-post evaluations are frequently done within large statistical or research organizations, in particular when a study is intended to be repeated. Such evaluations are recommended by the [Generic Statistical Business Process Model](#) (GSBPM). This section of the schema is used to describe the evaluation procedures and their outcomes.

- **Evaluation type** (*type*) The *type* attribute identifies the type of evaluation with or without the use of a controlled vocabulary.
 - **Evaluation process** (*evaluation_process*) A description of the evaluation process. This may include information on the dates the evaluation was conducted, cost/budget, relevance, institutional or legal arrangements, etc.
 - **Evaluators** (*evaluator*) The evaluator element identifies the person(s) and/or organization(s) involved in the evaluation.
 - **Name** (*name*) The name of the person or organization involved in the evaluation.
 - **Abbreviation** (*abbr*) An abbreviation for the organization mentioned in *name*.
 - **Affiliation** (*affiliation*) The affiliation of the individual or organization mentioned in *name*.
 - **Role** (*role*) The specific role played by the individual or organization mentioned in *name* in the evaluation process.
 - **Completion date** (*completion_date*) The date the ex-post evaluation was completed.
 - **Evaluation outcomes** (*outcomes*) A description of the outcomes of the evaluation. It may include a reference to an evaluation report.
- **Other data appraisal** (*data_appraisal*) This section is used to report any other action taken to assess the reliability of the data, or any observations regarding data quality. Describe here issues such as response variance, interviewer and response bias, question bias, etc. For a population census, this can include information on the main results of a post enumeration survey (a report should be provided in external resources and mentioned here); it can also include relevant comparisons with data from other sources that can be used as benchmarks.

DATA AVAILABILITY

This section describes the access conditions and terms of use for the dataset. This set of elements should be used when the access conditions are well-defined and are unlikely to change. An alternative option is to document the terms of use in the catalog where the data will be published, instead of "freezing" them in a metadata file.

- **Location of dataset** (*access_place*) Name of the location where the data collection is currently stored.
- **URL for location of dataset** (*access_place_url*) The URL of the website of the location where the data collection is currently stored.
- **Archive where study is originally stored** (*original_archive*) Archive from which the data collection was obtained, if any (the originating archive). Note that the schema we propose provides an element **provenance**, which is not part of the DDI, that can be used to document the origin of a dataset.
- **Extent of collection** (*coll_size*) Extent of the collection. This is a summary of the number of physical files that exist in a collection. We will record here the number of files that contain data and note whether the collection contains other machine-readable documentation and/or other supplementary files and information such as data dictionaries, data definition statements, or data collection instruments. This element will rarely be used.
- **Completeness** (*complete*) This item indicates the relationship of the data collected to the amount of data coded and stored in the data collection. Information as to why certain items of collected information were not included in the data file stored by the archive should be provided here. Example: "Because of embargo provisions, data values for some variables have been masked. Users should consult the data definition statements to see which variables are under embargo." This element will rarely be used.
- **Number of files** (*file_quantity*) The total number of physical files associated with a collection. This element will rarely be used.
- **Notes on data availability** (*notes*) Additional information on the dataset availability, not included in one of the elements above.

DEPOSITOR INFORMATION

- **Depositor** (*depositor*) The name of the person (or institution) who provided this study to the archive storing it.

- **Name** (*name*) The name of the depositor. It can be an individual or an organization.
- **Abbreviation** (*abbr*) The official abbreviation of the organization mentioned in **name**.
- **Affiliation** (*affiliation*) The affiliation of the person or organization mentioned in **name**.
- **URL** (*uri*) A URL to the depositor
- **Date of deposit** (*deposit_date*) The date that the study was deposited with the archive that originally received it. The date should be entered in the ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY). The exact date should be provided when possible.

DISTRIBUTOR INFORMATION

- **Distributors** (*distributors*) The organization(s) designated by the author or producer to generate copies of the study output including any necessary editions or revisions.
 - **Name** (*name*) The name of the distributor. It can be an individual or an organization.
 - **Abbreviation** (*abbr*) The official abbreviation of the organization mentioned in **name**.
 - **Affiliation** (*affiliation*) The affiliation of the person or organization mentioned in **name**.
 - **URL** (*uri*) A URL to the ordering service or download facility on a Web site.
- **Date of distribution** (*distribution_date*) The date that the study was made available for distribution/presentation. The date should be entered in the ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY). The exact date should be provided when possible.

DATA ACCESS

- **Access authority** (*contact*) Names and addresses of individuals responsible for the study. Individuals listed as contact persons will be used as resource persons regarding problems or questions raised by users.
 - **Name** (*name*) The name of the person or organization that can be contacted.
 - **Affiliation** (*affiliation*) The affiliation of the person or organization mentioned in **name**.
 - **URL** (*uri*) A URL to the contact mentioned in **name**.
 - **Email** (*email*) An email address for the contact mentioned in **name**.
 -
 - **Confidentiality declaration** (*conf_dec*) This element is used to determine if signing of a confidentiality declaration is needed to access a resource. We may indicate here what *Affidavit of Confidentiality* must be signed before the data can be accessed. Another option is to include this information in the next element (Access conditions). If there is no confidentiality issue, this field can be left blank.
 - **Required** (*required*) The "required" attribute is used to aid machine processing of this element. The default specification is "yes".
 - **Text** (*txt*) A statement on confidentiality and limitations to data use. This statement does not replace a more comprehensive data agreement (see **Access condition**). An example of statement could be the following: "Confidentiality of respondents is guaranteed by Articles N to NN of the National Statistics Act of [date]. Before being granted access to the dataset, all users have to formally agree:
 - To make no copies of any files or portions of files to which s/he is granted access except those authorized by the data depositor.
 - Not to use any technique in an attempt to learn the identity of any person, establishment, or sampling unit not identified on public use data files.
 - To hold in strictest confidence the identification of any establishment or individual that may be inadvertently revealed in any documents or discussion, or analysis.
 - That such inadvertent identification revealed in her/his analysis will be immediately and in confidentiality brought to the attention of the data depositor."

- **Form ID** (*form_id*) Indicates the number or ID of the confidentiality declaration form that the user must fill out.
- **Form URL** (*form_url*) The `"form_url"` element is used to provide a link to an online confidentiality declaration form.
- **Access conditions** (*conditions*) Indicates any additional information that will assist the user in understanding the access and use conditions of the data collection.
- **Citation requirement** (*cit_req*) A citation requirement that indicates the way that the dataset should be referenced when cited in any publication. Providing a citation requirement will guarantee that the data producer gets proper credit, and that results of analysis can be linked to the proper version of the dataset. The data access policy should explicitly mention the obligation to comply with the citation requirement. The citation should include at least the primary investigator, the name and abbreviation of the dataset, the reference year, and the version number. Include also a website where the data or information on the data is made available by the official data depositor. Ideally, the citation requirement will include a DOI (see the [DataCite](#) website for recommendations).
- **Deposit requirement** (*deposit_req*) Information regarding data users' responsibility for informing archives of their use of data through providing citations to the published work or providing copies of the manuscripts.
- **Availability status** (*status*) A statement of the data availability. An archive may need to indicate that a collection is unavailable because it is embargoed for a period of time, because it has been superseded, because a new edition is imminent, etc. This element will rarely be used.
- **Special permissions** (*spec_perm*) This element is used to determine if any special permissions are required to access a resource.
 - **Required** (*required*) The `required` is used to aid machine processing of this element. The default specification is "yes".
 - **Text** (*txt*) A statement on the special permissions required to access the dataset.
 - **Form ID** (*form_id*) The "form_id" indicates the number or ID of the special permissions form that the user must fill out.
 - **Form URL** (*form_url*) The `form_url` is used to provide a link to a special on-line permissions form.
- **Restrictions** (*restrictions*) Any restrictions on access to or use of the collection such as privacy certification or distribution restrictions should be indicated here. These can be restrictions applied by the author, producer, or distributor of the data. This element can for example contain a statement (extracted from the DDI documentation) like: "In preparing the data file(s) for this collection, the National Center for Health Statistics (NCHS) has removed direct identifiers and characteristics that might lead to identification of data subjects. As an additional precaution NCHS requires, under Section 308(d) of the Public Health Service Act (42 U.S.C. 242m), that data collected by NCHS not be used for any purpose other than statistical analysis and reporting. NCHS further requires that analysts not use the data to learn the identity of any persons or establishments and that the director of NCHS be notified if any identities are inadvertently discovered. Users ordering data are expected to adhere to these restrictions."
- **Notes on access and use** (*notes*) Any additional information related to data access that is not contained in the specific metadata elements provided in the section "Data access and use".

DISCLAIMER AND COPYRIGHT

- **Disclaimer** (*disclaimer*) 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: *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.*
- **Copyright** (*copyright*) Any additional information related to data access that is not contained in the specific metadata elements provided in the section "Data access and use".

OTHER INFORMATION

Notes on the study (*study_notes*) This element can be used to provide additional information on the study which cannot be accommodated in the specific metadata elements of the schema, in the form of a free text field.

CONTACTS

- **Contacts** (*contact*) Users of the data may need further clarification and information on the terms of use and conditions to access the data. This set of elements is used to identify the contact persons who can be used as resource persons regarding problems or questions raised by the user community.
 - **Name** (*name*) Name of the person. Note that in some cases, it might be better to provide a title/function than the actual name of the person. Keep in mind that people do not stay forever in their position.
 - **Affiliation** (*affiliation*) Affiliation of the person.
 - **Email** (*email*) The **email** element is used to indicate an email address for the contact individual mentioned in **name**. Ideally, a generic email address should be provided. It is easy to configure a mail server in such a way that all messages sent to the generic email address would be automatically forwarded to some staff members.
 - **URL** (*uri*) URI for the person; it can be the URL of the organization the person belongs to.

Data files

The **Data files** section of the DDI Metadata Editor is where **structural metadata** will be captured, including file-level and variable-level information.

Importing data files

Typically, the core of structural metadata will be automatically generated by importing the data files into the Metadata Editor. When no data file is available to be imported, this section will be left empty (in which case your metadata will be limited to descriptive and reference metadata), or a data dictionary can be entered manually, which may be a tedious process prone to errors.

To import data files, select *Data files* in the navigation tree, and click **IMPORT** in the *Data files* page. Select the file(s) to be imported.

The following formats are currently supported: Stata (.dta), SPSS (.sav), and CSV.

The Metadata Editor runs Python in the background (using the Pandas library). It will read the data files, and convert them to CSV format for internal use. The application will make a best guess to assign the type of each variable (continuous or categorical), extract the variable names and labels, extract the value labels, and generate (unweighted) summary statistics for all variables. The import process will result in a data dictionary for each data file.

The imported data files will now be listed in the navigation tree. Summary information on the imported data files is displayed in the *Data files* page.

| File# | File name | Variables | Cases | Modified | Data |
|-------|--|-----------|-------|------------|-------------------|
| F1 | WLD_2023_SYNTH-SVY-HLD-EN_v01.M WLD_2023_SYNTH-SVY-HLD-EN_v01.M.csv 1.57 MB | 49 | 8000 | 2025-02-12 | CLEAR DATA EXPORT |
| F2 | WLD_2023_SYNTH-SVY-IND-EN_v01.M WLD_2023_SYNTH-SVY-IND-EN_v01.M.csv 2.42 MB | 27 | 32396 | 2025-02-12 | CLEAR DATA EXPORT |

IMPORTANT: The CSV version of the data will be stored on the server that hosts the Metadata Editor. **All collaborators on your project will be able to see and download the data.** If the data are confidential and sensitive, they may be removed from the server at any time. See *Removing data files from server* below.

Reordering data files

If the order in which the files appear in the *Data files* page needs to be modified, use the handle icons to move the files up or down (drag and drop).

Exporting data files

At any time, you may export the imported data files, to any of the supported format (currently, Stata, SPSS, and CSV). Typically, you will want to export the data files AFTER editing the metadata (variable and value labels) to ensure full consistency between the data and the metadata.

If you made any change to the metadata (editing variable and/or value labels), exporting the data files before sharing them with the metadata will ensure that the data and the metadata are fully consistent. But you should **ALWAYS** preserve a copy of the original datasets that you imported into the Metadata Editor.

Deleting data files

Data files can be **deleted** from the project. When you delete a data file, all related metadata will be removed, and the data for the deleted file will also be removed from the server.

Removing data files from server

Imported data files are by default stored in the web server where the Metadata Editor has been installed. If you share a project with other users, you may grant them access to the data. If your files cannot be shared, you have the option to remove them from the server after the metadata extraction has been completed.

If you want to remove the data from the server, use the **Clear data** option provided in the *Data files* page. Note that after removing the data from the server, you will not be able to update the summary statistics anymore. If you plan to edit the value labels, or apply weights to generate weighted summary statistics, do that BEFORE you clear the data files from the server. And if you want to ensure that no collaborator has access to the data, clear the data files BEFORE sharing the project or adding it to a collection.

If you are not allowed to upload your data files to the server, one option is to extract one observation from each data file, anonymize it, and importing these single-observation files into the Metadata Editor. By doing this, the data dictionary will be generated as if you had imported the full dataset. But the summary statistics will be invalid. You will have to remove all summary statistics for all variables. See the section **Variables** below.

Replacing data files

You can replace a data file by selecting the option **Replace file** in the option menu for a data file in the *Data files* page. You can only replace a data file with a data file that has the exact same structure (same variables, in the same order). The application will replace the data, but leave the metadata already entered in the Metadata Editor untouched. Replacing a data file is thus a solution to "refresh the data" (and summary statistics) without losing updated metadata.

Data files

2 Files

[IMPORT FILES](#)

| <input type="checkbox"/> | <input type="checkbox"/> | File# | File name | Variables | Cases | Modified | Data ? | CLEAR DATA | EXPORT | ... |
|--------------------------|--------------------------|-------|--|-----------|-------|------------|----------------------------|----------------------------|------------------------|--|
| <input type="checkbox"/> | <input type="checkbox"/> | F1 | WLD_2023_SYNTH-SVY-HLD-EN_v01_M WLD_2023_SYNTH-SVY-HLD-EN_v01_M.csv 1.57 MB | 49 | 8000 | 2025-02-12 | CLEAR DATA | EXPORT | ... | Refresh summary statistics |
| <input type="checkbox"/> | <input type="checkbox"/> | F2 | WLD_2023_SYNTH-SVY-IND-EN_v01_M WLD_2023_SYNTH-SVY-IND-EN_v01_M.csv 2.42 MB | 27 | 32396 | 2025-02-12 | CLEAR DATA | EXPORT | ... | Replace file |

Refreshing summary statistics

Some actions on data files (modifying *Categories description* or applying weighting coefficients) will require that the summary statistics be updated. The Metadata Editor does not automatically update statistics each time a change is made. Instead, it will add visual clues (red icons) that the data does not fully correspond to the metadata anymore, and call for a refreshing of summary statistics. The *Refresh summary statistics* in the option menu next to the data file description (in the *Data files* page) serves this purpose. Another option to refresh statistics is to click on the [Refresh stats](#) icon in the *Variable list* frame of the Variable page (see *Refresh statistics* below).

File description

Clicking on a file name in the navigation tree will open the *File description* page where the information on the data file can be entered. Some information like the file name, number of variables and observations, are automatically generated when data files are imported.

The screenshot shows the Metadata Editor interface. On the left, there's a sidebar with navigation links: Home, Document description, Study description, Tags, Data files (which is expanded), Variables, Data, and another Variables link. Under 'Data files', the file 'WLD_2023_SYNTH-SVY-HLD-EN_v01_M' is selected and highlighted with a red box. The main right panel is titled 'Popstan Synthetic Household Survey 2023' and shows a form for the selected data file. The form includes fields for 'File name' (set to 'WLD_2023_SYNTH-SVY-HLD-EN_v01_M'), 'Description' (containing the text 'Contains all variables at the household level'), 'Producer' (set to 'World Bank Development Data Group'), and sections for 'Data checks' and 'Missing data'.

- **Description** (*description*) The `file_id` and `file_name` elements provide limited information on the content of the file. The `description` element is used to provide a more detailed description of the file content. This description should clearly distinguish collected variables and derived variables. It is also useful to indicate the availability in the data file of some particular variables such as the weighting coefficients. If the file contains derived variables, it is good practice to refer to the computer program that generated it. Information about the data file(s) that comprises a collection.
- **Producer** (*producer*) The name of the agency that produced the data file. Most data files will have been produced by the survey primary investigator. In some cases however, auxiliary or derived files from other producers may be released with a data set. This may for example be a file containing derived variables generated by a researcher.

- **Data checks** (*data_checks*) Use this element if needed to provide information about the types of checks and operations that have been performed on the data file to make sure that the data are as correct as possible, e.g. consistency checking, wildcode checking, etc. Note that the information included here should be specific to the data file. Information about data processing checks that have been carried out on the data collection (study) as a whole should be provided in the **Data editing** element at the study level. You may also provide here a reference to an external resource that contains the specifications for the data processing checks (that same information may be provided also in the **Data Editing** field in the **Study Description** section).
- **Missing data** (*missing_data*) A description of missing data (number of missing cases, cause of missing values, etc.)
- **Version** (*version*) The version of the data file. A data file may undergo various changes and modifications. File specific versions can be tracked in this element. This field will in most cases be left empty.
- **Notes** (*notes*) This field aims to provide information on the specific data file not covered elsewhere.

Variables

The variable description section of the DDI Codebook standard provides you with the possibility to create a very rich data dictionary. While some of the variable-level metadata will be imported from the data files, the DDI Codebook contains many elements that can enrich the data dictionary. This additional information is very useful both for discoverability and for increased usability of the data.

After importing a data file in the Metadata Editor, you can access information on the variables it contains by clicking on *Variables* below the file name in the navigation tree. This will open the *Variables* page, which contains four frames:

- **Variables.** Provides the list of variables in the data file, with their name, label, and status icons.
- **Variable categories.** Describes the categories (value labels) for a categorical variable selected in the list of variables.
- **Variable description.** Contains information on the status and type of the variable selected in the variable list, on its range, and on missing values.
- **Statistics, weights, documentation, and JSON.** Select the summary statistics to be included in the metadata, apply sampling weights if relevant, edit the documentation of the selected variable(s), and view the metadata in JSON format.

These four frames are described in detail below.

- **Variable list**

The variable list displays the list of variables in the selected data file, with their name and label. A search bar is provided to locate variables of interest (keyword search on variable name and label).

| V1 | hid | Unique household identifier | |
|----|--------|-----------------------------|--|
| V2 | geo1 | Geographic area - Admin 1 | |
| V3 | geo2 | Geographic area - Admin 2 | |
| V4 | ea | Enumeration area | |
| V5 | urbrur | Residence (urban/rural) | |

- The variable labels can be edited (double click on a label to edit it). 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 different label. File formats like Stata or SPSS often contain variable labels. Variable labels can also be found in data dictionaries in software applications like Survey Solutions or CsPro. Avoid using the question itself as a label (specific elements are available to capture the literal question text; see below). Think of a label as what you would want to see in a tabulation

of the variables. Keep in mind that software applications like Stata and others impose a limit to the number of characters in a label (often, 80).

- The variable names cannot be edited.
- A set of icons provides information on the status of each variable, as follows:



Variable is of type character



Variable is of type numeric



Variable is categorical



Variable is a sample weight



Sample weight has been applied



Summary statistics must be refreshed

The icons shown on top of the variable list perform the following actions on variables:



Refresh statistics



Change case (variable names and/or labels)



Spread metadata (from one file to another)



Delete the selected variables from the data file

- **Refresh statistics:** Instructs the Metadata Editor to re-read the data files and update the summary statistics for all variables. You will want to refresh the summary statistics after performing actions like applying sample weights or modifying the value labels (categories) of some variables. Note that this will not be possible if you have cleared the data from the server.
- **Change case:** This allows you to change the case of the variable names and/or variable labels (changing them to uppercase, lowercase, or sentence case). Note that changing variable names is not always a good idea; some software like Stata or R are case sensitive; changing the case of a variable name is equivalent to changing the variable name, which means that some scripts written to run on the original data will not run anymore on data with renamed variables.

Change case

Type

Title Case

Name
 Label

[APPLY](#)

[CANCEL](#)

- **Spread metadata:** In many micro-datasets, a small number of variables will be common to all data files. For example, in a sample household survey, the household identifier, and the variables that indicate the geographic location of the household, may be repeated in all files. In such case, you can document the variables in one data file, select these common variables, and "spread metadata" to automatically apply the metadata to variables with the same name in other data files.
- **Delete selection:** The Metadata Editor is not a data editor; data files cannot be modified, with the exception that variables can be deleted from a data file. This option is provided to drop variables like temporary variables or direct identifiers that may have been accidentally left in the data files. To drop variables, select them in the variable list (using the Shift and Ctrl key as useful to select multiple variables), then click on the **Delete selection** icon.
- **Variable categories** (value labels)

When the data are imported, the Metadata Editor makes a best guess about their type. If a variable is identified as categorical, the *Variable categories* will be automatically populated with codes and value labels. These codes and labels can be edited in the *Variable categories* frame. Codes can also be added if necessary.

The *Variable categories* frame contains the following icon toolbar:



Create categories from statistics



Clear all categories



Copy/paste (replace/append) categories

If a categorical variable was not identified as such when the data were imported, the Metadata Editor provides an option to automatically *Create categories*. This option is accessed by clicking on the **Create categories** icon at the top of the frame. The application will then extract a list of codes found for the variables in the data, and pre-populate the list of categories with the codes. The labels corresponding to each category can then be added by the data curator. An option (**Clear all** icon) is also provided to remove all content from the list of categories.

| Value | Label |
|-------|-----------------|
| 1 | Earth |
| 2 | Cement/concrete |
| 3 | Tile |
| 4 | Stone |
| 5 | Wood |
| 6 | Other |

A category can be removed from the list by clicking on the trash icon.

An option is also provided to copy/paste content in the list of value labels. This allows for example to copy code lists from an application like MS-Excel and to paste them in the *Variable categories* grid. The pasted information can either *replace* existing content of the grid (if any), or be *appended* to it.

- **Value**

- Copy**
- Paste (Replace)**
- Paste (Append)**
- Undo paste**

- **Variable information**

The **Variable information** frame provides options to tag a variable as being a sample weight, to edit the type of the variable, the range, and the format of the data it contains, and to inform the application of values to be treated as missing values.

Variable information

Is weight

Interval type

Discrete

Type

Numeric

| Min | Max | Decimal points |
|-----|-----|----------------|
| 1.0 | 6.0 | |

Missing

+ ADD ROW

- **Is weight** The Is weight toggle is used to tag variables that are sample weights, if any.

| Variables 49 | Search | Variable categories |
|---------------|---|--------------------------|
| V46 quint_urb | Per capita expenditure quintiles, urban | <input type="checkbox"/> |
| V47 quint_rur | Per capita expenditure quintiles, rural | <input type="checkbox"/> |
| V48 hhweight | Household weight | <input type="checkbox"/> |
| V49 popweight | Population weight | <input type="checkbox"/> |

- **Interval type** For numeric variables, the *interval type* can be set to *Continuous* or *Discrete*. This will impact how the data are documented and what summary statistics are produced (frequencies will not be calculated for continuous variables; variable categories will only be created for discrete variables). When data are imported, the application makes a best guess about the interval type. the data curator should browse through all variables to ensure the interval type has been properly set.
- **Type** This option informs the system of the format type of the variable, with options *numeric*, *character* (string variables), and *fixed*.
- **Min, Max, and Decimal point** A range of valid values can be entered in this element; values outside the range will trigger a validation error.
- **Missing** The proper use of data requires proper treatment of missing values. The application (and data users) needs to know what values in the data files are to be considered as *missing*. When data are imported from Stata or SPSS, if the system missing values were used in the origial files, the missing values will be shown as . But other values may

represent missing values. In some survey datasets, the code 99 for example may be used to represent "unknown or unreported age. When such values (other than system-missing) are used, they must be indicated as representing missing data, otherwise the summary statistics will be incorrect (e.g., 99 would be treated as a valid age even though it means "unreported"). One or multiple values representing *missing* can be entered for the variable.

- **Variable documentation**

The last frame of the Variable page is one where most of the additional metadata related to variables will be entered.

The screenshot shows the 'DOCUMENTATION' tab selected in the top navigation bar. On the left, there's a sidebar with a 'Settings' section containing three checkboxes: 'Universe' (checked), 'Variable definition' (checked), and 'Concepts'. To the right, there are two main sections: 'Universe' and 'Variable definition', each with a large empty text area for documentation.

Header bar The header bar of the frame shows the name and label of the variable selected in the Variable list. It also shows four icons, which allow you browse variables without having to select them in the Variable list.



Navigate through variables (first, previous, next, last)

Statistics

The STATISTICS tab shows the summary statistics generated by the application when the data are imported (or when the statistics are refreshed). If you have applied weighting coefficients (see **Weights** below), the summary statistics will include both weighted and unweighted statistics. Otherwise, the statistics are unweighted. Keep in mind to properly document values that should be interpreted as missing values, to ensure that statistics like means and standard deviations are valid (see **Missing** above).

The option is offered in this tab to select the summary statistics to be included in the metadata. Uncheck the statistics that you do not want to store in the metadata. For example, you do not want to keep statistics like mean or standard deviation for categorical variables. Only keep statistics that are meaningful. Note that you may select a group of variables in the *Variable list* and apply a selection to all selected variables at once.

Note that summary statistics that will be stored in the metadata are statistics for each variable taken independently. No cross tabulation is possible.

Variables 49

V7 statocc Status of occupancy of the dwelling

V8 rooms Number of rooms in dwelling

statooc - Status of occupancy of the dwelling

STATISTICS WEIGHTS • DOCUMENTATION JSON

Weighted statistics
 Frequencies
 List missings

Summary statistics:

- Valid
- Min
- Max
- Mean
- Weighted mean
- StdDev
- Weighted StdDev

Frequencies

| Value | Label | Cases | Weighted |
|-------|-------------------|-------|---------------|
| 1 | Owned | 6187 | 1933484 77.3% |
| 2 | Rented | 1021 | 316790 12.7% |
| 3 | Occupied for free | 792 | 251479 10.1% |

Summary statistics

| | |
|------------------|---------|
| Valid | 8000.00 |
| Min | 1.00 |
| Max | 3.00 |
| Mean | 1.33 |
| stdev | 0.65 |
| Mean (weighted) | 1.33 |
| stdev (weighted) | 0.65 |

Weights

If the dataset is a sample survey for which sample weights have been calculated, you should tag all variables that are sample weights (see **Weights** above). Variables that have been tagged as being weights can then be used to generate weighted statistics in tab STATISTICS. To do that, select the variables for which you want to generate weighted statistics (this should not include variables like unique identifiers or the weighting variables themselves). Use the Shift or Ctrl keys to select multiple variables in the list. Then click on *select weight variable*. The list of variables in the data file that have been tagged as weights will be displayed. Select the one to be used as weight, then close the selection popup menu. A green dot will appear next to the tab name, to indicate that a weight has been applied to the variable. The variable used as weight will be shown in the tab. An option to *remove* or *change* the weight is provided. A red icon will appear in the variable list to indicate that a change has been applied to the variable that requires refreshing the summary statistics. After you refresh the summary statistics, the STATISTICS tab will display the weighted values.

Variables 49

V6 hsize Household size

V7 statocc Status of occupancy of the dwelling

V8 rooms Number of rooms in dwelling

statooc - Status of occupancy of the dwelling

STATISTICS WEIGHTS • DOCUMENTATION JSON

| Name | Label |
|----------|------------------|
| hhweight | Household weight |

Documentation

The DOCUMENTATION tab is where you will enter variable-level metadata that will enrich the data dictionary. In this page, you will find a form allowing you to enter information on each variable like literal questions (for surveys), universe, concepts, interviewers' instructions, methodology, derivation and imputation, and more.

A **Settings** menu allows you to select the metadata elements to be used to document variables. This selection will reduce the length of the form used to enter the metadata. The only purpose of this selection is to make the metadata

entry page shorter and more convenient to use, by focusing only on the metadata elements for which content may be provided.

The screenshot shows two side-by-side views of the Metadata Editor interface for a variable named 'hid - Unique household identifier'.

Left View (DOC Tab):

- Header: hid - Unique household identifier
- Top navigation: STATISTICS, WEIGHTS, **DOC**
- Left sidebar: A red circle highlights the three-line menu icon.
- Content sections:
 - Universe**: Placeholder box.
 - Variable definition**: Placeholder box.
 - Concepts**: Sub-sections for **Title** and **Notes**.

Right View (DOCUMENTATION Tab):

- Header: hid - Unique household identifier
- Top navigation: STATISTICS, WEIGHTS, **DOCUMENTATION**, JSON
- Left sidebar: Placeholder box.
- Content sections:
 - Settings**: Placeholder box.
 - Universe**: Universe
 - Variable definition**: Variable definition
 - Concepts**: Concepts
 - Standard categories**: Standard categories
 - Security**: Security
 - Notes on variable**: Notes on variable
 - Questions and instructions**:
 - Respondent
 - Pre-question text
 - Literal question
 - Post question text
 - Forward skip
 - Backward skip
 - Interviewer instructions
 - Imputation and derivation**:
 - Coder instructions
 - Imputation
 - Derivation
 - Concepts**: Sub-section for **Title**.
 - Standard categories**: Sub-section for **Name**.

A large yellow arrow points from the left view to the right view, indicating the transition between the two tabs.

Note that you may select multiple variables in the *Variable list* (using the Shift or Ctrl key) and paste content in one of these metadata elements. It will then be pasted in the metadata for all selected variables. The existing metadata for other fields will be left unchanged.

The screenshot shows the 'DOCUMENTATION' tab selected in the Metadata Editor. The main content area is empty, indicating no documentation has been provided for this variable.

JSON

This last tab is only intended to display a JSON version of the variable metadata.

```
{
  "uid": "25354",
  "sid": "4709",
  "fid": "F1",
  "vid": "V7",
  "name": "statocc",
  "label": "Status of occupancy of the dwelling",
  "sort_order": "0",
  "var_intrvl": "discrete",
  "loc_width": 17,
  "var_valrng": {
    "range": {
      "UNITS": "REAL",
      "."": 2000
    }
  }
}
```

View data

The Metadata Editor allows you to view the content of the data files, by clicking on "Data" under the data file name in the navigation tree. You can view and export the data from this page, but you cannot edit the data. The Metadata Editor is not a data editor; changes to the data files must be made outside the Metadata Editor.

| # | hid | geo1 | geo2 | ea | urbrur | hhsiz | statocc | rooms | bedrooms | floor | walls | roof | water | piped_water | toilet | flush_toilet | electricity | cook_fuel | phone | cell | car | bicycle | m |
|---|-------------|------|------|-------|--------|-------|---------|-------|----------|-------|-------|------|-------|-------------|--------|--------------|-------------|-----------|-------|------|-----|---------|---|
| 1 | 00a191396a2 | 1 | 11 | 11066 | 2 | 1 | 1 | 1 | 1 | 2 | 3 | 5 | 61 | 0 | 21 | 2 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0111c7fa67c | 1 | 11 | 11066 | 2 | 1 | 2 | 1 | 0 | 2 | 6 | 5 | 11 | 1 | 22 | 2 | 1 | 3 | 0 | 1 | 0 | 0 | 0 |
| 3 | 01410d9c60d | 1 | 11 | 11087 | 2 | 2 | 1 | 4 | 1 | 2 | 3 | 1 | 11 | 1 | 11 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 4 | 015ff82b55c | 1 | 11 | 11094 | 2 | 2 | 1 | 1 | 1 | 2 | 3 | 5 | 11 | 1 | 21 | 2 | 1 | 2 | 0 | 1 | 0 | 0 | 0 |
| 5 | 01c39d3df0d | 1 | 11 | 11094 | 2 | 1 | 2 | 3 | 2 | 6 | 3 | 1 | 31 | 0 | 13 | 1 | 1 | 4 | 1 | 1 | 1 | 0 | 0 |
| 6 | 027fe2e901f | 1 | 11 | 11003 | 2 | 2 | 1 | 3 | 2 | 2 | 3 | 1 | 11 | 1 | 11 | 1 | 1 | 2 | 1 | 0 | 0 | 0 | 0 |

Metadata elements for variable description

The DDI Codebook metadata standard provides multiple elements to document variables contained in a micro-dataset. There is much value in documenting variables:

- it makes the data **usable** by providing users with a detailed data dictionary;
- it makes the data more **discoverable** as all keywords included in the description of variables are indexed in data catalogs;
- it allows users to assess the comparability of data across sources;
- it enables the development of question banks; and
- it adds transparency and credibility to the data especially when derived or imputed variables are documented. All possible effort should thus be made to generate and publish detailed variable-level documentation.

A micro-dataset can contain many variables. Some survey datasets include hundreds or even thousands of variables. Documenting variables can thus be a tedious process. The use of a specialized DDI metadata editor can make this process considerably more efficient. Much of the variable-level metadata can indeed be automatically extracted from the electronic data files. Data files in Stata, SPSS or other common formats include variable names, variable and value labels, and in some cases notes that can be extracted. And the variable-level summary statistics that are part of the metadata can be generated from the data files. Further, software applications used for capturing data like [Survey Solutions](#) from the World Bank or [CsPro](#) from the US Census Bureau can export variable metadata, including the variable names, the variable and value labels, and possibly the formulation of questions and the interviewers instructions when the software is used for conducting computer assisted personal interviews (CAPI). Survey Solutions and CsPro can export metadata in multiple formats, including the DDI Codebook. Multiple options exist to make the documentation of variables efficient. As much as possible, tedious manual curation of variable-level information should be avoided.

- **Universe** (`var_universe`) The universe at the variable level defines the population the question applied to. It reflects skip patterns in a questionnaire. This information can typically be copy/pasted from the survey questionnaire. Try to be as specific as possible. This information is critical for the analyst, as it explains why missing values may be found in a variable. In the example below (from the Malawi MICS 2006 survey questionnaire), the universe for questions ED1 to ED2 will be "*Household members age 5 and above*", and the universe for Question ED3 will be "*Household members age 5 and above who ever attended school or pre-school*".
- **Variable definition** This element provides a space to describe the variable in detail. Not all variables require a definition.
- **Concepts** (`var_concept`) The general subject to which the parent element may be seen as pertaining. This element serves the same purpose as the keywords and topic classification elements, but at the variable description level.
- **Title** (***) The name (label) of the concept.

- **Vocabulary** (***) The controlled vocabulary, if any, from which the concept 'title' was taken.
- **URL** (**) The location for the controlled vocabulary mentioned in 'vocab'.
- **Standard categories** This element is used to indicate that the codes used for a categorical variable are from a standard international or other classification, like COICOP, ISIC, ISO country codes, etc.
- **Security** (*var_security*) This element is used to provide information regarding levels of access, e.g., public, subscriber, need to know.
- **Notes on variable** (*var_notes*) This element is provided to record any additional or auxiliary information related to the specific variable.
- **Respondent** (*var_respunit*) Provides information regarding who provided the information contained within the variable, e.g., head of household, respondent, proxy, interviewer.
- **Pre-question text** (*var_qstn_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.
- **Literal question** (*var_qstn_qstnlit*) The literal question is the full text of the questionnaire as the enumerator is expected to ask it when conducting the interview. This does not apply to all variables (it does not apply to derived variables).
- **Post question text** (*var_qstn_postqtxt*) The post-question texts are instructions provided to the interviewers, printed in the questionnaire after the literal question. Post-question can be used to enter information on skips provided in the questionnaire. This does not apply to all variables. Do not confuse this with instructions provided in the interviewer's manual.

With the previous three elements, one should be able to understand how the question was formulated in a questionnaire. In the example below (extracted from the UNICEF [Malawi 2006 MICS](#) survey questionnaire), we find:

- a pre-question: "*Ask this question ONLY ONCE for each mother/caretaker (even if she has more children).*"
- a literal question: "*Sometimes children have severe illnesses and should be taken immediately to a health facility. What types of symptoms would cause you to take your child to a health facility right away?*"
- a post-question: "*Keep asking for more signs or symptoms until the mother/caretaker cannot recall any additional symptoms. Circle all symptoms mentioned. DO NOT PROMPT WITH ANY SUGGESTIONS*"
- **Forward skip** (*var_forward*) Contains a reference to the IDs of possible following questions. This can be used to document forward skip instructions.
- **Backward skip** (*var_backward*) Contains a reference to IDs of possible preceding questions. This can be used to document backward skip instructions.
- **Interviewer instructions** (*var_qstn_ivuinstr*) Specific instructions to the individual conducting an interview. The content will typically be entered by copy/pasting instructions in the interviewer's manual (or in the CAPI application). In cases where the same instructions relate to multiple variables, repeat the same information in the metadata for all these variables. NOTE: In earlier version of the documentation, due to a typo, the element was named *var_qstn_ivulnstr*.
- **Coder instructions** (*var_codinstr*) The coder instructions for the variable. These are any special instructions to those who converted information from one form to another (e.g., textual to numeric) for a particular variable.
- **Imputation** (*var_imputation*) Imputation is the process of estimating values for variables when a value is missing. The element is used to describe the procedure used to impute values when missing.
- **Derivation** (*var_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. The *var_derivation* element is used to provide a

brief description of this process. As full transparency in derivation processes is critical to build trust and ensure replicability or reproducibility, the information captured in this element will often not be sufficient. A reference to a document and/or computer program can in such case be provided in this element, and the document/scripts provided as external resources. For example, a variable "TOT_EXP" containing the annualized total household expenditure obtained from a household budget survey may be the result of a complex process of aggregation, de-seasonalization, and more. In such case, the information provided in the `var_derivation` element could be: "TOT_EXP was obtained by aggregating expenditure data on all goods and services, available in sections 4 to 6 of the household questionnaire. It contains imputed rental values for owner-occupied dwellings. The values have been deflated by a regional price deflator available in variable REG_DEF. All values are in local currency. Outliers have been fixed by imputation. Details on the calculations are available in Appendix 2 of the Report on Data Processing, and in the Stata program [generate_hh_exp_total.do]."

Variable groupings

Variables in a dataset are automatically grouped by data file. But for the convenience of users, they can also be grouped based on other criteria, e.g., by theme. These groupings will not alter the data structure in any way; variable groupings are strictly virtual. In this virtual grouping system:

- A variable can belong to more than a group
- Not all variables need to belong to a group
- A group can include variables from multiple files

Variable groups can be very useful for data discoverability, if variables are grouped by theme and if variable groups are defined. For example, a group of variables containing variables *age in month*, *sex*, *weight in grams*, and *height in cm* could be created and named "Anthropometric measures*", with a description informing users that these variables could be used to generate indicators like percentage of children stunting or wasting. This metadata, when indexed in data catalog, will improve the discoverability of the data by adding useful semantic content to the description of variables.

To group variables:

- Create a group (or subgroup) and give it a name.
- Select the variables (from any of the imported data files) to be included.
- Add a description of the group.

The screenshot shows the Metadata Editor interface for the Popstan Synthetic Household Survey 2023. The left sidebar has a red box around the 'Variable groups' item under 'Data files'. The main area displays the 'Variable Groups' section. A tree view on the left shows 'Variable Groups' with 'Demographics' selected. To the right, there are input fields: 'Group ID' (VG1), 'Group type' (Pragmatic), 'Label' (Demographics), and 'Description' (empty). The top right corner shows 'SAVE' and user information.

Variable selection

↳ **idno - PERSON IDENTIFICATION NUMBER**

| | |
|---|----|
| <input type="checkbox"/> relation - Relationship to the head of household | F2 |
| <input checked="" type="checkbox"/> sex - Sex | F2 |
| <input checked="" type="checkbox"/> age - Age in years | F2 |
| <input checked="" type="checkbox"/> age_month - Age in months | F2 |
| <input checked="" type="checkbox"/> marstat - Marital status | F2 |
| <input checked="" type="checkbox"/> religion - Religion | F2 |
| <input type="checkbox"/> school_attend - School attendance | F2 |
| <input type="checkbox"/> educ_attain - Education attainment | F2 |
| <input type="checkbox"/> yrs_school - Years of schooling | F2 |

CLOSE

Variables Select variables

| FID | Name | Label | |
|-----|-----------|----------------|--|
| F2 | sex | Sex | |
| F2 | age | Age in years | |
| F2 | age_month | Age in months | |
| F2 | marstat | Marital status | |
| F2 | religion | Religion | |

Provenance

The Provenance container is used to document how, from where, and when the dataset was acquired. It is used to ensure traceability. See section "Documenting - General instructions" for more information.

External resources

External resources are all materials (and links) that relate to the dataset. This includes documents and reports, scripts, photos and videos, data files, and any other resource available in digital format. These materials and links are added to the documentation of a dataset in the External resources container. Click on **External resources** in the navigation tree, then on CREATE RESOURCE. Enter the relevant information on the resource (at least a title), then provide either a filename (the file will then be uploaded on the server that hosts the Metadata Editor) or a URL to the resource.

External resources that have already been created for another project can also be imported. To do that, they must first be exported as JSON or RDF from the other project. The click on IMPORT in the External resources page, and select the file.

External resources will be part of the project ZIP package (when the ZIP package is generated - See the main menu).

See also chapter **General instructions**.

Importing metadata

In the main menu, an option is provided to **Import project metadata**. This allows you to import the metadata exported from another project (DDI/XML or JSON file exported from another project of type *microdata*). This option will be particularly useful when you have dataset that belong to a same series that share common study description or data dictionary. For example, if a country conducts a monthly labor force survey using the same questionnaire over time, most of the metadata will be common to all instances of the survey. Only a few fields will need to be edited (temporal coverage, summary statistics, and a few others). In such case, the most effective way to document a new instance of the survey is to load the dataset for the latest survey, apply metadata from the previous instance, and edit the few metadata elements that need to be updated.

This option will also be used to import metadata compliant with the DDI codebook that may have been generated using another tool. For example, if a survey was conducted by CAPI interview (using tablets) using the Survey Solutions CAPI application, a DDI containing the literal questions, universe, interviewers instructions (stored in the CAPI questionnaire) can be generated. This information can be imported in the Metadata Editor, avoiding the tedious process of copy/pasting this information.

The screenshot shows the Metadata Editor interface with the following layout:

- Top Bar:** Includes the 'Metadata Editor' logo, 'About', 'English', and 'John Doe' dropdown.
- Left Sidebar:** Buttons for 'Required' (checked), 'Recommended', and 'Empty'.
- Project Section:**
 - Export package (ZIP)**
 - Export DDI Codebook**
 - Export JSON**
 - Publish to NADA**
 - PDF documentation**
 - Change log**
- Metadata Section:**
 - Apply default values from template** (checkbox checked)
 - Import project metadata**
 - Import external resources**
- External resources Section:**
 - Export RDF/XML**
 - Export RDF/JSON**
- Right Side (Buttons):** 'SAVE' and a three-dot menu icon.

The components of the metadata to be imported can be selected.

Import project metadata

Choose DDI/XML or a JSON file

No file chosen

Options

- Document description
- Study description
- File description
- Variable information
- Variable documentation
- Variable categories
- Variable questions
- Variable weights
- Variable groups

Spreading metadata

Spreading metadata consists of copying the metadata for one or multiple variables selected in the variable list, then clicking on the **Spread metadata** icon. The metadata entered for the selected variables will automatically be applied to the variables that have the same name, in the other data files. The metadata fields to be spread can be selected.



Popstan Synthetic Household Survey 2023

| Variables 49 | | Search | |
|--------------|------|-----------------------------|--|
| V1 | hid | Unique household identifier | |
| V2 | geo1 | Geographic area - Admin 1 | |
| V3 | geo2 | Geographic area - Admin 2 | |
| V4 | ea | Enumeration area | |
| V5 | | | |

Spread metadata [1 matches]

CLOSE

| | FID | Dataset | Variable | Type | Type match |
|--------------------------|-----|---------------------------------|----------|------|------------|
| <input type="checkbox"/> | F2 | WLD_2023_SYNTH-SVY-IND-EN_v01_M | hid | | true |

Spread metadata

- Variable information - Labels
- Variable documentation - Texts, notes, universe, etc.
- Categories
- Question texts - Pre, post, literal, etc.
- Weights

Spread metadata **Cancel**

Exporting metadata

The project options menu allows you to export the metadata into different formats:

The screenshot shows the 'Project' and 'Metadata' sections of the project options menu. Several items are highlighted with red boxes or circles:

- Project section:**
 - Export package (ZIP)** (highlighted with a red box)
 - Export DDI Codebook**
 - Export JSON**
 - Publish to NADA**
 - PDF documentation** (highlighted with a red box)
 - Change log**
- Metadata section:**
 - Apply default values from template**
 - Import project metadata**
 - Import external resources**
 - External resources** section:
 - Export RDF/XML** (highlighted with a red box)
 - Export RDF/JSON** (highlighted with a red box)
- Top right corner:** A red circle highlights the three-dot menu icon.

- **Export package (ZIP):** A project contains the metadata you enter, the data files you may have imported (for microdata), the project thumbnail, and possibly external resources of different types. The Metadata Editor stores this information in a database and on the webserver that hosts the Editor. You may export a *package* that contains all materials related to the project. This package will consist of a ZIP file containing all files (including the metadata you

entered, even if you did not export them). The ZIP file can be archived, or shared. A package can be imported in the Metadata Editor.

- **Export DDI Codebook:** This option will generate a DDI Codebook 2.5 file (XML format) containing the metadata, **not including** the metadata related to external resources.
- **Export JSON:** This option will generate a metadata file in JSON format. The file will contain the core metadata and other components based on the options you will select: including external resources, including metadata elements marked as *private* in the metadata template, and including the administrative metadata.
- **PDF documentation:** This option will generate a formatted PDF document containing the project metadata, including metadata on external resources.
- **Export RDF/XML:** This option will export the metadata related to the external resources as an Resource Description Framework (RDF) / XML file.
- **Export RDF/JSON:** This option will export the metadata related to the external resources as an Resource Description Framework (RDF) / JSON file.

Change log

This option will open a page that shows all actions taken on the project, with identification of who took the action (change log). This option can be used to undo some actions. See chapter **General instructions** for more information.

Publish to NADA

This option in the Project home page options menu allows you to publish your metadata (and related materials, optionally including data) to a NADA catalog. See chapter **Publish to NADA** for more information.

Documenting a publication or report

Note: The schema we describe here is the schema used to document publications and reports to be catalogued, not the schema used to document resources that may be attached as external resources to data of any type, for which the *External resource* metadata schema is used (see [Documenting Data: General instructions](#)).

Librarians have developed various standards to describe and catalog documents. The [MARC21](#) (MAchine-Readable Cataloging) standard used by the United States Library of Congress is one of them. It provides a detailed structure for documenting bibliographic resources, and is the recommended standard for well-resourced document libraries.

For the purpose of cataloguing documents in a less-specialized data repository, we propose a simpler schema, based on the [Dublin Core Metadata Element Set](#). The Dublin Core standard, developed by the [Dublin Core Metadata Initiative](#), consists of a list of fifteen core metadata elements to which more specialized elements can be added. These fifteen elements are the following (with their definition extracted from the Dublin Core [website](#)):

| No | Element name | Description |
|----|--------------------------|---|
| 1 | <code>contributor</code> | An entity responsible for making contributions to the resource. |
| 2 | <code>coverage</code> | The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant. |
| 3 | <code>creator</code> | An entity primarily responsible for making the resource. |
| 4 | <code>date</code> | A point or period of time associated with an event in the life cycle of the resource. |
| 5 | <code>description</code> | An account of the resource. |
| 6 | <code>format</code> | The file format, physical medium, or dimensions of the resource. |
| 7 | <code>identifier</code> | An unambiguous reference to the resource within a given context. |
| 8 | <code>language</code> | A language of the resource. |
| 9 | <code>publisher</code> | An entity responsible for making the resource available. |
| 10 | <code>relation</code> | A related resource. |
| 11 | <code>rights</code> | Information about rights held in and over the resource. |
| 12 | <code>source</code> | A related resource from which the described resource is derived. |
| 13 | <code>subject</code> | The topic of the resource. |
| 14 | <code>title</code> | A name given to the resource. |

| No | Element name | Description |
|----|----------------------|--------------------------------------|
| 15 | type | The nature or genre of the resource. |

Due to its simplicity and versatility, the Dublin Core is widely used for multiple purposes. It can be used to document not only documents but also resources of other types like images or others. Documents that can be described using the MARC21 standard can be described using the Dublin Core, although not with the same granularity of information. The US Library of Congress provides a [mapping between the MARC and the Dublin Core](#) metadata elements.

Another schema, [BibTex](#), has been developed for the specific purpose of recording bibliographic citations. BibTex is a list of fields that may be used to generate bibliographic citations compliant with different bibliography styles.

The metadata schema implemented in the Metadata Editor to document publications and reports is a combination of Dublin Core, MARC21, and BibTex elements. The technical documentation of the schema and its API is available at <https://ihsn.github.io/nada-api-redoc/catalog-admin/#tag/Documents>.

Documenting the publication or report

See the [Quick start: Document](#) chapter for a quick introduction to the documentation of a publication or report.

Create a new project

The first step in documenting a publication or report is to create a new project. You do that by clicking on [CREATE NEW PROJECT](#) in the *My projects* page. Select *Document* as data type. This will open a new, untitled *Project Home* page.

The screenshot shows the 'Metadata Editor' application interface. On the left is a dark sidebar with navigation links: 'Required', 'Recommended', 'Empty', 'Search...', 'Home', 'Metadata information', 'Document description' (expanded to show 'Title statement', 'Dates', 'Authors and contributors', 'Bibliographic information', 'Content description', 'Spatial and temporal context', 'Access and rights', 'Other information', 'Tags', and 'External resources'), and 'External resources'. The main content area has a title 'The Analysis of Household Surveys' and a 'SAVE' button. Below the title is a thumbnail image of the document cover. To the right of the thumbnail are details: 'Project owner: John Doe', 'Created on: 2025-02-12 10:25:09', 'Last changed by: John Doe', 'Changed on: 2025-02-12 10:52:36', and 'Project IDNO: 8e05b200-1753-4c41-8fca-5ee50197694b'. There are three main sections: 'Template' (Project template: DOCUMENT IHSN SCHEMA 1.0 EN -), 'Collaborators' (None), and 'Collections' (None). In the 'Template' section, there are two sub-sections: 'Administrative metadata templates:' (with a settings icon) and 'Project validation' (Schema validation: No validation errors found; Template validation: No validation errors found). The 'Data and Documentation' section shows a table with one row for 'The Analysis of Household Surveys' under 'DOCUMENTATION' tab, with columns for Title and Type.

In that page, edit the thumbnail (optional, not required). It is recommended to use a screenshot of the document's cover page as thumbnail.

Then, in the *Template* frame of the project *Home* page, select the project template you want to use to document the publication or report. A default template is proposed; no action is needed if you want to use the default template. Otherwise, switch to another template by clicking on the template name in the **Templates** frame. Note that you can at any time change the template used for the documentation of a project. The selected template will determine what you see in the navigation tree and in the metadata entry pages, but switching from one template to another will not impact the metadata that has already been entered; no information will be deleted from the metadata.

Enter metadata

The metadata schema comprises two main blocks of elements, `metadata_information` and `document_description`. It also contains a few sets of elements common to all data types (*Tags*, *DataCite*, *Provenance*, *External resources*, and *Administrative metadata*).

Metadata information

The `metadata_information` section contains information not related to the document itself but to its metadata. In other words, it contains metadata on the metadata. This information is optional but we recommend to enter content at least in the `name` and `date` sub-elements, which indicate who generated the metadata and when. This information is not useful to end-users of document catalogs, but is useful to catalog administrators for two reasons:

- metadata compliant with standards are intended to be shared and used by inter-operable applications. Data catalogs offer opportunities to harvest (pull) information from other catalogs, or to publish (push) metadata in other catalogs. Metadata information helps to keep track of the provenance of metadata.
- metadata for a same document may have been generated by more than one person or organization, or one version of the metadata can be updated and replaced with a new version. The `metadata_information` helps catalog

administrators distinguish and manage different versions of the metadata.

Document description

The **Document description** block contains the metadata elements used to describe the document. We provide below instructions or recommendations for some of the metadata elements of the metadata schema.

In the list of metadata elements below, the *key* of each element in the metadata standard is provided between brackets next to the corresponding element's label in the template.

TITLE STATEMENT

- **Primary ID** (*idno*; a required element): A unique identifier of the document, which serves as the "primary ID". *idno* is a unique identification number used to identify the database. A unique identifier is required for cataloguing purpose, so this element is declared as "Required". The identifier will allow users to cite the indicator/series properly. The identifier must be unique within the catalog. Ideally, it should also be globally unique; the recommended option is to obtain a Digital Object Identifier (DOI) for the study. Alternatively, the *idno* can be constructed by an organization using a consistent scheme. Note that the schema allows you to provide more than one identifier for a same study (in element *identifiers*); a catalog-specific identifier is thus not incompatible with a globally unique identifier like a DOI. The *idno* should not contain blank spaces.
- **Other identifiers** (*identifiers*) This element is used to enter document identifiers (IDs) other than the catalog ID entered in the **Title statement** (*idno*). It can for example be a Digital Object Identifier (DOI), an International Standard Book Number (ISBN), or an International Standard Serial Number (ISSN). The ID entered in the **title_statement** can be repeated here (the **Title statement** does not provide a **type** parameter; if a DOI, ISBN, ISSN, or other standard reference ID is used as *idno*, it is recommended to repeat it here with the identification of its **type**). The information on an identifier includes two components: the **type** of identifier (for example "DOI", "ISBN", or "ISSN"), and the **identifier** itself.
- **Title** (*title*; a required element): The title of the book, report, paper, or other document. Pay attention to the use of capitalization in the title, to ensure consistency across documents listed in your catalog. Pay attention to the consistent use of capitalization in the title. It is recommended to use sentence capitalization.
- **Subtitle** (*sub_title*) The document subtitle can be used when there is a need to distinguish characteristics of a document. Pay attention to the consistent use of capitalization in the subtitle.
- **Alternate title** (*alternate_title*) An alternate version of the title, possibly an abbreviated version. For example, the World Bank's World Development Report is often referred to as the WDR; the alternate title for the "World Development Report 2021" could then be "WDR 2021".
- **Translated title** (*translated_title*) A translation of the title of the document. Special characters should be properly displayed, such as accents and other stress marks or different alphabets.

DATES

- **Date created** (*date_created*) The date, preferably entered in ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY), when the document was produced. This can be different from the date the document was published, made available, and from the temporal coverage.
- **Date available** (*date_available*) The date, preferably entered in ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY), when the document was made available. This is different from the date it was published (see element **date_published** below).
- **Date modified** (*date_modified*) The date, preferably entered in ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY), when the document was last modified.
- **Date published** (*date_published*) The date, preferably entered in ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY), when the document was published.

AUTHORS AND CONTRIBUTORS

- **Authors** (*authors*) The authors should be listed in the same order as they appear in the source itself, which is not necessarily alphabetical. The information on authors include each author's `first_name`, `initials`, `last_name`, `affiliation`, and `author_id`. The author ID is an identifier in a registry of academic researchers such as the [Open Researcher and Contributor ID \(ORCID\)](#). This is a repeatable element, as a person may have multiple IDs. When entered in the Metadata Editor, this information should therefore include the `type` of identifier (for example "ORCID") and the identifier itself (`id`). An option is also provided to enter the author's `full_name`. This element should only be used when the first and last name of an author cannot be distinguished, i.e. when elements `first_name` and `last_name` cannot be filled out. This element can also be used when the author of a document is an entity, not a person.
- **Editors** (*editors*) If the source is a text within an edited volume, it should be listed under the name of the author of the text used, not under the name of the editor. The name of the editor should however be provided in the bibliographic citation, in accordance with a [reference style](#). The information on an editor includes the editor's `first_name`, `initial`, `last_name`, and `affiliation`.
- **Translators** (*translators*) Information on translators, for publications that are translations of publication originally created in another language.
 - **First name** (`first_name`) The first name of the translator.
 - **Initial** (`initial`) The initials of the translator.
 - **Last name** (`last_name`) The last name of the translator.
 - **Affiliation** (`affiliation`) The affiliation of the translator.
- **Other contributors** (*contributors*) These elements are used to acknowledge contributions to the production of the document, other than the ones for which specific metadata elements are provided (like `authors` or `translators`).
 - **First name** (`first_name`) The first name of the contributor.
 - **Initial** (`initial`) The initials of the contributor.
 - **Last name** (`last_name`) The last name of the contributor. If the contributor is an organization, enter the name of the organization here.
 - **Affiliation** (`affiliation`) The affiliation of the contributor.
 - **Contribution** (`contribution`) A brief description of the specific contribution of the person to the document, e.g. "Design of the cover page", or "Proofreading".

BIBLIOGRAPHIC INFORMATION

- **Bibliographic citation** (`bibliographic_citation`) The bibliographic citation provides relevant information about the author and the publication. When using the element `bibliographic_citation`, the citation is provided as a single item. It should be provided in a standard style: Modern Language Association ([MLA](#)), American Psychological Association ([APA](#)), or [Chicago](#). Note that the schema provides an itemized list of all elements (BibTex fields) required to build a citation in a format of their choice.
 - **Style** (`style`) The citation style, e.g. "MLA", "APA", or "Chicago".
 - **Citation** (`citation`) The citation in the style mentioned in `style`.

The elements that are required to form a complete bibliographic citation depend on the type of document. The table below, adapted from the [BibTex templates](#), provides a list of required and optional fields by type of document:

| Document type | Required fields | Optional fields |
|------------------------------------|------------------------------|---|
| Article from a journal or magazine | author, title, journal, year | volume, number, pages, month, note, key |

| Document type | Required fields | Optional fields |
|---|--|---|
| Book with an explicit publisher | author or editor, title, publisher, year | volume, series, address, edition, month, note, key |
| Printed and bound document without a named publisher or sponsoring institution | title | author, howpublished, address, month, year, note, key |
| Part of a book (chapter and/or range of pages) | author or editor, title, chapter and/or pages, publisher, year | volume, series, address, edition, month, note, key |
| Part of a book with its own title | author, title, book title, publisher, year | editor, pages, organization, publisher, address, month, note, key |
| Article in a conference proceedings | author, title, book title, year | editor, pages, organization, publisher, address, month, note, key |
| Technical documentation | title | author, organization, address, edition, month, year, key |
| Master's thesis | author, title, school, year | address, month, note, key |
| Ph.D. thesis | author, title, school, year | address, month, note, key |
| Proceedings of a conference | title, year | editor, publisher, organization, address, month, note, key |
| Report published by a school or other institution, usually numbered within a series | author, title, institution, year | type, number, address, month, note, key |
| Document with an author and title, but not formally published | author, title, note | month, year, key |

- **Book title** (*booktitle*) Title of a book, part of which is being cited. If you are documenting the book itself, this element will not be used; it is only used when part of a book is being documented.
- **Chapter** (*chapter*) A chapter (or section) number. This element is only used to document a resource which has been extracted from a book.
- **Edition** (*edition*) The edition of a book - for example "Second". When a book has no edition number/name present, it can be assumed to be a first edition. If the edition is other than the first, information on the edition of the book being documented must be mentioned in the citation. The edition can be identified by a number, a label (such as "Revised edition" or "Abridged edition"), and/or a year. The first letter of the label should be capitalized.

- **Institution** (*institution*) The sponsoring institution of a technical report. For citations of Master's and Ph.D. thesis, this will be the name of the school.
- **Journal** (*journal*) A journal name. Abbreviations are provided for many journals.
- **Volume** (*volume*) The volume of a journal or multi-volume book. Periodical publications, such as scholarly journals, are published on a regular basis in installments that are called issues. A volume usually consists of the issues published during one year.
- **Number** (*number*) The number of a journal, magazine, technical report, or of a work in a series. An issue of a journal or magazine is usually identified by its **volume** (see previous element) and **number**; the organization that issues a technical report usually gives it a number; and sometimes books are given numbers in a named series.
- **Pages** (*pages*) One or more page numbers or range of numbers, such as 42-111 or 7,41,73-97 or 43+ (the '+' indicates pages following that don't form a simple range).
- **Publisher** (*publisher*) The entity responsible for making the resource available. For major publishing houses, the information can be omitted. For small publishers, providing the complete address is recommended. If the company is a university press, the abbreviation UP (for University Press) can be used. The publisher is not stated for journal articles, working papers, and similar types of documents.
- **Publisher address** (*publisher_address***) The address of the publisher. For major publishing houses, just the city is given. For small publishers, the complete address can be provided.
- **Series** (*series*) The name of a series or set of books. When citing an entire book, the title field gives its title and an optional series field gives the name of a series or multi-volume set in which the book is published.
- **Cross reference** (*crossref*) The catalog identifier ("database key") of another catalog entry being cross referenced. This element may be used when multiple entries refer to a same publication, to avoid duplication.
- **Key** (*key*) A key is a field used for alphabetizing, cross referencing, and creating a label when the 'author' information is missing.
- **Organization** (*organization*) The organization that sponsors a conference or that publishes a manual.
- **Annotation** (*annotate*) An annotation. This element will not be used by standard bibliography styles like the MLA, APA or Chicago, but may be used by others that produce an annotated bibliography.
- **How published** (*howpublished*) The **howpublished** element is used to store the notice for unusual publications. The first word should be capitalized. For example, "WebPage", or "Distributed at the local tourist office".
- **URL** (*url*) The URL of the document, preferably a permanent URL.

CONTENT DESCRIPTION

- **Document type** (*type*) This describes the nature of the resource. It is highly recommended to select a value from a controlled vocabulary. The vocabulary can be entered in the metadata template (see section *Designing templates*), and could for example include the following options:
 - article
 - book
 - booklet
 - collection
 - conference proceedings
 - manual
 - master thesis
 - patent

- PhD thesis
- proceedings
- technical report
- working paper
- website
- other

Specialized agencies may want to create more specific controlled vocabularies. For example, a national statistical agency may use options like:

- Statistical Yearbooks
- Survey and Census Reports
- Analytical and Thematic Reports
- Statistical Bulletins
- News Releases
- Methodological and Technical Reports
- Metadata and Documentation Reports
- Annual Reports
- Administrative Documents
- Legal Documents
- Tenders

The `type` element can be used to create a "Document type" facet (filter) in a data catalog. If the controlled vocabulary is such that it contains values that are not mutually exclusive (i.e. if a document could possibly have more than one type), the element `type` cannot be used as it is not repeatable. In such case, the solution is to provide the type of document as `tags`, in a `tag_group` that could for example be named `type` or `document_type`. Note also that the Dublin Core provides a controlled vocabulary ([the DCMI Type Vocabulary](#)) for the `type` element, but this vocabulary is related to the types of resources (dataset, event, image, software, sound, etc.), not the type of document which is what we are interested in here.

- **`publication_frequency`** (***) Some documents are published regularly. The frequency of publications can be documented using this element. It is recommended to use a controlled vocabulary, for example the [PRISM Publishing Frequency Vocabulary](#) which identifies standard publishing frequencies for a serial or periodical publication.

| Frequency | Description |
|--------------|---|
| annually | Published once a year |
| semiannually | Published twice a year |
| quarterly | Published every 3 months, or once a quarter |
| bimonthly | Published twice a month |
| monthly | Published once a month |
| biweekly | Published twice a week |
| weekly | Published once a week |

| Frequency | Description |
|-------------|---|
| daily | Published every day |
| continually | Published continually as new content is added; typical of websites and blogs, typically several times a day |
| irregularly | Published on an irregular schedule, such as every month except July and August |
| other | Published on another schedule not enumerated in this controlled vocabulary |

- **Language** (*languages*) The language(s) in which the document is written. For the language codes and names, the use of the ISO 639-2 standard is recommended. This is a block of two elements (at least one must be provided for each language): `name` (the name of the language), and `code`. The use of [ISO 639-2](#) (the alpha-3 code in Codes for the representation of names of languages) is recommended. Numeric codes must be entered as strings.
- **Description** (*description*) This element is used to provide a brief description of the document (not an abstract, which would be provided in the field `abstract`). It should not be used to provide content that is contained in other, more specific elements. As stated in the [Dublin Core Usage Guide](#), "Since the `description` field is a potentially rich source of indexable terms, care should be taken to provide this element when possible. Best practice recommendation for this element is to use full sentences, as description is often used to present information to users to assist in their selection of appropriate resources from a set of search results."
- **Abstract** (*abstract*) The abstract is a summary of the document, usually about one or two paragraph(s) long (around 150 to 300 words).
- **Scope** (*scope*) A textual description of the topics covered in the document, which complements (but does not duplicate) the elements `description` and `topics` available in the schema.
- **Keywords** (*keywords*) A list of keywords that provide information on the core content of the document. Keywords provide a convenient solution to improve the discoverability of the document, as it allows terms and phrases not found in the document itself to be indexed and to make a document discoverable by text-based search engines. A controlled vocabulary can be used (although not required), such as the [UNESCO Thesaurus](#). The list provided here can combine keywords from multiple controlled vocabularies and user-defined keywords.
 - **Keyword** (*name*) The keyword itself.
 - **Vocabulary** (*vocabulary*) The controlled vocabulary (including version number or date) from which the keyword is extracted, if any.
 - **URL** (*uri*) The URL of the controlled vocabulary from which the keyword is extracted, if any.
- **Topics** (*topics*) Information on the topics covered in the document. A controlled vocabulary will preferably be used, for example the [CESSDA Topics classification](#), a typology of topics available in 11 languages; or the [Journal of Economic Literature \(JEL\) Classification System](#), or the [World Bank topics classification](#). The list provided here can combine topics from multiple controlled vocabularies and user-defined topics. The element is a block of five fields:
 - **ID** (*id*) The identifier of the topic, taken from a controlled vocabulary.
 - **Topic** (*name*) The name (label) of the topic, preferably taken from a controlled vocabulary.
 - **Parent ID** (*parent_id*) The parent identifier of the topic (identifier of the item one level up in the hierarchy), if a hierarchical controlled vocabulary is used.
 - **Vocabulary** (*vocabulary*) The name (including version number) of the controlled vocabulary used, if any.
 - **URL** (*uri*) The URL to the controlled vocabulary used, if any.
- **Themes** (*themes*) A list of themes covered by the document. A controlled vocabulary will preferably be used. The list provided here can combine themes from multiple controlled vocabularies and user-defined themes. Note that

`themes` will rarely be used as the elements `topics` and `disciplines` are more appropriate for most uses. This is a block of five fields:

- `ID` (*id*) The ID of the theme, taken from a controlled vocabulary.
 - `Theme` (*name*) The name (label) of the theme, preferably taken from a controlled vocabulary.
 - `Parent ID` (*parent_id*) The parent ID of the theme (ID of the item one level up in the hierarchy), if a hierarchical controlled vocabulary is used.
 - `Vocabulary` (*vocabulary*) The name (including version number) of the controlled vocabulary used, if any.
 - `URL` (*uri*) The URL to the controlled vocabulary used, if any.
- `Disciplines` (*disciplines*) Information on the academic disciplines related to the content of the document. A controlled vocabulary will preferably be used, for example the one provided by the list of academic fields in [Wikipedia](#). The list provided here can combine disciplines from multiple controlled vocabularies and user-defined disciplines. This is a block of five elements:
- `ID` (*id*) The identifier of the discipline, taken from a controlled vocabulary.
 - `Discipline` (*name*) The name (label) of the discipline, preferably taken from a controlled vocabulary.
 - `Parent ID` (*parent_id*) The parent identifier of the discipline (identifier of the item one level up in the hierarchy), if a hierarchical controlled vocabulary is used.
 - `vocabulary` (*vocabulary*) The name (including version number) of the controlled vocabulary used, if any.
 - `URL` (*uri*) The URL to the controlled vocabulary used, if any.
- `Table of contents` (*toc*) The table of content of the document, provided as a single string element, i.e. with no structure (an structured alternative is provided with the field `toc_structured` described below). This element is also a rich source of indexable terms which can contribute to document discoverability; care should thus be taken to use it (or the `toc_structured` alternative) whenever possible.
- `Table of contents (structured)` (*toc_structured*) This element is used as an alternative to `toc` to provide a structured table of content. The element contains a repeatable block of sub-elements which provides the possibility to define a hierarchical structure:
- `ID` (*id*) A unique identifier for the element of the table of content. For example, the `id` for Chapter 1 could be "1" while the `id` for section 1 of chapter 1 would be "11".
 - `Parent ID` (*parent_id*) The `id` of the parent section (e.g., if the table of content is divided into chapters, themselves divided into sections, the `parent_id` of a section would be the id of the chapter it belongs to.)
 - `Name` (*name*) The label of this section of the table of content (e.g., the chapter or section title)

SPATIAL AND TEMPORAL COVERAGE

- `Countries` (*ref_country*) The list of countries (or regions) covered by the document, if applicable. This is a repeatable block of two elements:
- `Name` (*name*) The country/region name. Note that many organizations have their own policies on the naming of countries/regions/economies/territories, which data curators will have to comply with.
 - `Code` (*code*) The country/region code. It is recommended to use a standard list of countries codes, such as the [ISO 3166] (https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes).

Considering the importance of the geographic coverage of a document as a filter, the `ref_country` element deserves particular attention. Consistency in the use of country codes and country names is essential in a data catalog. The field `ref_country` will often be used as a filter (facet) in data catalogs. Therefore, make sure that you do not refer to a same country using different names (for example, the Democratic Republic of Congo should not be named "Congo, D.R." in some instances and "Congo, Dem.Rep." or "Democratic Republic of congo" in other instances). When a document is related to only part of a country, we still want to capture this information in

the metadata. For example, the `ref_country` element for the document "[Sewerage and sanitation : Jakarta and Manila](#)" will list "Indonesia" (code IDN) and "Philippines" (code PHL).

The document title will often but not always provide the necessary information. Using R, Python or other programming languages, a list of all countries mentioned in a document can be automatically extracted, with their frequencies. This approach (which requires a lookup file containing a list of all countries in the world with their different denominations and spelling) can be used to extract the information needed to populate the `ref_country` element (not all countries in the list will have to be included; some threshold can be set to only include countries that are "significantly" mentioned in a document). Tools like the R package [countrycode](#) are available to facilitate this process.

When a document is related to a region (not to specific countries), or when it is related to a topic but not a specific geographic area, the `ref_country` might still be applicable. Try and extract (possibly using a script that parses the document) information on the countries mentioned in the document. For example, `ref_country` for the World Bank document "[The investment climate in South Asia](#)" should include Afghanistan (mentioned 81 times in the document), Bangladesh (113), Bhutan (94), India (148), Maldives (62), Nepal (64), Pakistan (103), and Sri Lanka (98), but also China (not a South-Asian country, but mentioned 63 times in the document).

If a document is not specific to any country, the element `ref_country` would be ignored (not included in the metadata) if the content of the document is not related to any geographic area (for example, the user's guide of a software application), or would contain "World" (code WLD) if the document is related but not specific to countries (for example, a document on "Climate change mitigation").

- **Geographic areas** (`geographic_units`) A list of geographic units covered in the document, other than the countries listed in `ref_country`. The geographic units will be identified by their `name`, `code`, and `type` (for example, "province", "state", "district", or "town").
- **Bounding box** (`bbox`) This element is used to define one or multiple geographic bounding box(es), which are the rectangular fundamental geometric description of the geographic coverage of the data. A bounding box is defined by west and east longitudes and north and south latitudes, and includes the largest geographic extent of the dataset's geographic coverage. The bounding box provides the geographic coordinates of the top left (north/west) and bottom-right (south/east) corners of a rectangular area. This element can be used in catalogs as the first pass of a coordinate-based search. The valid range of latitude in degrees is -90 and +90 for the southern and northern hemisphere, respectively. Longitude is in the range -180 and +180 specifying coordinates west and east of the Prime Meridian, respectively. This element will rarely be used for documenting publications. Bounding boxes are an optional element, but when a bounding box is defined, all four coordinates are required.
- **Spatial coverage** (`spatial_coverage`) This element provides another space for capturing information on the spatial coverage of a document, which complements the `ref_country`, `geographic_units`, and `bbox` elements. It can be used to qualify the geographic coverage of the document, in the form of a free text. For example, a report on refugee camps in the Cox's Bazar district of Bangladesh would have Bangladesh as reference country, "Cox's Bazar" as a geographic unit, and "Rohingya's refugee camps" as spatial coverage.
- **Temporal coverage** (`temporal_coverage`) Not all documents have a specific time coverage. When they do, it can be specified in this element.

ACCESS AND RIGHTS

- **Status** (`status`) The status of the document. The status of the document should (but does not have to) be provided using a controlled vocabulary, for example with the following options:
 - first draft
 - draft
 - reviewed draft
 - final draft
 - final Most documents published in a catalog will likely be "final".

- **Rights** (*rights*) A statement on the rights associated with the document (others than the copyright, which should be described in the element `copyright` described below).
- **Copyright** (*copyright*) A statement and identifier indicating the legal ownership and rights regarding use and re-use of all or part of the resource. If the document is protected by a copyright, enter the information on the person or organization who owns the rights.
- **License** (*license*) Information on the license(s) attached to the document, which defines the terms of use. A license is identified by its `name` (for example, CC BY 4.0 International) and `uri` (the URL of the license, where detailed information on the license can be obtained).
- **Usage_terms** (*usage_terms*) This element is used to provide a description of the legal terms or other conditions that a person or organization who wants to use or reproduce the document has to comply with.
- **Disclaimer** (*disclaimer*) A disclaimer limits the liability of the author(s) and/or publisher(s) of the document. A standard legal statement should be used for all documents from a same agency.
- **Security classification** (*security_classification*) Information on the security classification attached to the document. The different levels of classification indicate the degree of sensitivity of the content of the document. This field should make use of a controlled vocabulary, specific or adopted by the organization that curates or disseminates the document. Such a vocabulary could contain the following levels: `public, internal only, confidential, restricted, strictly confidential`
- **Access restrictions** (*access_restrictions*) A textual description of access restrictions that apply to the document.
- **Pricing** (*pricing*) The current price of the document in any defined currency. As this information is subject to regular change, it will often not be included in the document metadata.

OTHER INFORMATION

- **contacts** (*contacts*) Contact information for a person or organization that can be contacted for inquiries related to the document.
 - `name` (*name*) The name of the contact. This can be a person or an organization..
 - `role` (*role*) The specific role of the person or organization mentioned in `contact` .
 - **affiliation** (*affiliation*) The affiliation of the contact person.
 - `email` (*email*) The email address of the contact person or organization. Personal emails should be avoided.
 - `telephone` (*telephone*) The telephone number for the contact person or organization. Personal phone numbers should be avoided.
 - `uri` (*uri*) A link to an on-line resource related to the contact person or organization.
- **Sources** (*sources*) This element is used to describe the sources of different types (except data sources, which must be listed in the element "Data sources") that were used in the production of the document.
 - **source_origin** (*source_origin*) For historical materials, information about the origin(s) of the sources and the rules followed in establishing the sources should be specified.
 - **source_char** (*source_char*) Characteristics of the source. Assessment of characteristics and quality of source material.
 - **source_doc** (*source_doc*) Documentation and access to the source.
- **Data sources** (*data_sources*) Used to list the machine-readable data file(s) -if any- that served as the source(s) of data.
 - `name` (*name*) Name (title) of the dataset used as source. For example: "Bangladesh Demographic and Health Survey 2017-18"

- **uri** (*uri*) Link (URL) to the dataset or to a web page describing the dataset. For example: "<https://www.dhsprogram.com/methodology/survey/survey-display-536.cfm>"
- **note** (*note*) Additional information on the data source. For example: "Household survey conducted by the National Institute of Population Research and Training, Medical Education and Family Welfare Division and Ministry of Health and Family Welfare. Data and documentation available at <https://dhsprogram.com/>"
- **Reproducibility statement** (*reproducibility.statement*) The "Reproducibility statement" is a general statement on reproducibility and replicability of the analysis (including data processing, tabulation, production of visualizations, modeling, etc.) being presented in the document.
- **Reproducibility links** (*reproducibility.links*) The "Reproducibility links" provides links to web pages where reproducible materials and the related information can be found.
- **Audience** (*audience*) Information on the intended audience for the document, i.e. the category or categories of users for whom the resource is intended in terms of their interest, skills, status, or other.
- **Mandate** (*mandate*) The legislative or other mandate under which the resource was produced.
- **Related resources** (*relations*) References to related resources with a specification of the type of relationship.
 - **name** (*name*) The related resource. Recommended practice is to identify the related resource by means of a URL. If this is not possible or feasible, a string conforming to a formal identification system may be provided.
 - **type** (*type*) The type of relationship. The use of a controlled vocabulary is recommended. The Dublin Core proposes the following vocabulary: { `isPartOf` , `hasPart` , `isVersionOf` , `isFormatOf` , `hasFormat` , `references` , `isReferencedBy` , `isBasedOn` , `isBasisFor` , `replaces` , `isReplacedBy` , `requires` , `isRequiredBy` }.
- **Notes** (*notes*) This field can be used to provide information on the document that does not belong to the other, more specific metadata elements provided in the schema.

DataCite

See section **Documenting - General instructions**.

Tags

See section **Documenting - General instructions**.

Provenance

The **Provenance** container is used to document how and when the dataset was acquired. It is used to ensure traceability. See chapter **Documenting - General instructions**.

External resources

If you intend to publish the metadata and the publication or report in a catalog, you may want to provide users not only with a description of the document, but also with the document itself (or with a link to the document), and possibly other related materials (such as a link to electronic annexes, images, tabulations in Excel format, related news releases, or other). You will do that by attaching and documenting one or multiple external resources to the metadata (for example, a PDF copy of the document).

External resources are all materials (and links) that relate to the indicator. This may include documents on methodology, scripts, photos and videos, and any other resource available in digital format. These materials and links are added to the documentation of an indicator in the External resources container. Select *External resources* in the navigation tree, then

on [CREATE RESOURCE](#). Enter the relevant information on the resource (at least a title), then provide either a filename (the file will then be uploaded on the server that hosts the Metadata Editor) or a URL to the resource.

External resources that have already been created for another project can also be imported. To do that, they must first be exported as JSON or RDF from the other project. Then click on [IMPORT](#) in the External resources page, and select the file.

Administrative metadata

One or multiple administrative metadata templates can be attached to the project. See [See chapter Documenting - General instructions](#) and [Administrative metadata](#).

Documenting indicators and databases

Two metadata schemas are employed to document indicators and their respective databases. The first schema is utilized to individually document each indicator. The second standard is applied to document collections or databases of indicators. We refer to these metadata as schemas and not standards, as they have been developed by the World Bank and are not maintained by an international community.

Metadata schemas

Metadata schema for documenting indicators

Indicators are summary measures that capture key issues or phenomena, derived from observed data. When indicators are presented for a specific geographic area and include a temporal dimension — such as annual, quarterly, monthly, or daily values — they form time series.

To facilitate standardized documentation of indicators, the World Bank developed a metadata standard by compiling and structuring metadata elements commonly used by various organizations, including the World Bank itself, United Nations agencies, the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), and Eurostat.

The World Bank metadata schema aligns with the Statistical Data and Metadata Exchange standard (SDMX), a standard designed to enable machine-to-machine data exchanges.

- The *descriptive and reference metadata* component of the World Bank schema aligns with SDMX Metadata Structure Definitions (MSDs). MSDs provide a framework for structuring metadata, but do not define the metadata elements to be included. The World Bank's schema therefore complements SDMX by specifying detailed content that can be incorporated into SDMX MSDs and metadatasets.
- The *structural metadata* in the World Bank schema, which describes how indicator data are organized in the data file or database, aligns with the SDMX Data Structure Definitions (DSDs).

The Metadata Editor supports this integration by offering tools to:

- Export metadata templates as MSDs – ensuring compatibility with SDMX metadata structures.
- Export indicator metadata as meta-datasets – enabling seamless incorporation of metadata into SDMX-compliant systems.

Metadata schema for documenting databases of indicators

In addition to documenting individual indicators, the World Bank has developed a complementary metadata schema for documenting databases of indicators (i.e., collections of indicators). This schema provides additional metadata at the database level, which enriches the contextual information available for each indicator.

To establish a clear link between an indicator and its associated database, the indicator metadata schema includes a dedicated element that stores the database identifier, enabling seamless association between indicator metadata and database metadata.

In data catalogs like a NADA catalog, the metadata on indicators and the related databases can be combined.

Age dependency ratio (% of working-age population)
1960 - 2020 - SP.POP.DPND (Source database: World Development Indicators)

Created on October 15, 2021 Last modified October 15, 2021 Page views 251 Metadata JSON

SERIES DESCRIPTION

Overview

SERIES UNIQUE ID
SP.POP.DPND

SERIES NAME
Age dependency ratio (% of working-age population)

DATABASE ID
WLD_2021_WDI_v01_M

PERIODICITY OF DATA

SOURCE DATABASE

TITLE
World Development Indicators

ALTERNATE TITLE
WDI

DATABASE ID
WLD_2021_WDI_v01_M

AUTHORING ENTITY

| Name | Role | Affiliation |
|------------------------|----------|----------------------|
| Development Data Group | Producer | The World Bank Group |

ABSTRACT
The primary World Bank collection of development indicators, compiled from officially-recognized international sources. It presents the most current and accurate global development data available, and includes national, regional and global estimates.

Documenting an indicator

This section describes in detail the process of documenting an indicator, and the various functionalities provided in the Metadata Editor for generating and publishing indicators metadata. The section focuses on the functionalities specific to the documentation of indicators. Refer to the *General instructions* for guidance on the components of the application that are common to all data types.

Key principles for producing high-quality metadata

The following general principles should be followed to ensure the production of high-quality metadata. They apply broadly across metadata elements used to document statistical indicators. Subsequent sections of this Guide provide detailed instructions and quality criteria for specific metadata elements. These elements and criteria form the foundation of an AI-enabled system for metadata quality assurance and augmentation.

- **Standardize terminology.** Use consistent terms across all metadata. Where applicable, adopt internationally recognized controlled vocabularies or code lists (e.g., for topics, sectors, geographic areas) to enhance interoperability and comparability.
- **Maintain consistency in terminology and jargon.** Avoid mixing technical jargon with general terms unless clearly explained. Use consistent language and avoid introducing synonyms or alternative phrases for the same concept without clear justification.
- **Link related indicators.** Where applicable, establish links among related indicators. Group variants, disaggregations, or derived indicators using relationships such as: Is part of, Has disaggregation, Derived from. This helps users navigate the catalog and understand the relationships among indicators.
- **Spell out acronyms and abbreviations.** Avoid using unexplained acronyms, especially in definitions. Widely recognized acronyms (e.g., GDP, USD) may be used in titles, but must be spelled out or explained in the metadata definition or notes.
- Use accessible and non-technical language. Write metadata in a clear and accessible style. Avoid excessive technical detail. Complex methodological information should be included in referenced documentation, not embedded in core metadata fields.

Create a new project

The first step in documenting an indicator is to create a new project. You do that by clicking on **CREATE NEW PROJECT** in the *My projects* page, then selecting *Indicator* as data type when prompted. This will open a new, untitled indicator Project page.

In that page, **select the template** you want to use to document the indicator. A default template is proposed; no action is needed if you want to use the default template. Otherwise, switch to another template by clicking on the template name in the *Templates* frame. Note that you can at any time change the template used for the documentation of a project. The selected template will determine what you see in the navigation tree and in the metadata entry pages, but switching from one template to another will not impact the metadata that has already been entered; no information will be deleted from the metadata.

Enter information on metadata

The *Metadata information* section in the navigation tree (in the *Project page*) contains elements intended to document the metadata being generated, i.e., metadata about the metadata. All content in this section is optional; it is however recommended practice to document the metadata as precisely as possible. This information will not be useful to data users, but it will be to catalog administrators. When metadata is shared across catalogs, the information entered in the *Information on metadata* provides transparency and clarity on the origin of the metadata.

Enter the indicator description (descriptive metadata)

The documentation of an indicator is a relatively straightforward process. The navigation tree in the *Project page* provides access to various metadata entry pages where the information about the indicator can be captured. These metadata entry pages are defined by the selected template.

The template includes a description of each metadata element it contains, which can be used as instructions to data curators. Templates can be exported to PDF format, and used as a reference documents by data curators. The instructions contained in templates are also displayed in the metadata entry pages, by clicking on the  icon shown next to each metadata element's label.

We provide below some guidance on a selection of the metadata elements found in the metadata schema. A small number of these elements are marked as *required*. When documenting an indicator, it is advised to provide the most comprehensive information possible.

In the list of metadata elements below, the *key* of each element in the metadata standard is provided between brackets next to the corresponding element's label in the template.

TITLE STATEMENT

- **Primary ID** (`idno`) A unique identifier (ID) for the indicator. This is a required element. Most agencies will (and should) use a coherent coding convention to generate their indicators identifiers
- **DOI** (`doi`) A Digital Object Identifier (DOI) for the the indicator. See the *Documenting data - General instructions* section of the User Guide for more information on DOIs.
- **Other identifiers** (`alternate_identifiers`) The element `idno` described above is the reference unique identifier for the catalog in which the metadata is intended to be published. But the same indicator/metadata may be published in other catalogs. For example, a data catalog may publish metadata for indicator extracted from the World Bank World Development Indicators (WDI) database. And the WDI itself contains indicators generated and published by other organizations, such as the World Health Organization or UNICEF. Catalog administrators may want to assign a unique identifier specific to their catalog (the `idno` element), but keep track of the identifier of the indicator or indicator in other catalogs or databases. The `alternate_identifiers` element serves that purpose. It includes the following sub-elements:
 - **Type** (`type`) The type of identifier. For example: "ISBN".
 - **Identifier** (`name`) This element will be used to provide the identifier.
 - **Database** (`database`) The name of the database (or catalog) where this alternative identifier is used, e.g. "IMF, International Financial Statistics (IFS)".
 - **URL** (`uri`) A link (URL) to the database mentioned in `database`.

- **Notes** (notes) Any additional information on the alternate identifier.
- **Name** (name) The name (label) of the indicator. Note that a field **aliases** is provided (see below) to capture alternative names for the indicator.
- **Display name** (display_name) The name (label) of the indicator as it should be displayed in a data catalog.
- **Aliases** (aliases) An indicator can be referred to using different names. The **aliases** element is provided to capture the multiple names and labels that may be associated with (i.e. synonyms of) the documented indicator.
- **Database ID** (database_id) The unique identifier of the database the indicator belongs to. This field must correspond to the element `database_description > title_statement > idno` in the database schema. This is the field that will be used to establish the link between the database metadata and the indicator metadata.
- **Database name** (database_name) The name of the database the indicator belongs to (name of the database identified in **Database ID**).
- **Date released** (date_released) The date, in ISO format (YYYY-MM-DD) when the indicator was released.

SOURCES, CONCEPTS, AND METHODS

- **Definition short** (definition_short) A short definition of the indicator. The short definition captures the essence of the indicator.
- **Definition long** (definition_long) A long(er) version of the definition of the indicator. If only one definition is available (not a short/long version), it is recommended to capture it in the **definition_short** element. Alternatively, the same definition can be stored in both **definition_short** and **definition_long**.
- **Definition references** (definition_references) This element is provided to link to an external resources from which the definition was extracted.
 - **Source** (source) The source of the definition (title, or label).
 - **URL** (uri) A link (URL) to the source of the definition.
 - **Note** (note) This element provides for annotating or explaining the reason the reference has been included as part of the metadata.
- **Relevance** (relevance) This field documents the relevance of an indicator in relation to a social imperative or policy objective.
- **Mandate**
 - **Mandate** (mandate) Description of the institutional mandate or of a set of rules or other formal set of instructions assigning responsibility as well as the authority to an organization for the collection, processing, and dissemination of statistics for this indicator.
 - **Mandate URL** (URI) A link to a resource (document, website) describing the mandate.
- **Data collection** (data_collection) This group of elements can be used to document data collection activities that led to or allowed the production of the indicator. This element will typically be used for the description of surveys or censuses. Note: the schema also contains an element "sources". That element will be used to document the organization and/or main data production program from which the indicator is derived.
 - **Data source** (data_source) A concise and standardized name (label) for the data source, e.g. "National Labor Force Survey, 1st quarter 2022". If multiple data sources were used, they can all be listed here. Note that if an indicator has values obtained from many different sources, the source for each value (or group of values) will not be part of the indicator metadata, but will be stored as an attribute in the data file where the information can be associated with a specific observation ("cell note" or group of observation (e.g. attached to an indicator for avv values for a same year or for a same area).

- **Data collection method** (*method*) Brief information on the data collection method, e.g. "Sample household survey".
- **Data collection period** (*period*) Information on the period of the data collection, e.g. "January to March 2022".
- **Data collection note** (*note*) Additional information on the data collection.
- **Data collection URL** (*uri*) A link to a resource (website, document) where more information on the data collection can be found.
- **Methodology** (*methodology*) Methodological details on the production of the series or indicator.
- **Methodology references** (*methodology_references*) This element is provided to link to an external resources from which the definition was extracted.
 - **Source** (*source*) The source of the information on methodology.
 - **URL** (*uri*) A link (URL) to the source of the information on methodology.
 - **Note** (*note*) This element provides for annotating or explaining the reason the reference has been included as part of the metadata.
- **Derivation** (*derivation*) Description of the derivation method (not including imputations, which should be described in element "Imputation").
 - **Source** (*source*) The source of the information on derivation.
 - **URL** (*uri*) A link (URL) to the source of the information on derivation.
 - **Note** (*note*) This element provides for annotating or explaining the reason the reference has been included as part of the metadata.
- **Derivation references** (*derivation_references*) This element is provided to link to an external resources from which the definition was extracted.
- **Imputation** (*imputation*) Data may have been imputed to account for data gaps or for other reasons (harmonization/standardization, and others). If imputations have been made, this element provides the space for their description.
- **Imputation references** (*imputation_references*) This element is provided to link to an external resources from which the definition was extracted.
 - **Source** (*source*) The source of the information on imputation.
 - **URL** (*uri*) A link (URL) to the source of the information on imputation.
 - **Note** (*note*) This element provides for annotating or explaining the reason the reference has been included as part of the metadata.
- **Statistical concept** (*statistical_concept*) This element allows to insert a reference of the indicator with content of a statistical character. This can include coding concepts or standards that are applied to render the data statistically relevant.
- **Concept references** (*concept_references*) This element is provided to link to an external resources from which the definition was extracted.
 - **Source** (*source*) The source of the information on statistical concepts.
 - **URL** (*uri*) A link (URL) to the source of the information on statistical concepts.
 - **Note** (*note*) This element provides for annotating or explaining the reason the reference has been included as part of the metadata.
- **Related concepts** (*concepts*) This repeatable element can be used to document concepts related to the indicators (other than the main statistical concept that may have been entered in **statistical_concept**). For example, the

concept of *malnutrition* could be documented in relation to the indicators "Prevalence of stunting" and "Prevalence of wasting".

- **Name** (*name*) A concise and standardized name (label) for the concept.
- **Definition** (*definition*) The definition of the concept.
- **URL** (*uri*) A link (URL) to a resource providing more detailed information on the concept.

- **Aggregation method** (*aggregation_method*) The *aggregation_method* element describes how values can be aggregated from one geographic level (for example, a country) to a higher-level geographic area (for example, a group of country defined based on a geographic criteria (region, world) or another criteria (low/medium/high-income countries, island countries, OECD countries, etc.). The aggregation method can be simple (like "sum" or "population-weighted average") or more complex, involving weighting of values.

- **Aggregation references** (*aggregation_references*) This element is provided to link to an external resources from which the aggregation method was extracted.
 - **Source** (*source*) The source of the information on aggregation.
 - **URL** (*uri*) A link (URL) to the source of the information on aggregation.
 - **Note** (*note*) This element provides for annotating or explaining the reason the reference has been included as part of the metadata.

- **Sources of data** (*sources*) This element provides information on the source(s) of data that were used to generate the indicator. A source can refer to an organization (e.g., "Source: World Health Organization"), or to a dataset (e.g., for a national poverty headcount indicator, the sources will likely be a list of sample household surveys). In *sources*, we are mainly interested in the latter. When an indicator in a database is an indicator extracted from another database (e.g., when the World Bank World Development Indicators include an indicator from the World Health Organization in its database), the source organization should be mentioned in the *authoring_entity* element of the schema. The *sources* element is a repeatable element. Note 1: In some cases, the source of a specific value in a database will be stored as an attribute of the data file (e.g., as a "footnote" attached to a specific cell. If the sources are listed in the data file, they may but do not need to be stored in the metadata. Note 2: the schema also contains an element "data_collection" that would be used to describe a specific data collection activity from which an indicator is derived.
 - **Identifier** (*idno*) This element records the unique identifier of a source. It is a required element. If the source does not have a specific unique identifier, a sequential number can be used. If the source is a dataset or database that has its own unique identifier (possibly a DOI), this identifier should be used.
 - **Other identifiers** (*other_identifiers*) This repeatable element is used to enter identifiers (IDs) other than the primary ID.
 - **Type** (*type*) The type of identifier. For example: "DOI", or "ISBN".
 - **Identifier** (*identifier*) The identifier itself.
 - **Type** (*type*) The type of source, e.g. "household survey", "administrative data", or "external database".
 - **Name** (*name*) The name (title, or label) of the source.
 - **Organization** (*organization*) The organization responsible for the source data.
 - **Authors** (*authors*) Authors (with detailed information on authors)
 - **Datasets** (*datasets*)
 - **Identifier** (*idno*) The identifier of the dataset.
 - **Title** (*idno*) Name (title) of the dataset.
 - **URL** (*idno*) Link to a dataset website.
 - **Publisher** (*publisher*)
 - **Publication date** (*publication_date*)
 - **URL** (*uri*)
 - **Date accessed** (*access_date*)

- **Notes** (note) This element can be used to provide additional information regarding the source data.
- **Notes on data source** (sources_note) Additional information on the source(s) of data used to generate the indicator or indicator.

STANDARDS AND FRAMEWORKS

- **Standards** (compliance) For some indicators, international standards have been established. This is for example the case of indicators like the unemployment or unemployment rate, for which the International Conference of Labour Statisticians defines the standards concepts and methods. The **compliance** element is used to document the compliance of an indicator with one or multiple national or international standards.
 - **Name** (standard) The name of the standard that the indicator complies with. This name will ideally include a label and a version or a date. For example: "International Standard Industrial Classification of All Economic Activities (ISIC) Revision 4, published in 2007"
 - **Abbreviation** (abbreviation) The acronym of the standard that the indicator complies with.
 - **Custodian** (custodian) The organization that maintains the standard that is being used for compliance. For example: "United Nations Statistics Division".
 - **URL** (uri) A link to a public website site where information on the compliance standard can be obtained. For example: "<https://unstats.un.org/unsd/classifications/Family/Detail/27>"
- **Frameworks** (framework) Some national, regional, and international agencies develop monitoring frameworks, with goals, targets, and indicators. Some well-known examples are the [Millennium Development Goals](#) and the [Sustainable Development Goals](#) which establish international goals for human development, or the World Summit for Children (1990) which set international goals in the areas of child survival, development and protection, supporting sector goals such as women's health and education, nutrition, child health, water and sanitation, basic education, and children in difficult circumstances. The **framework** element is used to link an indicator to the framework, goal, and target associated with it.
 - **Name** (name) The name of the framework.
 - **Abbreviation** (abbreviation) The abbreviation of the name of the framework.
 - **Custodian** (custodian) The name of the organization that is the official custodian of the framework.
 - **Description** (description) A brief description of the framework.
 - **Goal ID** (goal_id) The identifier of the Goal that the indicator is associated with.
 - **Goal name** (goal_name) The name (label) of the Goal that the indicator is associated with.
 - **Goal description** (goal_description) A brief description of the Goal that the indicator is associated with.
 - **Target ID** (target_id) The identifier of the Target that the indicator is associated with.
 - **Target name** (target_name) The name (label) of the Target that the indicator is associated with.
 - **Target description** (target_description) A brief description of the Target that the indicator is associated with.
 - **Indicator ID** (indicator_id) The identifier of the indicator, as provided in the framework (this is not the **idno** identifier).
 - **Indicator name** (indicator_name) The name of the indicator, as provided in the framework (which may be different from the name provided in **name**)
 - **Indicator description** (indicator_description) A brief description of the indicator, as provided in the framework.
 - **URL** (uri) A link to a website providing detailed information on the framework, its goals, targets, and indicators.
 - **Notes** (notes) Any additional information on the relationship between the indicator and the framework.

QUALITY

- **Limitation** (limitation) This element is used to communicate to the user any limitations or exceptions in using the data. The limitations may result from the methodology, from issues of quality or consistency in the data source, or

other.

- **Validation rules** (*validation_rules*) Description of the set of rules (itemized) used to validate values for the indicator, e.g. "Is within range 0-100", or "Is the sum of indicator X + indicator Y".
- **Quality checks** (*quality_checks*) Data may have gone through data quality checks to assure that the values are reasonable and coherent, which can be described in this element. These quality checks may include checking for outlying values or other. A brief description of such quality control procedures will contribute to reinforcing the credibility of the data being disseminated.
- **Quality note** (*quality_note*) Additional notes or an overall statement on data quality. These could for example cover non-standard quality notes and/or information on independent reviews on the data quality.
- **Discrepancies** (*sources_discrepancies*) This element is used to describe and explain why the data in the indicator may be different from the data for the same indicator published in other sources. International organizations, for example, may apply different techniques to make data obtained from national sources comparable across countries, in which cases the data published in international databases may differ from the data published in national, official databases.
- **Adjustments** (*adjustments*) Description of any adjustments with respect to use of standard classifications and harmonization of breakdowns for age group and other dimensions, or adjustments made for compliance with specific international or national definitions.
- **Missing** (*missing*) Information on missing values in the indicator or indicator. This information can be related to treatment of missing values, to the cause(s) of missing values, and others.
- **Errata** (*errata*) This element is used to provide information on detected errors in the data or metadata for the indicator, and on the measures taken to remedy them.
 - **Date** (*date*) The date the erratum was published.
 - **Description** (*description*) A description of the error and remedy measures.
- **Acknowledgements** (*acknowledgements*) Itemized list of persons and organizations being acknowledged
 - **Name** (*name*) Name of the person or organization being acknowledged.
 - **Affiliation** (*affiliation*) Affiliation of the person being acknowledged.
 - **Role** (*role*) Role of the person or organization being acknowledged.
- **Acknowledgement statement** (*acknowledgement_statement*) Overall statement of acknowledgement.
- **Disclaimer** (*disclaimer*) A disclaimer statement that applies to the indicator.

GEOGRAPHIC AND TIME COVERAGE

- **Time coverage** (*time_periods*) The time period covers the entire span of data available for the indicator. The time period has a start and an end and is reported according to the periodicity provided in a previous element. The dates should be entered in ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY).
 - **Start** (*start*) The initial date of the indicator in the dataset.
 - **End** (*end*) The end date is the latest date for which an estimate for the indicator is available.
- **Countries** (*ref_country*) A list of countries or economies for which data are available in the indicator. This element is somewhat redundant with the next element (**geographic_units**) which may also contain a list of countries. Identifying geographic areas of type "country" is important to enable filters and facets in data catalogs (country names are among the most frequent queries submitted to catalogs).
 - **Name** (*name*) The name of the country.
 - **Code** (*code*) The code of the country. The use of the [ISO 3166-1 alpha-3](#) codes is recommended.

- **Geographic areas** (*geographic_units*) List of geographic units (regions, countries, states, provinces, etc.) for which data are available for the indicator.
 - **Name** (*name*) Name of the geographic unit e.g. "World", "Africa", "Afghanistan", "OECD countries", "Bangkok".
 - **Code** (*code*) Code of the geographic unit. The [ISO 3166-1 alpha-3](#) code is preferred when the unit is a country.
 - **Type** (*type*) Type of geographic unit e.g. "country", "state", "region", "province", "city", etc.
- **Bounding box** (*bbox*) This element is used to define one or multiple bounding box(es), which are the rectangular fundamental geometric description of the geographic coverage of the data. A bounding box is defined by west and east longitudes and north and south latitudes, and includes the largest geographic extent of the dataset's geographic coverage. The bounding box provides the geographic coordinates of the top left (north/west) and bottom-right (south/east) corners of a rectangular area. This element can be used in catalogs as the first pass of a coordinate-based search. This element is optional, but if the **bound_poly** element (see below) is used, then the **bbox** element must be included.
 - **West** (*west*) West longitude of the bounding box.
 - **East** (*east*) East longitude of the bounding box.
 - **South** (*south*) South latitude of the bounding box.
 - **North** (*north*) North latitude of the bounding box.

DESCRIPTION

- **Authoring entity** (*authoring_entity*) This set of five elements is used to identify the organization(s) or person(s) who are the main producers/curators of the indicator. Note that a similar element is provided at the database level. The authoring_entity for the indicator can be different from the authoring_entity of the database. For example, the World Bank is the authoring entity for the World Development Indicators database, which contains indicators obtained from the International Monetary Fund, World Health Organization, and other organizations that are thus the authoring entities for specific indicators.
 - **Name** (*name*) The name of the person or organization who is responsible for the production of the indicator. Write the name in full (use the element **abbreviation** to capture the acronym of the organization, if relevant).
 - **Affiliation** (*affiliation*) The affiliation of the person or organization mentioned in **name**.
 - **Abbreviation** (*abbreviation*) Abbreviated name (acronym) of the organization mentioned in **name**.
 - **Email** (*email*) The public email contact of the person or organizations mentioned in **name**. It is good practice to provide a service account email address, not a personal one.
 - **URL** (*uri*) A link (URL) to the website of the entity mentioned in **name**.
- **Measurement unit** (*measurement_unit*) The unit of measurement. Note that in many databases the measurement unit will also be mentioned in the indicator name. The World Bank's World Development Indicators (WDI) for example, contains indicators named as follows: "CO2 emissions (kg per 2010 US\$ of GDP)", "GDP per capita (current US\$)", "GDP per capita (current LCU)", or "Population density (people per sq. km of land area)".
- **Dimensions** (*dimensions*) An indicator can be made available with different levels of disaggregation. For example, an indicator containing annual estimates of the "Resident population (mid-year)" can be provided by urban/rural area of residence, by sex, and by age group. The **dimension** element may be used to document such disaggregations. Note that the element should only be used when the **Data Structure Definition** of the indicator is not documented (see below the section on structural metadata). Documenting the Data Structured Definition is the preferred way of documenting the dimensions of an indicator.
 - **Name** (*name*) The name of the dimension.
 - **Label** (*label*) The label of the dimension, for example "sex", or "urban/rural".
 - **Description** (*description*) A description of the dimension (for example, if the label is "age group", the description can provide detailed information on the age groups, e.g. "The age groups in the database are 0-14, 15-49, 50-64, and 65+ years old".)

- **Release calendar** (*release_calendar*) Information on when updates for the indicators can be expected. This will usually not consist of exact dates (which would have to be updated regularly), but of more general information like "Every first Monday of the Month", or "Every year on June 30", or "The last week of each quarter".
- **Periodicity** (*periodicity*) The periodicity of the indicator. It is recommended to use a controlled vocabulary with values like *annual*, *quarterly*, *monthly*, *daily*, etc.
- **Base period** (*base_period*) The base period for the indicator. This field will only apply to indicators that require a base year (or other reference time) used as a benchmark, like a Consumer Price Index (CPI) which will have a value of 100 for a reference base year.
- **Breaks in series** (*series_break*) Breaks in statistical series occur when there is a change in the standards, sources of data, or reference year used in the compilation of a series. Breaks in series must be well documented. The documentation should include the reason(s) for the break, the time it occurred, and information on the impact on comparability of data over time.
- **Keywords** (*keywords*) Words or phrases that describe salient aspects of a data collection's content. Can be used for building keyword indexes and for classification and retrieval purposes. A controlled vocabulary can be employed. Keywords should be selected from a standard thesaurus, preferably an international, multilingual thesaurus.
 - **Keyword** (*name*) Keyword (or phrase). Keywords summarize the content or subject matter of the study.
 - **Vocabulary** (*vocabulary*) Controlled vocabulary from which the keyword is extracted, if any.
 - **URL** (*uri*) The URI of the controlled vocabulary used, if any.
- **Topics** (*topics*) The **topics** field indicates the broad substantive topic(s) that the indicator covers. A topic classification facilitates referencing and searches in electronic survey catalogs. Topics should be selected from a standard controlled vocabulary such as the [Council of European Social Science Data Archives \(CESSDA\) topics classification](#).
 - **ID** (*id*) The unique identifier of the topic. It can be a sequential number, or the ID of the topic in a controlled vocabulary.
 - **Label** (*name*) The label of the topic associated with the data.
 - **Parent ID** (*parent_id*) When a hierarchical (nested) controlled vocabulary is used, the **parent_id** field can be used to indicate a higher-level topic to which this topic belongs.
 - **Vocabulary** (*vocabulary*) The name of the controlled vocabulary used, if any.
 - **URL** (*uri*) A link to the controlled vocabulary mentioned in field **vocabulary**.
- **Themes** (*themes*) Themes provide a general idea of the research that might guide the creation and/or demand for the indicator. A theme is broad and is likely also subject to a community based definition or list. A controlled vocabulary should be used. This element will rarely be used (the element **topics** described below will be used more often).
 - **ID** (*id*) The unique identifier of the theme. It can be a sequential number, or the ID of the theme in a controlled vocabulary.
 - **Name** (*name*) The label of the theme associated with the data.
 - **Parent ID** (*parent_id*) When a hierarchical (nested) controlled vocabulary is used, the **parent_id** field can be used to indicate a higher-level theme to which this theme belongs.
 - **Vocabulary** (*vocabulary*) The name of the controlled vocabulary used, if any.
 - **URL** (*uri*) A link to the controlled vocabulary mentioned in field 'vocabulary'.
- **Disciplines** (*disciplines*) Information on the academic disciplines related to the content of the document. A controlled vocabulary will preferably be used, for example the one provided by the list of academic fields in [Wikipedia](#). This is a block of five elements:
 - **ID** (*id*) The ID of the discipline, preferably taken from a controlled vocabulary.

- **Name** (*name*) The name (label) of the discipline, preferably taken from a controlled vocabulary.
- **Parent ID** (*parent_id*) The parent ID of the discipline (ID of the item one level up in the hierarchy), if a hierarchical controlled vocabulary is used.
- **Vocabulary** (*vocabulary*) The name (including version number) of the controlled vocabulary used, if any.
- **URL** (*uri*) The URL to the controlled vocabulary used, if any.
- **Disaggregation** (*disaggregation*) This element is intended to inform users that an indicator is available at various levels of disaggregation. The related indicator should be listed (by name and/or identifier). For indicator "Population, total" for example, one may inform the user that the indicator is also available (in other indicators) by sex, urban/rural, and age group (in indicator "Population, male" and "Population, female", etc.).
- **Language** (*language*) The language of the indicator/series metadata. For the language codes and names, the use of the ISO 639-2 standard is recommended.
- **Acronyms** (*acronyms*) The **acronyms** element is used to document the meaning of all acronyms used in the metadata of an indicator. If some acronyms are well known (like "GDP", or "IMF" for example), others may be less obvious or could be uncertain (does "PPP" mean "public-private partnership", or "purchasing power parity"?). In any case, providing a list of acronyms with their meaning will help users and make your metadata more discoverable. Note that acronyms should not include country codes used in the documentation of the geographic coverage of the data.
 - **Acronym** (*acronym*) An acronym referenced in the indicator metadata (e.g. "GDP").
 - **Expansion** (*expansion*) The expansion of the acronym, i.e. the full name or title that it represents (e.g., "Gross Domestic Product").
 - **Occurrence** (*occurrence*) This numeric element can be used to indicate the number of times the acronym is mentioned in the metadata. The element will rarely be used.
- **Related indicators** (*related_indicators*) This element allows to reference indicators that are often associated with the indicator being documented.
 - **Code** (*code*) The code for the indicator that is referenced in the document. It will likely be an ID that is used by that indicator.
 - **Label** (*label*) The name or label of the indicator that is associated with the indicator being documented.
 - **URL** (*uri*) A link to the related indicator.
- **Indicator group** (*series_group*) The group(s) the indicator belongs to. Groups can be created to organize indicators by theme, producer, or other.
 - **Group name** (*name*) The name of the group.
 - **Description** (*description*) A brief description of the group.
 - **Version** (*version*) The version of the grouping.
 - **URL** (*uri*) A link to a public website site where information on the grouping can be obtained.
- **Notes** (*notes*) This element is open and reserved for explanatory notes deemed useful to the users of the data. Notes should account for additional information that might help: replicate the indicator; access the data and research area, or discoverability in general.
 - **note** (*note*) The note itself.

ACCESS AND USE

- **License** (*license*) The license refers to the accessibility and terms of use associated with the data. Providing a license and a link to the terms of the license allows data users to determine, with full clarity, what they can and cannot do with the data.

- **Name** (*name*) The name of the license, e.g. "Creative Commons Attribution 4.0 International license (CC-BY 4.0)".
- **URL** (*uri*) The URL of a website where the licensed is described in detail, for example "<https://creativecommons.org/licenses/by/4.0/>".
- **Note** (*note*) Any additional information on the license.
- **Confidentiality statement** (*confidentiality*) A statement of confidentiality for the indicator.
- **Confidentiality status** (*confidentiality_status*) This indicates a confidentiality status for the indicator. A controlled vocabulary should be used with possible options "public", "official use only", "confidential", "strictly confidential". When all indicators are made publicly available, and belong to a database that has an open or public access policy, this element can be ignored.
- **Confidentiality note** (*confidentiality_note*) This element is reserved for additional notes regarding confidentiality of the data. This could involve references to specific laws and circumstances regarding the use of data.
- **Citation requirement** (*citation_requirement*) This element is used to provide information on how the data should be cited. This can include the preferred citation format, the name of the author, the title of the data, the date of publication, the version of the data, and the URL of the data.
- **Links** (*links*) This element provides links to online resources of any type that could be useful to the data users. This can be links to description of methods and reference documents, analytics tools, visualizations, data sources, or other.
 - **Type** (*type*) This element allows to classify the link that is provided.
 - **Description** (*description*) A description of the link that is provided.
 - **URL** (*uri*) The uri (URL) to the described resource.
- **Contacts** (*contacts*) The **contacts** element provides the public interface for questions associated with the production of the indicator.
 - **Name** (*name*) The name of the contact person that should be contacted. Instead of the name of an individual (which would be subject to change and require frequent update of the metadata), a title can be provided here (e.g. "data helpdesk").
 - **Role** (*role*) The specific role of the contact person mentioned in **name**. This will be used when multiple contacts are listed, and is intended to help users direct their questions and requests to the right contact person.
 - **Affiliation** (*affiliation*) The organization or affiliation of the contact person mentioned in **name**.
 - **Email** (*email*) The email address of the person or organization mentioned in **name**. Avoid using personal email accounts; the use of an anonymous email is recommended (e.g, "helpdesk@....org")
 - **Phone** (*telephone*) The phone number of the person or organization mentioned in **name**.
 - **URL** (*uri*) The URI of the agency (typically, a URL to a "Contact us" web page).
- **API documentation** (*api_documentation*) Increasingly, data are made accessible via Application Programming Interfaces (APIs). The API associated with an indicator must be documented. The documentation will usually not be specific to an indicator, but apply to all indicators in a same database.
 - **Description** (*description*) This element will not contain the API documentation itself, but information on what documentation is available.
 - **URL** (*uri*) The URL of the API documentation.
- **Periodicity** (*periodicity*) The periodicity of the series. It is recommended to use a controlled vocabulary with values like annual, quarterly, monthly, daily, ad-hoc, etc.

VERSION STATEMENT

- **Version** (*version*)

- **Version date** (`version_date`)
- **Version notes** (`version_notes`)
- **Version responsibility** (`version_resp`)

Use of AI to enhance descriptive metadata

You may take advantage of generative AI tools like ChatGPT or equivalent to suggest content for selected fields. For example, generative AI can help:

- Suggest **keywords** (ask ChatGPT to "Please suggest 20 keywords related to an indicator titled *(enter the title of the indicator here)* defined as *(enter the definition of the indicator here)*".)
- Propose a **definition** (ask ChatGPT to either *Suggest a definition for a statistical indicator titled (...)* or to *Improve the following definition of a statistical indicator titled (title) and currently defined as (definition)*.)
- Propose a description of the indicator's **relevance** (ask ChatGPT to either *Please describe the relevance of a statistical indicator titled (...) and defined as (definition)*. *In the description, indicate to whom and for what purposes the indicator is relevant.* Do not blindly accept suggestions formulated by AI models. Carefully review and improve (or reject) the proposed content.)

Data structure (structural metadata)

The data structure definition (DSD) allows you to describe how the data are organized in the data file. This is the structural metadata for the indicator. The DSD in the Metadata Editor follows the SDMX standard.

In a data file, the indicator is supposed to be stored in long (not wide) format, i.e. in a database type of format where each row corresponds to one observation value. Each row must contain the core information required to define what the observation value represents, complemented by additional (optional) information.

Typically, a data file for an indicator will look like one of the two options below, which use the population by sex of two countries as an example.

- **Option 1.** In option 1, the population by sex is provided as three different indicators: "Population, total", "Population, male", and "Population, female". The data file may look like this:

| <code>country_name</code> | <code>country_code</code> | <code>name</code> | <code>indicatorID</code> | <code>year</code> | <code>value</code> | <code>source</code> | <code>note</code> |
|---------------------------|---------------------------|--------------------|--------------------------|-------------------|--------------------|---------------------|-----------------------------------|
| Malawi | MWI | Population, total | SP_POP_TOTL | 2022 | 20568728 | NSO | Estimate; last census was in 2018 |
| Malawi | MWI | Population, female | SP_POP_TOTL_FE_IN | 2022 | 10537752 | NSO | Estimate; last census was in 2019 |
| Malawi | MWI | Population, male | SP_POP_TOTL_MA_IN | 2022 | 10030976 | NSO | Estimate; last census was in 2020 |
| Malawi | MWI | Population, total | SP_POP_TOTL | 2023 | 21104482 | NSO | Estimate; last census was in 2021 |
| Malawi | MWI | Population, female | SP_POP_TOTL_FE_IN | 2023 | 10809386 | NSO | Estimate; last census was in 2022 |
| Malawi | MWI | Population, male | SP_POP_TOTL_MA_IN | 2023 | 10295096 | NSO | Estimate; last census was in 2023 |
| Mauritania | MRT | Population, total | SP_POP_TOTL | 2022 | 4875637 | ANSADE | Estimate; last census was in 2013 |
| Mauritania | MRT | Population, female | SP_POP_TOTL_FE_IN | 2022 | 2487363 | ANSADE | Estimate; last census was in 2013 |
| Mauritania | MRT | Population, male | SP_POP_TOTL_MA_IN | 2022 | 2388274 | ANSADE | Estimate; last census was in 2013 |
| Mauritania | MRT | Population, total | SP_POP_TOTL | 2023 | 5022441 | ANSADE | Estimate; last census was in 2013 |
| Mauritania | MRT | Population, female | SP_POP_TOTL_FE_IN | 2023 | 2560475 | ANSADE | Estimate; last census was in 2013 |
| Mauritania | MRT | Population, male | SP_POP_TOTL_MA_IN | 2023 | 2461966 | ANSADE | Estimate; last census was in 2013 |

- **Option 2.** In option 2, the information is provided as a single indicator ("Population" with a dimension "sex" with values M, F or T for Male, Female, and Total).

| country_name | country_code | name | indicatorID | year | sex | value | source | note |
|--------------|--------------|------------|-------------|------|-----|----------|--------|-----------------------------------|
| Malawi | MWI | Population | SP_POP | 2022 | T | 20568728 | NSO | Estimate; last census was in 2018 |
| Malawi | MWI | Population | SP_POP | 2022 | F | 10537752 | NSO | Estimate; last census was in 2019 |
| Malawi | MWI | Population | SP_POP | 2022 | M | 10030976 | NSO | Estimate; last census was in 2020 |
| Malawi | MWI | Population | SP_POP | 2023 | T | 21104482 | NSO | Estimate; last census was in 2021 |
| Malawi | MWI | Population | SP_POP | 2023 | F | 10809386 | NSO | Estimate; last census was in 2022 |
| Malawi | MWI | Population | SP_POP | 2023 | M | 10295096 | NSO | Estimate; last census was in 2023 |
| Mauritania | MRT | Population | SP_POP | 2022 | T | 4875637 | ANSADE | Estimate; last census was in 2013 |
| Mauritania | MRT | Population | SP_POP | 2022 | F | 2487363 | ANSADE | Estimate; last census was in 2013 |
| Mauritania | MRT | Population | SP_POP | 2022 | M | 2388274 | ANSADE | Estimate; last census was in 2013 |
| Mauritania | MRT | Population | SP_POP | 2023 | T | 5022441 | ANSADE | Estimate; last census was in 2013 |
| Mauritania | MRT | Population | SP_POP | 2023 | F | 2560475 | ANSADE | Estimate; last census was in 2013 |
| Mauritania | MRT | Population | SP_POP | 2023 | M | 2461966 | ANSADE | Estimate; last census was in 2013 |

Organizations make their own decisions on how to organize their indicators. Using dimensions has the advantage of reducing the number of indicators. But it makes data discovery (indexing and search in data catalogs) somewhat more complex. In some cases, the use of dimensions is the only practical option. For example, data on population by age group (with 9 age groups + total), sex (2 options + total), and urban/rural (2 options with total) would require $10 \times 3 \times 3 = 90$ different indicators. Maintaining a single indicator "population" with 3 dimensions would be more efficient (one indicator can have multiple dimensions).

Properly documenting the data structure of an indicator provides users (and machines) with the information they need to query and use the data.

The screenshot shows the Metadata Editor interface with the following details:

- Header:** Metadata Editor, About, English, John Doe.
- Left sidebar:** Required, Recommended, Empty, Search bar, Home, Metadata information, Indicator description, DataCite, Tags, Data structure definition (selected), Data columns (sub-selected), Provenance, External resources, Administrative metadata.
- Main content:**
 - Title:** Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)
 - Data columns:**
 - 1 - Data columns:** indicator_id (Indicator identifier)
 - Name of the column:** indicator_id
 - Label:** Indicator identifier
 - Description of the column:** Unique identifier of the indicator in the ABC123 database
 - Data type:** String
 - Column type (SDMX):** Indicator ID
 - Time period format:** (empty)
 - Codelist:** (empty)

A data structure definition consists of providing the following information about the indicator and the way it is organized in the data file. This information will be provided for each column of the data file:

- Name:** The column name
- Label:** The column label
- Description:** A brief description of the column

- **Data type:** The type of variable, with the following possible values: String, Integer, Float, Date, and Boolean.
- **Column type:** The type of column, with the following options:
 - **Dimension:** In our Option 2 example, column "sex" is a dimension.
 - **Time period:** The column indicates the time period to which the observation value applies. In our Option 2 example, "year" is a time period.
- **Attribute:**
- **Indicator ID:** The column is the indicator unique identifier. Only one column can be an Indicator ID. In our Option 2 example, "IndicatorID" is the Indicator ID.
- **Indicator name:** The column is the name (or title) of the indicator. In our Option 2 example, "Name" is the indicator name.
- **Annotation:**
- **Geography:** The geographic area to which the value corresponds. In our example, column "country" is the geography.
- **Observation value:** The observation value (the "data" itself). In our example, column "value" is the observation value.
- **Periodicity:**
- **Time period format:** Time period format is used to indicate the format of the date in the column identified as Time period. In our Option 2 example, Time period format is "YYYY" as we have data by year.
- **Codelist:**
- **Codelist reference:**

All columns in the data file will be documented.

Poverty headcount ratio at \$2.15 a day (2017 PPP) (% of population)

Data columns

| | | | | |
|------------------|----------------|-------------------|---------------------------|--------|
| 1 - Data columns | IndicatorID | indicator_id | Indicator identifier ... | Remove |
| 2 - Data columns | indicator_name | indicator_name | Indicator n... | Remove |
| 3 - Data columns | country | Country or region | Code of the coun... | Remove |
| 4 - Data columns | year | Year | integer time_period | Remove |
| 5 - Data columns | value | Observation value | float observation_v... | Remove |
| 6 - Data columns | sex | Sex | Sex code string dimension | Remove |

Add section - Data columns

Data notes

DataCite

See section "Documenting - General instructions".

Provenance

The **Provenance** container is used to document how and when the dataset was acquired. It is used to ensure traceability. See section "Documenting - General instructions".

Tags

See section "Documenting - General instructions".

Additional

External resources

External resources are all materials (and links) that relate to the indicator. This may include documents on methodology, scripts, photos and videos, and any other resource available in digital format. These materials and links are added to the documentation of an indicator in the External resources container. Click on **External resources** in the navigation tree, then on CREATE RESOURCE. Enter the relevant information on the resource (at least a title), then provide either a filename (the file will then be uploaded on the server that hosts the Metadata Editor) or a URL to the resource.

External resources that have already been created for another project can also be imported. To do that, they must first be exported as JSON or RDF from the other project. The click on IMPORT in the External resources page, and select the file.

External resources will be part of the project ZIP package (when the ZIP package is generated - See the main menu).

Export and publish

Database-level metadata

Not to be documented for each indicator. Document once, give it an ID, and enter it in each indicator in field *databaseID*. When published in NADA, this will create a one-to-many relationship.

Create a new project

The first step in documenting a database is to create a new project. You do that by clicking on **CREATE NEW PROJECT** in the *My projects* page, then selecting *Database* as data type when prompted. This will open a new, untitled *Project page*.

In that page, **select the template** you want to use to document the database. A default template is proposed; no action is needed if you want to use the default template. Otherwise, switch to another template by clicking on the template name in the *Templates* frame. Note that you can at any time change the template used for the documentation of a project. The selected template will determine what you see in the navigation tree and in the metadata entry pages, but switching from one template to another will not impact the metadata that has already been entered; no information will be deleted from the metadata.

Enter information on metadata

The *Metadata information* section in the navigation tree (in the *Project page*) contains elements intended to document the metadata being generated, i.e., metadata about the metadata. All content in this section is optional; it is however recommended practice to document the metadata as precisely as possible. This information will not be useful to data

users, but it will be to catalog administrators. When metadata is shared across catalogs, the information entered in the *Information on metadata* provides transparency and clarity on the origin of the metadata.

Enter a description of the database (descriptive metadata)

Export and publish metadata

[Save as ZIP package](#) [Export metadata in different formats](#) [Publish to NADA](#)

Documenting geographic datasets and services

The ISO 19100 metadata standards

To make geographic information discoverable and to facilitate their dissemination and use, the ISO Technical Committee on Geographic Information/Geomatics (ISO/TC211) created a set of metadata standards to describe geographic **datasets** (ISO 19115-1 for vector datasets and ISO 19115-2 for raster datasets) and geographic **data services** (ISO 19119).

The ISO 19115-2 provides the necessary metadata elements to describe the structure of raster data. The ISO 19115-1 standard does not provide all necessary metadata elements needed to describe the structure of vector datasets. The description of data structures for vector data (also referred to as *feature types*) is therefore often omitted. The ISO 19110 standard solves that issue, by providing the means to document the structure of vector datasets (column names and definitions, codes and value labels, measurement units, etc.), which will contribute to making the data more discoverable and usable.

This set of standards, known as the ISO 19100 series, have been "unified" into a common XML specification (ISO 19139).

These standards served as the cornerstone of multiple initiatives to improve the documentation and management of geographic information such as the [Open Geospatial Consortium \(OGC\)](#), the [US Federal Geographic Data Committee \(FDGC\)](#), the [European INSPIRE directive](#), or more recently the [Research Data Alliance \(RDA\)](#), among others.

The ISO 19100 standards have been designed to cover the large scope of geographic information. The level of detail they provide goes beyond the needs of most data curators. What we use in the Metadata Editor is a subset of the standards, which focuses on what we consider as the most relevant metadata elements to describe and catalog geographic datasets and services.

Documenting geographic datasets, series, and services

Geographic datasets

Geographic datasets Geographic datasets refers to the actual stored data about the Earth's features, phenomena, or events. Geographic datasets "identify and depict geographic locations, boundaries and characteristics of features on the surface of the earth. Geographic datasets can be vector data (points, lines, polygons) or raster data (grids, pixels, imagery). They include geographic coordinates (e.g., latitude and longitude) and data associated to geographic locations (...)" (Source: <https://www.fws.gov/gis/>) The ISO 19115 standard defines the structure and content of the metadata to be used to document geographic datasets. The ISO 19115 standard is split into two parts covering:

- **vector data** (ISO 19115-1), and
- **raster data** including imagery and gridded data (ISO 19115-2). The elements of ISO 19115 are included in the XML specification ISO 19139.

Vector and raster spatial datasets are built with different structures and formats. The following summarizes how these two categories differ and how they can be processed using the R software. The descriptions of vector and raster data provided in this chapter are adapted from:

- <https://gisgeography.com/spatial-data-types-vector-raster/>
- <https://datacarpentry.org/organization-geospatial/02-intro-vector-data/index.html>

Vector data

Vector data are comprised of **points**, **lines**, and **polygons** (areas).

A vector **point** is defined by a single x, y coordinate. Generally, vector points are a latitude and longitude with a spatial reference frame. A point can for example represent the location of a building or facility. When multiple dots are connected in a set order, they become a vector **line** with each dot representing a **vertex**. Lines usually represent features that are linear in nature, like roads and rivers. Each bend in the line represents a vertex that has a defined x, y location. When a set of 3 or more vertices is joined in a particular order and closed (i.e. the first and last coordinate pairs are the same), it becomes a **polygon**. Polygons are used to show boundaries. They will typically represent lakes, oceans, countries and their administrative subdivisions (provinces, states, districts), building footprints, or outline of survey plots. Polygons have an area (which will correspond to the square-footage for a building footprint, to the acreage for an agricultural plot, etc.)

Vector data are often provided in one of the following file formats:

- ESRI Shapefile (actually a zip set of files; not standard and limited as it is based on an outdated DBF format, but still widely used);
- ESRI GeoDatabase file (not a standard format, but widely used);
- GML: the Official OGC geospatial standard format, used by standard spatial data services;
- GeoPackage: the OGC recommended standard for handling vector data;
- GeoJSON: another OGC standard, often used when a service is associated to the data;
- KML/KMZ: [Keyhole Markup Language](#), an XML notation for expressing geographic annotation and visualization within two-dimensional maps and three-dimensional Earth browsers;
- CSV file: Comma-separated values files, with geometries provided in OGC Well-Known-Text (WKT);
- OSM: An XML-formatted file containing "nodes" (points), "ways" (connections), and "relations" from [OpenStreetMap](#) format.

The figure below provides an example of vector data extracted from [Open Street Map](#) for a part of the city of Thimphu, Bhutan (as of 17 May, 2021).



Raster data

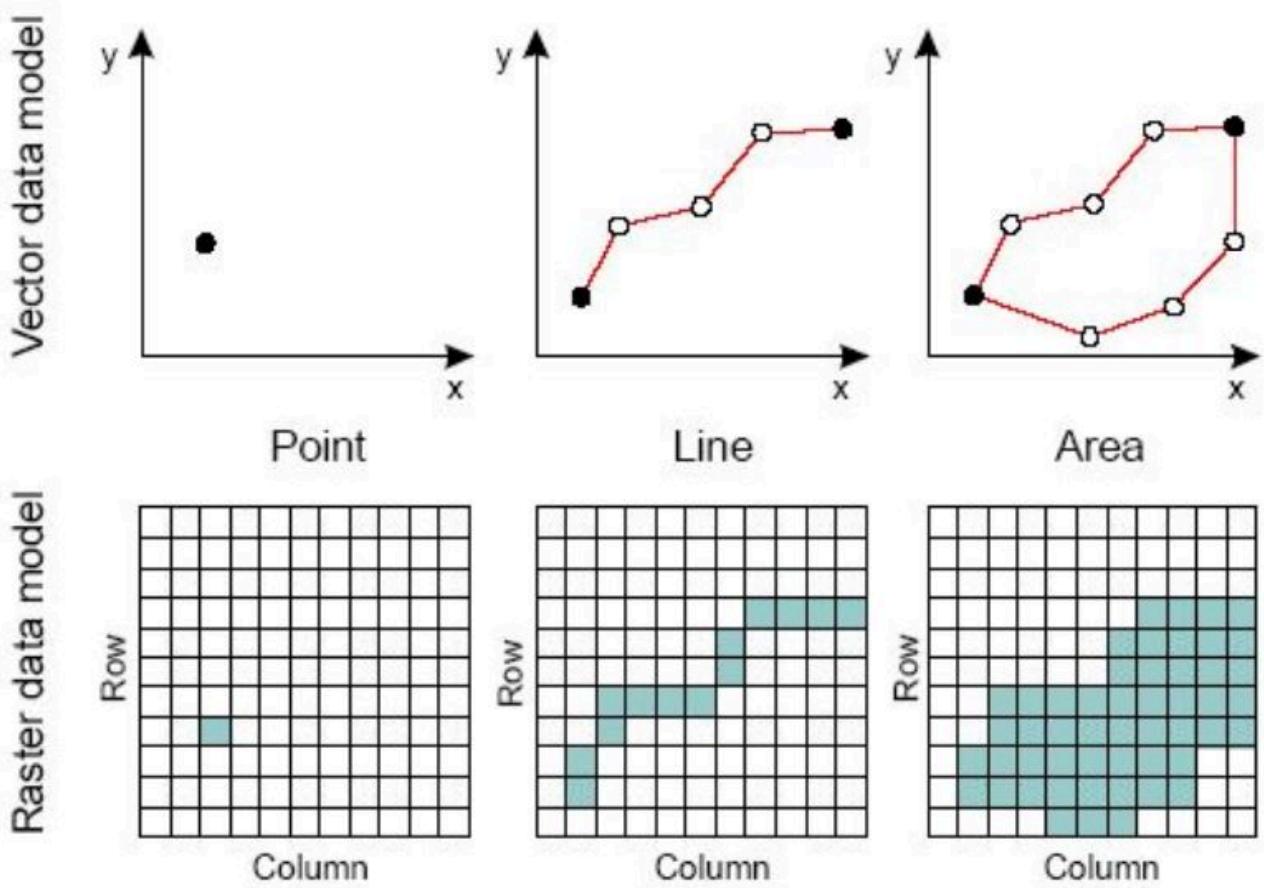
Raster data are made up of pixels, also referred to as *grid cells*. Satellite imagery and other remote sensing data are raster datasets. Grid cells in raster data are usually (but not necessarily) regularly-spaced and square. Data stored in a raster format is arranged in a grid without storing the coordinates of each cell (pixel). The coordinates of the corner points and the spacing of the grid can be used to calculate (rather than to store) the coordinates of each location in a grid.

Any given pixel in a grid stores one or more values (in one or more bands). In geographic raster datasets, ***bands*** represent separate layers of data values within a raster. Each band corresponds to a matrix of values for the same spatial extent, often representing measurements in different spectral ranges or thematic data layers. For example:

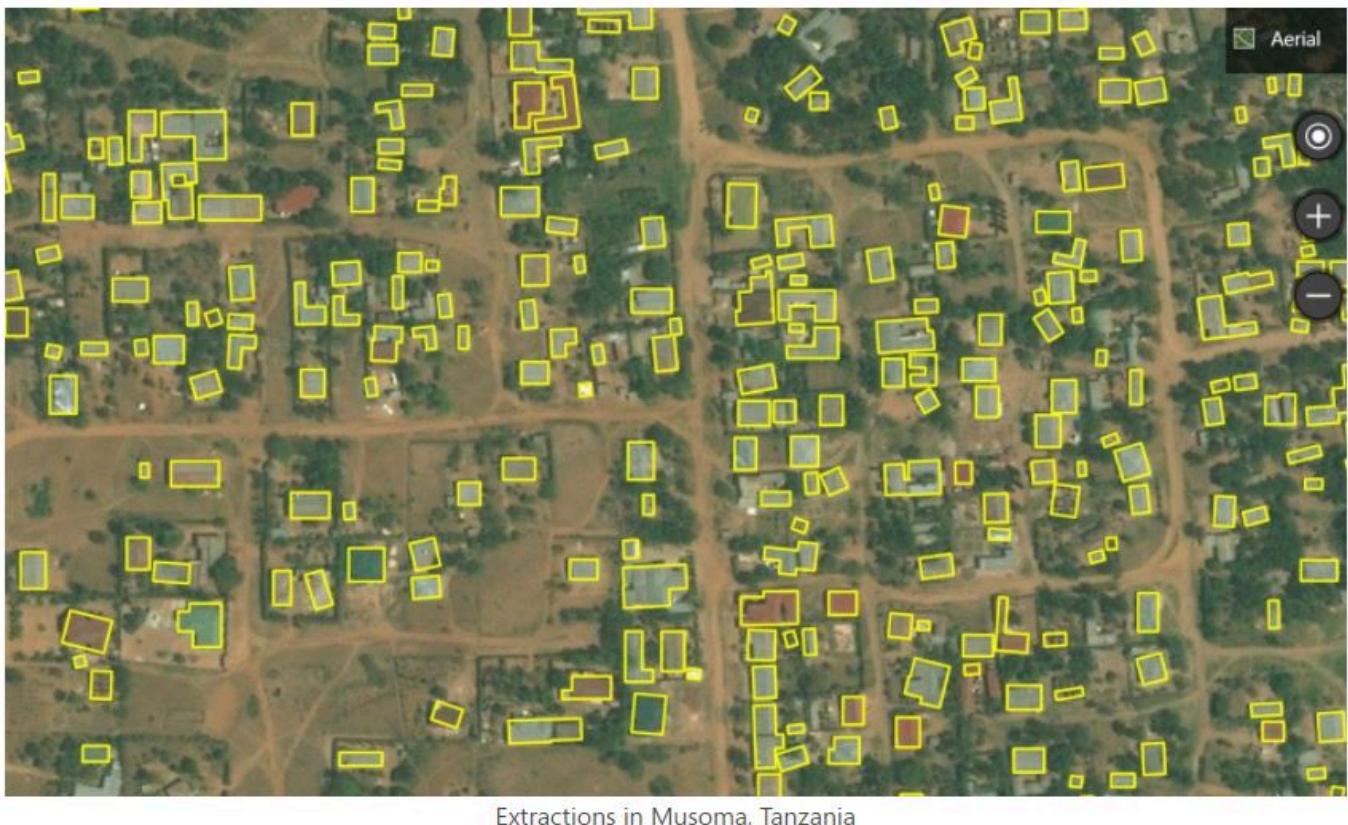
- In satellite imagery, one band may capture red light, another green, another blue, and others might capture infrared or thermal information.
 - In a digital elevation model, a single band might contain elevation values only.
 - In a multi-variable raster dataset, different bands may represent different variables (e.g., temperature, salinity).

Raster data can be **discrete** or **continuous**. Discrete rasters have distinct themes or categories. For example, one grid cell can represent a land cover class, or a soil type. In a discrete raster, each thematic class can be discretely defined (usually represented by an integer) and distinguished from other classes. In other words, each cell is definable and its value applies to the entire area of the cell. For example, the value 1 for a class might indicate "urban area", value 2 "forest", and value 3 "others". Continuous (or non-discrete) rasters are grid cells with gradual changing values, which could for example represent elevation, temperature, or an aerial photograph.

The difference between vector and raster data, and between different types of vectors, is clearly illustrated in the figure below taken from the World Bank's [Light Every Night GitHub repository](#).



Raster data are sometimes converted into vector data. For example, a building footprint layer (vector data, composed of polygons) can be derived from a satellite image (raster data). Such conversions can be implemented in a largely automated manner using machine learning algorithms.



Source: <https://blogs.bing.com/maps/2019-09/microsoft-releases-18M-building-footprints-in-uganda-and-tanzania-to-enable-ai-assisted-mapping>

Raster data are often provided in one of the following file formats:

- GeoTIFF (standard): Most of the remote sensing data are stored as GeoTIFF files.
<https://www.ogc.org/standards/geotiff>
- NetCDF (standard) https://www.unidata.ucar.edu/software/netcdf/docs/netcdf_introduction.html
- ECW: [https://en.wikipedia.org/wiki/ECW_\(file_format\)](https://en.wikipedia.org/wiki/ECW_(file_format))
- JPEG 2000: https://fr.wikipedia.org/wiki/JPEG_2000
- MrSid: <https://en.wikipedia.org/wiki/MrSID>
- ArcGrid (ESRI Grid format)

GeoTIFF is a popular file format for raster data. A *Tagged Image File Format* (TIFF or TIF) is a file format designed to store raster-type data. A GeoTIFF file is a TIFF file that contains specific tags to store structured geospatial metadata including:

- Spatial extent: the area coverage of the file
- Coordinate reference system: the projection / coordinate reference system used
- Resolution: the spatial extent of each pixel (spatial resolution)
- Number of layers: number of layers or bands available in the file

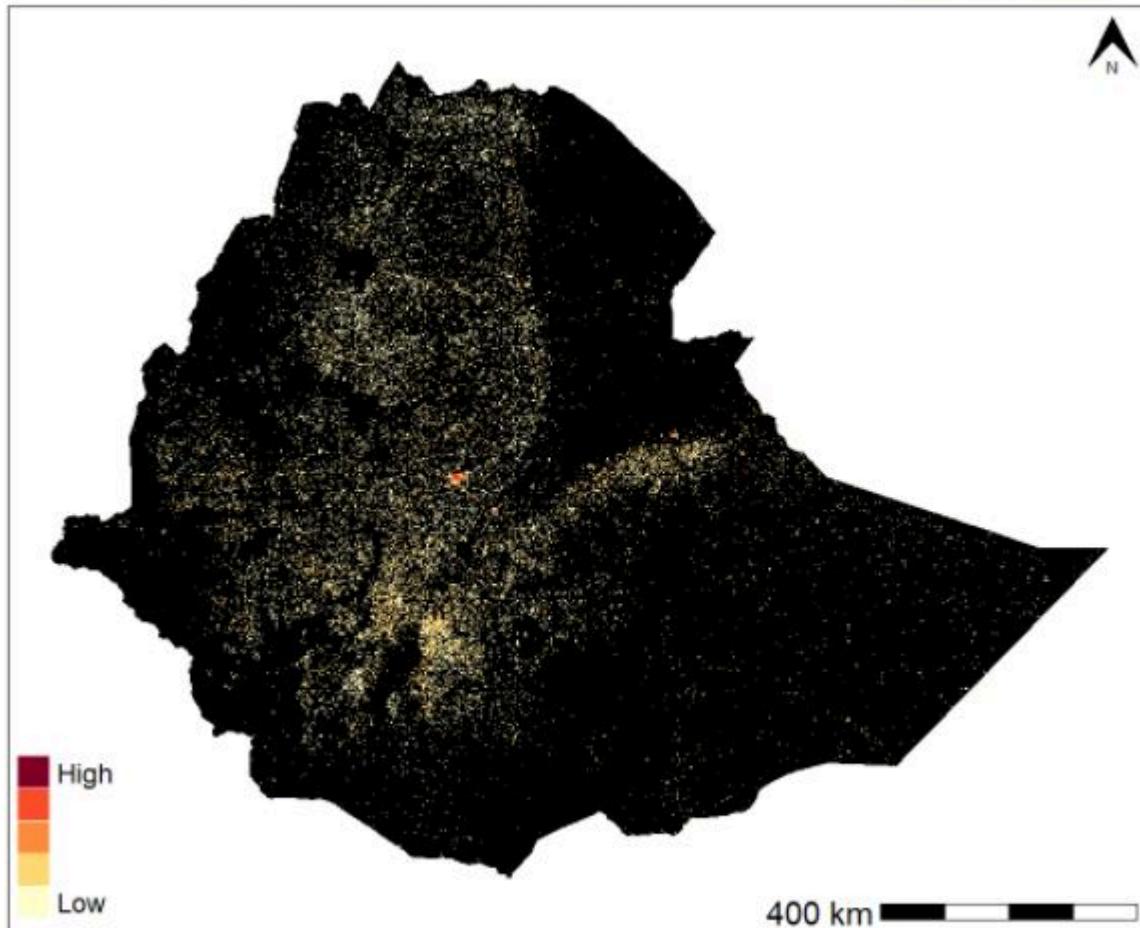
TIFF files can be read using (among other options) the R package [raster](#) or the Python library [rasterio](#).

GeoTIFF files can also be provided as **Cloud Optimized GeoTIFFS (COGs)**. In COGs, the data are structured in a way that allows them to be shared via web services which allow users to query, visualize, or download a user-defined subset of the content of the file, without having to download the entire file. This option can be a major advantage, as geoTIFF files generated by remote sensing/satellite imagery can be very large. Extracting only the relevant part of a file can save significant time and storage space.

The example below shows the spatial distribution of the Ethiopian population in 2020. The data file was downloaded from the [WorldPop](#) website on 17 May 2021.

Ethiopia population 2020

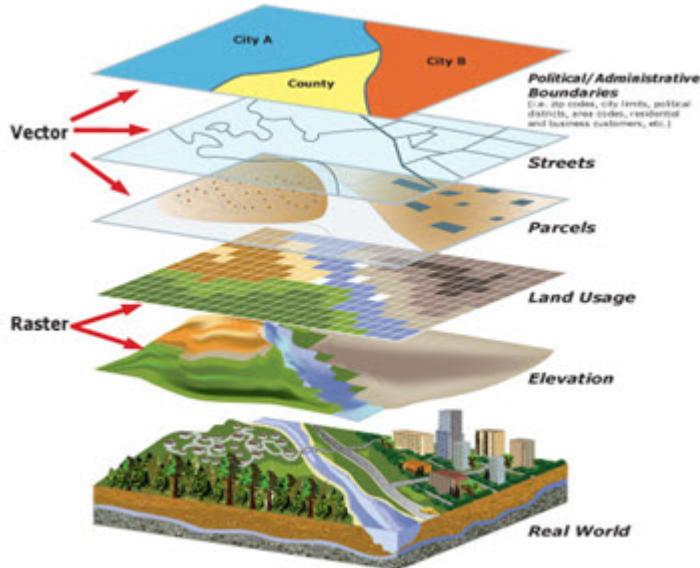
Estimated total number of people per grid-cell at a resolution of 3 arc seconds (approximately 100m at the equator)



WorldPop (worldpop.org - School of Geography and Environmental Science, University of Southampton
©2020 This work is licensed under a Creative Commons Attribution 4.0 International License

Combining vector and raster data in GIS applications

In GIS applications, vector and raster data are often combined into multi-layer datasets, as shown in the figure below extracted from the [County of San Bernardino \(US\) website](#).



Geographic datasets series

Geographic datasets series Geographic datasets can be organized in *series*. In ISO 19139 (and ISO 19115), *series* refers to a collection of related datasets that are grouped together because they share a common purpose, theme, method of production, or spatial/temporal extent. In other words, a series is a set of datasets that are logically connected and can be described together at a higher level.

Examples:

- A series of annual land use maps for different years (2000, 2005, 2010, etc.).
- A collection of satellite images covering different tiles of a country.
- A set of topographic maps produced at the same scale for different regions.

In the metadata, some elements in the ISO 19139 are dedicated to provide information on series. You can describe the series itself at a general level. Then you can have metadata for individual datasets linked to that series. This helps avoid duplication — common information is stored once at the series level, and dataset-level metadata can focus on the differences. In ISO 19139 XML, you will often see fields like:

- resourceScope set to series (to show you are describing a series, not a single dataset);
- aggregationInfo to link datasets to their series or services.

Geographic data services

Geographic data services refers to operations or set of operations that allows users to access, manipulate, transform, analyze, or visualize geographic data over a network or system. It's not the data itself — it's the functionality provided to interact with the data. In other words, a geographic data service is something you can call or use to work with geographic data without downloading the full dataset first. Geographic data services are documented using metadata elements from the ISO 19119 metadata standard. The elements of ISO 19119 are included in the XML specification ISO 19139.

Unified metadata specification - The ISO/TS 19139 standard

The three metadata standards previously described - ISO 19115 for vector and raster datasets, ISO 19110 for vector data structures, and ISO 19119 for data services, provide a set of concepts and definitions useful to describe the geographic information. To facilitate their practical implementation, a digital specification, which defines how this information is stored and organized in an electronic metadata file, is required. The ISO/TS 19139 standard, an XML specification of the ISO 19115/10110/19119/, was created for that purpose.

The ISO/TS 19139 is a standard used worldwide to describe geographic information. It is the backbone for the implementation of [INSPIRE](#) dataset and service metadata in the European Union. It is supported by a wide range of tools, including desktop applications like [Quantum GIS](#), [ESRI ArcGIS](#), and OGC-compliant metadata catalogs (e.g., [GeoNetwork](#)) and geographic servers (e.g., [GeoServer](#)).

ISO 19139-compliant metadata can be generated and edited using specialized metadata editors such as [CatMDEdit](#) or [QSphere](#), or using programmatic tools like Java Apache SIS or the R packages [geometa](#) and [geoflow](#), among others.

The Metadata Editor uses the ISO 19139 to provide a solution compatible with the ISO 19115, ISO 19110, and ISO 19119 standards.

Metadata templates or profiles for the practical implementation of ISO 19139

The ISO 19139 specification is complex. To enable and simplify its use in the Metadata Editor, we produced a JSON version of (part of) the standard. We selected the elements we considered most relevant for our purpose, and organized them into the JSON schema described below. For data curators with limited expertise in XML and geographic data documentation, this JSON schema will make the production of metadata compliant with the ISO 19139 standard easier.

Some organizations have sought to make the use of ISO 19139 manageable, by defining *templates* or *profiles* that consist of subsets. This includes the INSPIRE set of elements defined by the European Union, and the GEMINI set of elements defined by the United Kingdom.

The Metadata Editor is also provided with recommended templates that only contain the elements of the ISO 19139 that are considered the most useful for the documentation of geographic datasets and services. The template provided in the Metadata Editor contains all elements from the INSPIRE and GEMINI profiles, and a few more.

INSPIRE - Infrastructure for Spatial Information in Europe

The EU INSPIRE Directive (Infrastructure for Spatial Information in the European Community), adopted in 2007, establishes a legal framework for the creation of a unified spatial data infrastructure across the European Union. Its goal is to enable the sharing, discovery, and use of interoperable spatial data among public authorities, policymakers, and the public, particularly to support environmental policies and activities that may impact the environment. INSPIRE requires EU Member States to document, harmonize, and provide access to spatial datasets and services related to 34 environmental themes, including land use, transport networks, and biodiversity. The Directive mandates the use of standardized metadata (based on ISO 19115 and ISO 19139), common data specifications, and network services (discovery, view, download, etc.), ensuring that spatial information is easily searchable and accessible across borders and administrative levels.

See the INSPIRE knowledge base at https://knowledge-base.inspire.ec.europa.eu/index_en

UK GEMINI 2.3

The UK GEMINI (GEo-spatial Metadata INteroperability Initiative) standard is the United Kingdom's national metadata standard for describing geographic datasets and services. Developed by the Association for Geographic Information

(AGI) and maintained by the UK Metadata Working Group, GEMINI ensures consistent and interoperable metadata across UK public sector organizations. It is based on and fully compatible with international standards ISO 19115 for geographic information and ISO 19139 for XML encoding. GEMINI supports the requirements of the UK Location Programme and the European INSPIRE Directive, enabling efficient discovery, access, and use of spatial data through standardized metadata elements such as dataset identification, spatial and temporal extent, quality, access constraints, and responsible organizations.

See:

- https://guidance.data.gov.uk/publish_and_manage_data/harvest_or_add_data/harvest_data/gemini/#gemini-and-iso-19139-metadata
- A description of UK GEMINI 2.3 2020-04-07 dataset or series: <https://agiorguk.github.io/gemini/1062-gemini-datasets-and-data-series.html>

Documenting a geographic dataset

This section describes how to document a geographic dataset (vector or raster) using the Metadata Editor. Documenting a data service follows the exact same principles, except that the option to extract metadata from data files does not apply when documenting a data service.

The metadata template provided with the Metadata Editor identifies metadata elements that are specific to vector or raster datasets, or to data services.

Prepare your materials

As a data curator preparing a geographic dataset for documentation using the ISO 19139 standard, it is essential to ensure that the dataset and all supporting materials are complete, well-organized, and ready for standardized metadata creation. Below are core recommendations for preparing a dataset prior to its documentation:

1. Ensure dataset completeness and consistency

- Verify that the dataset is finalized, with complete records and no missing key attributes or geometries.
- Check for consistency in coordinate systems, units of measurement, and naming conventions across all layers or features.
- Ensure the dataset is in a recognized and interoperable format (e.g., GeoPackage, shapefile, GML, GeoJSON).

2. Confirm spatial reference information

- Identify and record the Coordinate Reference System (CRS) used, including the full name and EPSG code.
- If multiple CRSs are used, provide clear documentation and justification.
- Check the vertical reference system if vertical data is present.

3. Prepare descriptive and contextual information

- Write a clear and concise title and abstract for the dataset that accurately describes its content and purpose.
- Define the geographic extent (bounding box or polygon), temporal extent, and data collection period.
- Identify the data theme or subject area (e.g., hydrography, land cover, administrative boundaries).

4. Identify data lineage and quality information

- Document the origin of the data, including source datasets, data capture methods, and any processing steps applied.

- Record known limitations, accuracy, and quality assessments (e.g., positional accuracy, completeness).
- Note any validation or quality control procedures performed.

5. Organize supporting files and graphics

- Prepare any graphic overviews (e.g., thumbnail maps, sample visualizations) to be referenced in metadata.
- Assemble distribution files and note their formats and versions (e.g., ZIP, GeoTIFF, CSV).
- Collect legal documents, such as terms of use, licenses, and access conditions.

6. Identify responsible parties

- Record the organization or person responsible for creating, maintaining, and distributing the dataset.
- Include contact information with roles (e.g., point of contact, metadata author, distributor).

7. Determine access and use constraints

- Identify and clearly state any restrictions on access or limitations on use, including intellectual property rights.
- Prepare text for disclaimers, preferred citations, or user guidance.

8. Review related resources and linkages

- List any related datasets, services, or documents (e.g., methodology reports, feature catalogues).
- Collect unique identifiers, such as UUIDs, for related resources when available.

9. Validate the dataset's technical readiness

- Ensure file names and field names follow naming conventions (e.g., no special characters, reasonable length).
- Verify that all attribute fields are well-documented, with field names, definitions, and value domains.
- Check that feature types are defined consistently and that geometry types are valid and appropriate.

10. Maintain versioning and update information

- Record the dataset version, date of last update, and frequency of updates.
- Keep track of historical versions if applicable, for transparency and reproducibility.

11. Generate images for the Graphic Overview section

- Create at least one graphic overview image that visually summarizes the dataset (e.g., a map showing coverage or key features).
- Ensure the image is clear, appropriately scaled, and provides useful context for users.
- Include a descriptive file name, brief caption, and file format (e.g., PNG, JPEG).
- Save the image in a standard resolution and aspect ratio suitable for web display and embedding in metadata tools.

Create a project

The first step in documenting a dataset is to create a new project. You do that by clicking on **CREATE NEW PROJECT** in the *My projects* page. Select *Geospatial* as data type. This will open a new, untitled project *Home* page.

In the **Templates** frame, select the template you want to use to document the dataset. A default template is proposed; no action is needed if you want to use that template. Otherwise, switch to another template by clicking on the template name. Note that you can at any time change the template used for the documentation of a project. The selected template will determine what you see in the navigation tree and in the metadata entry pages.

Switching from one template to another will not impact the metadata that has already been entered; no information will be deleted from the metadata.

Once a project has been created, you can import the data files (if available) and start documenting the dataset.

Extract metadata from data files [under development - to be available in version 1.1 of the application]

The Metadata Editor allows you to extract metadata contained in geographic datasets. The following formats are supported:

- Vector datasets: GeoJSON, Shapefile (SHP), KML, GPKG, GDB
- Raster datasets: NetCDF (NC), geoTIF, BIL, ASCII (XYZ), GeoPDF (PDF), JPG, GIF, ADF, OVR

To extract information from data files:

- Click on DATA in the navigation bar. The data import page will be displayed. Select the data file(s) to be imported, then click **IMPORT**.

The metadata that will be extracted will be automatically entered in the relevant metadata elements. What can be extracted depends on the data format, and on what the data producer may have included in the data files. Typically, the following metadata will be extracted: Filename, Bounding boxes, Reference system, Features (for vector datasets), Bands, and more.

Enter additional metadata

We describe below the metadata elements included in the recommended INSPIRE + GEMINI + ADDITIONAL ELEMENTS template. This is only a subset of the elements contained in the ISO 19139. If you developed or imported a different, more comprehensive template, consult the description of metadata elements provided by the ISO 19139 documentation.

DOCUMENT DESCRIPTION

This section is not specific to geographic datasets. It corresponds to the *Document description* section of the DDI Codebook metadata standard (for microdata) and to the *Information on metadata* section in other metadata standards and schemas. This section contains metadata on the metadata, structured in a format consistent across metadata standards supported by the Metadata Editor. The content of this section is mainly intended to be used by catalog administrators, and will not be exported to ISO 19139 metadata files.

DESCRIPTION

METADATA

- **Hierarchy level** This is the type of resource being described by the metadata and it is filled in with a value from a classification of the resource based on its scope. The choice of Resource Type will be probably the first decision made by the user and it will define the metadata elements that should be filled.

The hierarchy level defines the scope of the resource. It indicates whether the resource is a collection, a dataset, a series, a service, or another type of resource. The ISO 19139 provides a controlled vocabulary for this element. It is recommended but not mandatory to make use of it. The most relevant levels for the purpose of cataloguing geographic data and services are dataset (for both raster and vector data), service (a capability which a service provider entity makes available to a service user entity through a set of interfaces that define a behavior), and series. Series will be used when the data represent an ordered succession, in time or in space; this will typically apply to time series, but it can also be used to describe other types of series (e.g., a series of ocean water temperatures collected at a succession of depths). Note that:

- dataset: is an identifiable data that can be accessed separately. A dataset can be a part of a whole (series) or a segregate resource.
- series: is a collection of resources or related datasets that share the same product specification.
- service: technologies providing availability and access to spatial information, for example, web map services, web feature services, web coverage services, web processing services, catalogue web services, etc.
- **Primary ID** The "Primary ID" (also referred to as IDNO) is a unique identification number used to identify the study (geographic dataset or service). A unique identifier is required for cataloguing purpose, so this element is declared as "Required". The identifier will allow users to cite the study properly. The identifier must be unique within the catalog. Ideally, it should also be globally unique; the recommended option is to obtain a Digital Object Identifier (DOI) for the study. Alternatively, the "Primary ID" can be constructed by an organization using a consistent scheme. The identifier should not contain blank spaces.
- **Parent identifier** (*for hierarchy level = series, not for datasets*) A geographic data resource can be a subset of a larger dataset. For example, an aquatic species distribution map can be part of a data collection covering all species, or the 2010 population census dataset of a country can be part of a dataset that includes all population censuses for that country since 1900. In such case, the parent identifier metadata element can be used to identify this higher-level resource. As for the fileIdentifier, the parentIdentifier must be a unique identifier persistent in time. In a data catalog, a parentIdentifier will allow the user to move from one dataset to another. The parentIdentifier is generally applied to datasets, although it may in some cases be used in data services descriptions.
- **Metadata date** Date and time when the metadata record was created or updated. Requires an extended ISO 8601 formatted combined UTC date and time string (2009-11-17T10:00:00).
- **Metadata language** Main language used in the metadata description. It is recommended to select a value from a controlled vocabulary, for example that provided by ISO 639-2.
- **Metadata contact**
 - **Individual name** The responsible party (person) in charge of the metadata production.
 - **Organisation name** The responsible party (organization) in charge of the metadata production.
 - **Email** Enter the email address of the contact person for this metadata. To ensure continuity and long-term accessibility, avoid using personal email addresses. Use a role-based or institutional email account (e.g., help@myorganization.org) that remains valid even if individual staff members change.
 - **Phone** Enter the phone number for contacting the person or team responsible for the metadata. To ensure continuity and accessibility, avoid using personal or direct mobile numbers. Instead, provide a general or role-based contact number (e.g., a departmental line or help desk number) that will remain valid even if individual staff members change.
 - **Address**
 - **Delivery point** Physical address - Street, building number, etc.
 - **City** Physical address - City name
 - **Postal code** Physical address - Postal code
 - **Country** Physical address - Country name
 - **Online resource**
 - **Name** Name of the online resource.
 - **Description** Description of the online resource
 - **URL** URL of the online resource.
 - **Metadata standard name** The name of the geographic metadata standard used to describe the resource. The recommended values are:
 - In the case of vector dataset metadata: ISO 19115 Geographic information - Metadata

- In the case of grid/imagery dataset metadata: ISO 19115-2 Geographic Information - Metadata Part 2 Extensions for imagery and gridded data
- In the case of service metadata: ISO 19119 Geographic information - Services
- **Metadata standard version** The version of the metadata standard being used. It is good practice to enter the standard's inception/revision year. ISO standards are revised with an average periodicity of 10-year. Although the ISO TC211 geographic information metadata standards have been reviewed, it is still accepted to refer to the original version of the standard as many information systems/catalogs still make use of that version. The recommended values are:
 - In the case of vector dataset metadata: ISO 19115:2003
 - In the case of grid/imagery dataset metadata: ISO 19115-2:2009
 - In the case of service metadata: ISO 19119:2005
- **Dataset URI** A unique resource identifier for the dataset, such as a web link that uniquely identifies the dataset. The use of a Digital Object Identifier (DOI) is recommended.
- **Metadata maintenance and update frequency** The metadata maintenance and update frequency elements provide information on the maintenance of the metadata including the frequency of updates. This is a free text element. The information should be chosen from values recommended by the ISO 19139 controlled vocabulary with the following options: continual, daily, weekly, fortnightly, monthly, quarterly, biannually, annually, asNeeded, irregular, notPlanned, unknown.
- **Metadata character set** This element specifies the character encoding used in the metadata record itself — i.e., how the text in the metadata is encoded. The purpose is to ensure that software applications can correctly interpret and display the metadata text, especially in multilingual or international contexts.
 - **Character set code** This is the value element within the Metadata character set field. It contains the actual code (or abbreviation) representing the character set. It contains a code from a predefined list of values (controlled vocabulary), for example: "utf8".
 - **Codelist used** Refers to the controlled vocabulary or authoritative list from which the Character set code value is taken. The purpose is to provide a reference to the standardized list of valid character set codes, ensuring interoperability and consistent interpretation across systems.

IDENTIFICATION

DATASET IDENTIFICATION

These elements are used to clearly and uniquely identify a geographic dataset and establish its relationship to other resources. Proper identification is essential for users to understand the nature, scope, and context of the dataset.

- **Title** The primary name by which the dataset, series, or service is known. It should be unique, concise, and descriptive, reflecting the content and scope of the resource. The title helps users discover and distinguish the dataset from others. Example: "Land Use Map of Northern Ireland, 2022 Edition".
- **Alternate title** An alternative name used to refer to the dataset. This may include abbreviations, translated titles, project names, or legacy titles. Including alternate titles improves discoverability, especially when datasets are known by different names in different contexts or languages. Example: "NI_LU_2022", or "Carte de l'occupation du sol – Irlande du Nord, 2022".
- **Collective title** The title of a larger resource or collection of which this dataset is a part. This element is used when the dataset is a component of a series, such as a national spatial data infrastructure collection, a thematic series, or a multi-part publication. It helps group related datasets under a common collection title and supports hierarchical or thematic navigation. Example: "UK National Land Use Dataset Series".
- **Responsible party** The Responsible party element identifies the individual or organization that has a specific role in relation to the dataset being documented. This includes responsibilities such as authoring, publishing, maintaining, or distributing the dataset. The purposes are to: (i) ensure users know who to contact for more information, assistance,

or clarification; (ii) Enhance transparency, traceability, and accountability ; and (iii) Support data governance and helps with data stewardship and maintenance over time.

- **Individual name** The responsible party (person) in charge of the dataset or service.
- **Organisation name** The responsible party (organization) in charge of the dataset or service.
- **Email** Enter the email address of the contact person for this dataset. To ensure continuity and long-term accessibility, avoid using personal email addresses. Use a role-based or institutional email account (e.g., help@myorganization.org) that remains valid even if individual staff members change.
- **Phone** Enter the phone number for contacting the person or team responsible for the dataset or service. To ensure continuity and accessibility, avoid using personal or direct mobile numbers. Instead, provide a general or role-based contact number (e.g., a departmental line or help desk number) that will remain valid even if individual staff members change.
- **Address**
 - **Delivery point** Physical address - Street, building number, etc.
 - **City** Physical address - City name
 - **Postal code** Physical address - Postal code
 - **Country** Physical address - Country name
 - **Online resource**
 - **Name** Name of the online resource. In case of a geographic standard data services, this should be filled with the identifier of the resource as published in the service. Example, for an OGC Web Map Service (WMS), we will use the layer name.
 - **Description** Description of the online resource.
 - **URL** URL of the online resource. In case of a geographic data services, only the base URL should be provided, without any service parameter.
- **Reference dates** Date(s) associated to the resource. This may include different types of dates. The metadata shall contain a date of publication, revision or creation of the resource.
 - **Date** The date, in ISO format.
 - **Type** The type of date should be provided, and selected from the controlled vocabulary proposed by the ISO 19139. The following date types will often be used:
 - *Date of publication*: This is the date of publication of the resource when available, or the date of entry into force. There may be more than one date of publication. Date of publication differs from the temporal extent. For example, a dataset might have been published in March 2009 (2009-03-15) but the covered information was collected over the year 2008 (temporal extent from 2008-01-01 to 2008-12-31).
 - *Date of last revision*: This date describes when the resource was last revised, if the resource has been revised. Date of revision differs from the temporal extent. For example, a dataset might have been revised in April 2009 (2009-04-15) but the covered information was collected over the year 2008 (temporal extent from 2008-01-01 to 2008-12-31).
 - *Date of creation*: This date describes when the resource was created. Date of creation differs from the temporal extent. For example, a dataset might have been created in February 2009 (2009-02-15) but the covered information was collected over the year 2008 (temporal extent from 2008-01-01 to 2008-12-31).
- **Abstract** This is a brief narrative summary of the content of the resource. The abstract provides a clear and concise statement that enables the reader to understand the content of the data or service. The following is recommended:
 - The resource abstract must be a succinct description that can include:
 - A brief summary with the most important details that summarize the data or service
 - Coverage: linguistic transcriptions of the extent or location in addition to the bounding box

- Main attributes
 - Data sources
 - Legal references
 - Importance of the work
 - Do not use unexplained acronyms.
 - Summarize the most important details in the first sentence or first 100 characters.
- **Additional information** Any other descriptive information about the resource that doesn't fit into other elements.
- **Topics** The Topics element refers to the broad thematic categories under which the dataset can be classified. These topics are often linked to established taxonomies or classification schemes, such as those used in national or international data repositories. The purpose is to help users quickly identify the general subject area of a dataset (e.g., environment, economics, transportation, etc.). Topics are often selected from predefined, standardized lists or ontologies, such as the United Nations' "Statistical Theme" classification or domain-specific controlled vocabularies (e.g., "land use," "biodiversity"). Examples: "Agriculture"; "Urban Planning"; "Climate Change". This classification is useful for large data portals, data aggregators, or metadata systems where users may want to filter or explore datasets based on high-level thematic areas.
- **Keywords** The Keywords element allows for more granular and specific terms that describe the dataset's content. These can be used to identify particular concepts, places, techniques, or any other relevant aspect that the dataset addresses. Keywords help users narrow down search results to find datasets that closely match their specific needs or queries. They are often free-text terms chosen by the dataset creators or curators. Keywords should include relevant technical terms, geographical locations, data formats, measurement units, or any other descriptors that might aid in discovering datasets. Example: "Soil Quality"; "Flood Risk Modeling"; "Carbon Emissions". In contrast to topics, keywords are typically more flexible and specific, often tailored to a particular dataset or user needs. They are usually entered as free text but may also follow controlled vocabulary schemes depending on the application.
- **Type** Keywords type. The ISO 19139 provides a recommended controlled vocabulary.
 - **Keyword** The keyword itself. When possible, existing vocabularies should be preferred to writing free-text keywords. An example of global vocabulary is the Global Change Master Directory that could be a valuable source to reference data domains / disciplines, or the UNESCO Thesaurus.
 - **Thesaurus name** A reference to a thesaurus (if applicable) from which the keywords are extracted. The thesaurus itself should then be documented as a citation.
- **Resource identifiers** The Resource identifiers element is used to specify unique identifiers for a resource, ensuring that it can be consistently referenced and accessed. These identifiers are critical for distinguishing between different datasets or resources and for enabling their retrieval or citation.
- **Identifier** The Identifier is a unique value assigned to an object (in this case, a dataset or resource) within a particular namespace. It is used to uniquely reference the resource, ensuring that it can be identified and distinguished from other resources. The purpose is to provide a permanent and unique way to reference a dataset, resource, or entity, making it easier for users, systems, or software to locate or retrieve the resource. The identifier can take different formats, such as a DOI (Digital Object Identifier), URI (Uniform Resource Identifier), or URN (Uniform Resource Name). It could also be a custom identifier depending on the institution or system. Examples: "doi:10.1234/abcd1234"; "urn:example:dataset:123456" ; "<http://example.org/datasets/12345>". The identifier is crucial for ensuring that datasets or resources are easily findable and retrievable, especially in large repositories or distributed systems.
 - **Authority** The Authority refers to the organization or system that is responsible for assigning or managing the Identifier. This ensures that the identifier has been issued by a recognized, trusted authority and can be linked back to an official registry or system. The purpose is to identify the entity responsible for managing and ensuring the persistence of the identifier. It provides context to users about where the identifier originates and whether it can be trusted for long-term access. Example: "International DOI Foundation (for DOI-based identifiers)"; "European Data Portal (for datasets related to EU resources)". The Authority typically represents a well-known organization or service that assigns and maintains identifiers, ensuring consistency and reliability for long-term data reference.

- **Dataset language** The dataset language, defaulted to the language of the metadata. This refers to the language(s) used within the resource (dataset, series, or service if relevant). It is recommended to use the alpha-3 codes of ISO 639-2. Use only three-letter codes from in ISO 639-2/B (bibliographic codes, example "eng" for English). The list of all the codes is defined at <http://www.loc.gov/standards/iso639-2/>. Regional languages also are included in this list.
- **Presentation form** The Presentation form element in the context of metadata (such as ISO 19139) refers to the way in which the dataset or resource is presented or made available for use. This element specifies the format or mode of delivery of the data, providing essential details about how the data is structured, represented, or visually displayed for users. Purpose: Presentation form helps users understand how the data will be consumed, interacted with, or visualized. It indicates the physical or logical structure of the data and guides users in selecting the appropriate tools or methods to work with the resource. Examples of Presentation form:
 - Digital: The data is in a digital format (e.g., a file that can be read by computer software).
 - Image: The data is represented as a graphic or map image, such as a JPEG, PNG, or TIFF format.
 - Vector data: The data is represented in a vector format (e.g., shapefiles, GeoJSON) suitable for use in Geographic Information Systems (GIS).
 - Text: The data is in textual form, such as a document, report, or database.
 - Spreadsheet: The data is presented in a tabular format, like an Excel spreadsheet or CSV file.
 - Service: The data is made available through a web service, such as a WMS (Web Map Service) or WFS (Web Feature Service).
 - Audio/Video: The data might be multimedia, such as a video or audio file.

The Presentation form element often uses a codelist, which is a set of predefined values that represent different formats and types of presentation. These codelists standardize the types of data presentation, ensuring consistency across datasets. For example, a dataset could have the following presentation forms:

- Map: The data is shown in a map format.
- Report: The data is presented in a report, perhaps in a PDF or DOC format.
- Tabular: The data is provided in a table format.

Why is it important?

- Clarifies data format: By specifying the presentation form, the metadata ensures that users know the format they are working with, allowing them to make decisions about how to access or use the data.
- Improves interoperability: Understanding the presentation form helps ensure that systems, tools, and users can interact with the data properly, reducing errors or confusion.
- Ensures proper visualization: If the dataset is visual (like a map or image), the presentation form indicates the type of visualization, helping users to prepare appropriate viewing tools.
- **Contacts** This is the description of the person or organization responsible for the establishment, management, maintenance or distribution of the resource. This description shall include at least the name of the organization and contact email address. The name of the organization should be given in full, without abbreviations. It is recommended to use institutional email instead of personal emails.
 - **Individual name** The responsible party (person).
 - **Organisation name** The responsible party (organization).
 - **Email** Enter the email address of the contact person. To ensure continuity and long-term accessibility, avoid using personal email addresses. Use a role-based or institutional email account (e.g., help@myorganization.org) that remains valid even if individual staff members change.
 - **Phone** Enter the phone number for contacting the person or team. To ensure continuity and accessibility, avoid using personal or direct mobile numbers. Instead, provide a general or role-based contact number (e.g., a departmental line or help desk number) that will remain valid even if individual staff members change.
 - **Address**
 - **Delivery point** Physical address - Street, building number, etc.
 - **City** Physical address - City name

- **Postal code** Physical address - Postal code
- **Country** Physical address - Country name
- **Online resource**
 - **Name** Name of the online resource.
 - **Description** Description of the online resource.
 - **URL** URL of the online resource.
- **Dataset character set** The Dataset character set refers to the character encoding standard used for representing the textual content of a geographic dataset. Character encoding is essential for ensuring that text data is stored, displayed, and interpreted correctly across different systems, software, and platforms. The Dataset character set element helps document which character encoding scheme is applied to the dataset, particularly if it deviates from the most common standards, such as UTF-8. The purpose is to specify the encoding format for text data within the dataset. This ensures compatibility and correct interpretation of text when the dataset is shared or accessed by different systems.
- **Codelist value** The Codelist value element is used to reference the specific value from a predefined codelist that corresponds to the dataset's character encoding. Codelists are collections of standardized values used to describe or categorize certain elements. In the context of character sets, the codelist provides standard identifiers for commonly used encoding schemes. The purpose is to identify the specific character set in use by referencing a standardized value from a predefined list. Examples: "UTF-8" (code representing the UTF-8 character set); "ISO-8859-1" (code representing ISO Latin-1 character set). The Codelist value ensures that the encoding used is recognized and standardized, facilitating the exchange and interpretation of datasets.
- **Codelist URI** The Codelist URI refers to the Uniform Resource Identifier (URI) that links to the specific codelist that contains the predefined values used in the Codelist value. A URI is a string that provides a unique reference to a resource, in this case, a codelist that defines the character sets used. The URI should point to a standardized, authoritative codelist that can be consulted for valid encoding options. The purpose is to provide a link to the codelist containing the relevant encoding options, enabling the reader or system to check and reference the standard list of accepted character encodings. Examples:
 - <http://www.iso.org/iso-10646-1> (link to the official ISO 10646-1 standard codelist)
 - <http://www.w3.org/TR/encoding> (link to W3C encoding codelist) By providing the Codelist URI, the metadata makes it clear where to find the official list of encoding schemes, ensuring consistency and reference to accepted standards.

SERVICE IDENTIFICATION (this section applies to services only, not to datasets)

- **Service type** This element defines the type or category of service being described. It identifies the general functionality the service provides, usually by referencing a controlled vocabulary such as the OGC service taxonomy (e.g., WMS for Web Map Service, WFS for Web Feature Service, CSW for Catalogue Service for the Web, etc.). The purpose is to allow users and systems to understand what kind of operations the service supports and how to interact with it. It is often expressed as a text string, and may optionally include a URI that identifies the type in a controlled vocabulary.
- **Service type version** This element specifies the version of the service type being described. Different versions may support different operations, parameters, or protocols. Example: For a WMS service, this might be 1.3.0 or 1.1.1. The purpose is to ensure clients can interact with the service using the correct protocol version and avoid compatibility issues.
- **Access properties** This is a general category that includes metadata about how users or systems can access and use the service.
 - **Fees** Indicates any costs associated with accessing or obtaining the dataset. If there is no charge, specify "None" or "Free of charge".
 - **Service availability date** The date when the dataset or related service becomes (or became) available to users. This helps users know when they can access the resource.

- **Ordering instructions** Describes how users can request or obtain the dataset, including any steps, forms, or systems required to place an order.
- **Turnaround** The expected time between placing an order and receiving the dataset or service. This helps users plan their data requests accordingly.
- **Restrictions** Information about any limitations or conditions on accessing or using the dataset or service. This may refer to legal, security, or other types of restrictions.
 - **Legal constraints** This element specifies legal restrictions or obligations governing access to and use of the service. These may include intellectual property rights, licensing terms, or usage policies.
 - **Use limitation** A free text field to describe any legal conditions or obligations that limit how the dataset or service can be used (e.g., "For non-commercial use only", "Must cite the source").
 - **Access constraints** Controlled vocabulary values (e.g., "copyright", "license", "intellectual property rights") that define legal limitations on accessing the service.
 - **Use constraints** Specifies legal restrictions on how the dataset or service can be used once accessed (e.g., no redistribution, attribution required).
 - **Other constraints** Lists any additional legal or contractual conditions not covered by the above fields. This can include terms of service, end-user license agreements, or disclaimers.
 - **Security constraints** This element describes any limitations or requirements related to the confidentiality, integrity, or availability of the service.
 - **Use limitation** Describes any restrictions based on security policies that affect how the dataset or service may be used.
 - **Classification** Indicates the security level assigned to the dataset or service (e.g., Unclassified, Confidential, Secret), following the organization's classification scheme.
 - **Note on security classification** Provides explanatory text or justification for the assigned security classification.
 - **Classification system** Specifies the formal system or policy under which the security classification is defined (e.g., national security guidelines, internal protocols).
 - **Handling description** Details instructions for how the dataset or service must be handled to comply with its security classification (e.g., encryption, access logging).
- **Keywords** A keyword is defined by (i) a keyword value ("keyword"); (ii) an optional originating controlled vocabulary which in ISO standard is referred to as "Thesaurus". If the keyword value originates from a controlled vocabulary (thesaurus, ontology), for example GEMET - Concepts; (iii) the citation of the originating controlled vocabulary shall be provided. It is better to select keyword values from a collection of terms linked and predefined (controlled vocabularies). If only one keyword is used, then for spatial dataset or spatial dataset series, the keyword (i) shall describe the relevant data theme; (ii) shall be expressed in the language of the metadata. For example, a keyword that comes from GEMET - Concepts shall be cited as follows: keyword: freshwater ; thesaurus name: GEMET - Concepts, version 2.4
 - **Type** Indicates the type or role of the keywords, using a controlled vocabulary. Common types include:
 - theme – describes the topic or subject matter (e.g., "elevation", "imagery")
 - place – refers to a geographic location (e.g., "France", "Amazon Basin")
 - stratum – refers to a layer or vertical component (e.g., "atmosphere", "surface")
 - temporal – relates to a time period (e.g., "2020", "historical") This allows metadata users and search engines to interpret keywords correctly and refine search results based on thematic, spatial, or temporal filters.
 - **Keyword** The keyword value is a commonly used word, formalized word or phrase used to describe the subject. While the topic category is too coarse for detailed queries, keywords help narrowing a full text search and they allow for structured keyword search. Examples:
 - Atmospheric conditions (INSPIRE Spatial Data Theme)

- humanCatalogueViewer (spatial data service subcategory)
- water springs (AGROVOC)
- rain water (GEMET -GEneral Multilingual Environmental Thesaurus- Concepts)
- **Thesaurus name** The thesaurus name shall include at least the title and a reference date (date of publication, date of last revision or of creation) of the originating controlled vocabulary. It is important to specify which version of the thesaurus was used to take the keyword value from.
- **Coupled resource** Refers to the dataset(s) that the service operates on or delivers. This links the service to the specific geographic data it provides access to or processes.
 - **Operation name** The name of a specific function or operation that the service provides (e.g., GetMap, GetFeature). This typically corresponds to standard service operations like those in OGC Web Services.
 - **Identifier** A unique code or reference that identifies the service or dataset in a catalog or registry. It ensures consistency and supports interoperability across systems.
- **Coupling type** Describes how tightly the service is connected to the dataset. Common values include:
 - Tight: The service is designed to work only with a specific dataset.
 - Loose: The service can operate on multiple datasets.
 - Mixed: The service has both tightly and loosely coupled operations.
- **Operations contained in service** The elements listed in this group are used to describe the operations supported by a service, how those operations can be invoked, and the technical parameters needed for execution. These elements are used to provide detailed documentation of service capabilities. Operations contained in service contains a list of all operations (i.e., service functions or actions) that the service supports.
 - **Operation name** The name of the operation supported by the service. Example: "GetFeature", "GetMap", "DescribeCoverage".
 - **DCP** (Distributed Computing Platform) Specifies the protocol or platform used to invoke the operation. The purpose is to indicate how the operation can be accessed over a network.
 - **Operation description** A free-text explanation of what the operation does.
 - **Invocation name** The formal name used to call the operation (e.g., in a WSDL or API). Example: "GetMap" (used in service interface)
 - **Parameters** Describes the input and output parameters of the operation. Each operation may have multiple parameters.
 - **Name** The name of the parameter. For example: "bbox", "format", "layers"
 - **Direction** Indicates whether the parameter is an input to, or output from, the operation.
 - **Description** A textual description of the purpose and content of the parameter.
 - **Optionality** Indicates whether the parameter is optional or mandatory.
 - **Repeatability** Indicates whether the parameter can occur multiple times (true or false).
 - **Value type** Describes the data type or value domain of the parameter. For example: string, integer, URI, geometry, enumeration of values.
 - **Connect point** Describes how and where to invoke the operation (i.e., its endpoint).
 - **Linkage** The URL or URI to access the operation.
 - **Name** A label or name for the resource (not the same as the operation name).
 - **Description** A textual description of the online resource.
 - **Protocol** The protocol used to communicate with the service endpoint.
 - **Function** Describes the intended function of the online resource (e.g., download, information, search).

- **Operates on** The element establishes a link between a service metadata record and the dataset(s) it serves. This element identifies the datasets that the service operates on — i.e., the resources that the service accesses, modifies, or delivers. It allows the metadata for a service to explicitly point to the metadata of the dataset(s) it supports, thereby facilitating dataset-service linkage. It can either contain embedded metadata for the dataset or (more commonly) a reference to a separate dataset metadata record.

- **uuidref** A UUID (Universal Unique Identifier) reference to an external dataset metadata record. The purpose is to link the service metadata to a separate dataset metadata record stored elsewhere in a metadata catalog or registry.

PURPOSE, CREDIT AND STATUS

- **Purpose** A description of why the dataset or resource was created—in other words, its intended use or application. The expectations are as follows:

- Content: A clear, concise explanation of the rationale behind the dataset's creation. This could include:
 - The objective of the data collection effort.
 - The decision-making context or policy needs the data supports.
 - Specific applications the dataset was designed for (e.g., urban planning, disaster risk assessment).
 - Any targeted users or communities (e.g., researchers, environmental agencies).
- Format:
 - It is a free-text field.
 - The text should be descriptive, but not excessively long—typically one to a few sentences.
- Best practices:
 - Avoid jargon; ensure it is understandable to a general audience.
 - Distinguish it from other elements like abstract (which describes the content) or useConstraints (which explains how the data may or may not be used).
 - Be specific enough to support discovery and evaluation by potential users.

- **Credit** This element is used to acknowledge individuals, organizations, or agencies that contributed to the creation, funding, or provision of the dataset or resource. The expectations are as follows:

- Content: A free-text acknowledgment of contributors, typically including:
 - Data producers
 - Funders or sponsors
 - Collaborating institutions
 - Partner organizations
 - Any person or group whose contribution should be recognized
- Best practices:
 - Be accurate and complete—credit all major contributors to avoid disputes or omission.
 - Use formal names of organizations or projects to support discoverability and proper attribution.
 - Do not use this field for legal constraints or licensing

- **Status** This element specifies the current state or progress of the dataset or resource. It is a mandatory element in ISO 19115 metadata, as it helps users understand whether the dataset is completed, ongoing, planned, or obsolete. The expectations are as follows:

- Content: A code from a controlled vocabulary defined by ISO 19115. The expected value is taken from the MD_ProgressCode code list. Common values include:
 - completed – data collection and processing are finished

- historicalArchive – no longer maintained but archived for reference
- obsolete – superseded by a newer dataset
- onGoing – being continually updated
- planned – data collection or production is intended but not started
- required – data is needed but not yet available
- underDevelopment – in the process of being created
- Best practices:
 - Use only values from the official code list.
 - Choose the code that best reflects the life cycle stage of the dataset.
 - Be consistent across related metadata records if the status applies to multiple datasets in a series.

EXTENT (GEOGRAPHIC, TEMPORAL, VERTICAL)

In metadata terms, Extent provides information about the geographic area, time period, and/or vertical range to which the data applies.

Types of extents:

- Geographic Extent: Describes the spatial coverage, typically using bounding boxes or geographic identifiers
- Temporal Extent: Specifies the time period the data covers.
- Vertical Extent: Indicates the vertical range (e.g., altitude or depth) covered by the data.

The Extent element is crucial because it:

- Helps users determine if a dataset is relevant for their area or time of interest.
- Supports discovery and filtering in spatial data catalogs.
- Enables automated data integration and analysis processes.
-

Geographic element

- **Bounding box** This is the extent of the resource in the geographic space, given as a bounding box. Defining the coordinates of a rectangle representing the resource area on a map allows the discovery by geographical area. Provide the coordinates bounding the limits of the dataset, by means of four properties:
 - **West bound longitude** Western-most coordinate of the limit of the dataset extent, expressed in longitude in decimal degrees.
 - **East bound longitude** Eastern-most coordinate of the limit of the dataset extent, expressed in longitude in decimal degrees.
 - **South bound latitude** Southern-most coordinate of the limit of the dataset extent, expressed in latitude in decimal degrees.
 - **North bound latitude** Northern-most coordinate of the limit of the dataset extent, expressed in latitude in decimal degrees.
- **Geohash** A short alphanumeric string representing a spatial location or bounding box using the Geohash encoding. The purpose is to provide a compact, human-readable, and indexable way to encode geographic areas.
 - **Geohash** The geohash is not explicitly defined in ISO 19139, but often used in profiles, extensions, or auxiliary metadata to support spatial search, indexing, or approximate location. For example, "u4pruydqqvj" represents a specific area near New York City. Geohashes can vary in length; more characters mean higher precision.
 - **Note** A descriptive or explanatory note about the Geohash value — e.g., resolution, use, limitations. For example, "Geohash at 8-character precision (~19 meters accuracy)."

- **Geographic description** A textual or coded description of the geographic area covered by the dataset or service. The purpose is to provide a named place or coded region (e.g., country, administrative unit) as a way to describe geographic extent, especially when precise coordinates or geometry are not available. This useful for indexing metadata by regions or matching data to known administrative areas.
- **Bounding polygon** A polygonal shape describing the precise spatial coverage of the dataset or service. The purpose is to provide more accurate spatial boundaries than a bounding box, especially for irregular or non-rectangular areas.
 - **Polygon identifier (ID)** A unique identifier assigned to the polygon for referencing or validation. The purpose is to allow referencing this specific geometry within the metadata or from external systems.
 - **Polygon** A geometric object that defines the shape and extent of the area covered. This is defined in GML (Geography Markup Language), typically using EPSG:4326 (WGS 84) as the coordinate system.
 - **Interior or exterior ring** Rings that define the boundary of the polygon. Each ring is defined with a list of coordinates.
 - Exterior ring: The outer boundary of the polygon.
 - Interior ring: One or more holes within the polygon (optional).
 - **Type** The type of geometry used to represent the area. Common types are Polygon (a single closed area), MultiPolygon (a set of multiple polygons), and Envelope (a bounding rectangle).
 - **Coordinates** Coordinates of the polygon. These are the longitude-latitude pairs that define the polygon boundary. Each pair is ordered as **longitude** **latitude**, separated by spaces. The first and last coordinates must be the same to close the polygon.

Temporal element The temporal extent defines the time period covered by the content of the resource. Depending on the temporal characteristics of the dataset, this will consist in a Time period (made of a begin position and end position) or a time instant (made of a single time position) referencing date/time information according to ISO 8601. This time period may be expressed as an individual date, an interval of dates (starting date and ending date), or a mix of individual dates and intervals of dates.

- **beginPosition** Begin time position. Requires an extended ISO 8601 formatted combined UTC date and time string (2009-11-17T10:00:00)
- **endPosition** End time position. Requires an extended ISO 8601 formatted combined UTC date and time string (2009-11-17T10:00:00)
- **Vertical element** The vertical extent of the dataset or service, specifying the range of altitudes or depths it covers. The purpose is to document the vertical dimension (e.g., elevation above sea level, ocean depth) relevant to the data or service. This is essential for datasets that include 3D or altitude-related information, such as terrain models, atmospheric data, oceanographic observations, or data collected at specific depths.
 - **Minimum value** The lowest vertical value covered by the dataset or service. This is the lower bound of the vertical extent, typically in meters. It can be negative for data below a reference level (e.g., sea level).
 - **Maximum value** The highest vertical value covered by the dataset or service. This is the upper bound of the vertical extent.
 - **Vertical CRS** The coordinate reference system used to interpret the vertical values. The purpose is to ensure clarity and interoperability by specifying the reference surface or datum for elevation or depth.

Spatial Representation Type and Resolution

- **Spatial representation type** The spatial representation type of the dataset. Values should be selected from the following controlled vocabulary: {vector, grid, textTable, tin, stereoModel, video}.
- **Spatial resolution** The Spatial resolution element describes the level of spatial detail in a dataset. It can refer to how close together observations are made (e.g., pixel size in raster data or minimum mapping unit in vector data). It shall be expressed as a set of zero to many resolution distances (typically for gridded data and imagery-derived products) or equivalent scales (typically for maps or map-derived products). An equivalent scale is generally

expressed as an integer value expressing the scale denominator. A resolution distance shall be expressed as a numerical value associated with a unit of length.

- **Spatial resolution UOM** The unit of measure for the spatial resolution distance. Common units: "m" for meters, "km" for kilometers, or custom URIs referencing a units catalog.
- **Spatial resolution value** This is the numerical value of the resolution, usually representing the smallest distance that can be reliably distinguished in the dataset. For example: a raster dataset with 30-meter resolution would have a spatial resolution value of 30. NOTES:
 - For services, it is not possible to express the restriction of a service concerning the spatial resolution in the current version of ISO 19119. While the problem is addressed by the standardization community, spatial resolution restrictions for services shall be expressed in the Abstract.
 - When two equivalent scales or two ground sample distances are expressed, the spatial resolution is an interval bounded by these two values.

GRAPHIC OVERVIEW

- **Graphic overview** A graphic or image that provides a quick visual representation of the dataset or service. The purpose is to help users quickly understand the nature, coverage, or content of the resource without needing to inspect the full data. Common examples include (i) a map thumbnail showing the dataset's spatial extent; (ii) a sample chart or diagram illustrating data features; and (iii) a logo or icon representing the service.
 - **File name** The location or URL of the image file. This can be a relative or absolute path.
 - **File description** A textual description of the image's content or purpose. The purpose is to provide context to help users interpret what the image represents. For example: "Overview map of dataset coverage" or "Elevation model sample visualization".
 - **File type** The format or file type of the image. For example, "jpeg" or "png".

FREQUENCY OF UPDATE

This section of the metadata describes how often the dataset or service is maintained, updated, or revised. It helps users understand how current or dynamic the resource is, and whether it is suitable for time-sensitive applications.

- **Resource maintenance** This is a container element that provides information about routine or planned maintenance of the dataset.
 - **Frequency** (Maintenance and update frequency) This indicates how often updates or maintenance actions are performed on the dataset or service. Common values include:
 - continual – Data is repeatedly and frequently updated.
 - daily – Updated every day.
 - weekly – Updated every week.
 - monthly – Updated every month.
 - annually – Updated every year.
 - asNeeded – Updated when necessary (no regular schedule).
 - irregular – Updated at unpredictable intervals.
 - notPlanned – No updates are planned.
 - unknown – Frequency is unknown.
- **Maintenance note** A free-text note providing additional information about the maintenance strategy, versioning policy, or details about how updates are applied. This can be used to explain specific conditions, responsible parties, data revision processes, or known limitations in maintenance.

SPECIFIC USAGE

- **Resource specific usage** This element describes how the resource (dataset or service) has been or can be used, including information on usage contexts, limitations, and responsible parties. The purpose is to help users understand

real-world applications of the resource and any constraints or conditions that affect its use.

- **Specific usage** A free-text description of a particular use or application of the dataset or service. For example, "Used for land cover classification analysis", or "Applied in coastal erosion monitoring projects".
- **Time (date)** The date or time period when the resource was used in the specific context.
- **User determined limitations** A description of limitations identified by the user during the actual use of the resource. Examples: (i) "Data resolution was insufficient for urban-scale planning."; (ii) "Dataset lacks recent updates beyond 2020."
- **Contacts** Information about the individual or organization who used the resource or can provide details about its use.
 - **Individual name** The responsible party (person).
 - **Organisation name** The responsible party (organization).
 - **Email** Enter the email address of the contact person. To ensure continuity and long-term accessibility, avoid using personal email addresses. Use a role-based or institutional email account (e.g., help@myorganization.org) that remains valid even if individual staff members change.
 - **Phone** Enter the phone number for contacting the person or team. To ensure continuity and accessibility, avoid using personal or direct mobile numbers. Instead, provide a general or role-based contact number (e.g., a departmental line or help desk number) that will remain valid even if individual staff members change.
 - **Address** Physical address of the person or organization.
 - **Delivery point** Physical address - Street, building number, etc.
 - **City** Physical address - City name
 - **Postal code** Physical address - Postal code
 - **Country** Physical address - Country name
 - **Online resource** Description and link to an online resource
 - **Name** Name of the online resource.
 - **URL** URL of the online resource.
 - **Description** Description of the online resource

LEGAL CONSTRAINTS

- **Legal constraints** A metadata section that documents legal or regulatory limitations on the access to, or use of, the dataset or service. The purpose is to inform users of any intellectual property rights, licensing terms, privacy protections, access restrictions, and required citations or disclaimers associated with the resource.
 - **Use limitation** A free-text description of limitations on the use of the resource imposed by law or policy. Examples: (i) "Restricted to academic research only."; (ii) "Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)." ; (iii) "Redistribution prohibited without prior written consent."
 - **Access constraints** Describes legal constraints on accessing the dataset or service (e.g. due to privacy, confidentiality, or licensing). The ISO 19139 provides a controlled vocabulary. These are the access constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on obtaining the resource. Common values from the controlled vocabulary are:
 - copyright – Copyright-protected
 - intellectualPropertyRights – IP rights apply
 - license – License restrictions apply
 - privacy – Limited to protect personal or sensitive information
 - confidential – Only authorized users may access
 - restricted – Access limited for other reasons
 - otherRestrictions – Other restrictions apply
 - unrestricted – No legal access constraints

- **Use constraints** Describes how the data may or may not be used, usually entered as free text. While Access constraints controls who can access, Use constraints controls what users can do with the data once accessed. Typical content:
 - Terms of use
 - Disclaimers
 - Preferred citations
 - Attribution requirements
 - Data limitations (if legal in nature) Examples: (i) "Users must cite the source when publishing results."; (ii) "Not for navigation or emergency use."; (iii) "Use of data is subject to the terms described at <https://example.org/license>".
- **Other constraints** Any other legal restrictions and legal prerequisites for accessing and using the resource or metadata. Use cases include:
 - Mandatory citation format
 - Mandated attribution wording
 - Data sharing policies
 - Jurisdiction-specific legal notes

RESOURCE FORMAT¹

- **Resource format** The format of the resource, describing the file format, encoding, or structure used to store or represent the dataset or service. This metadata element helps users understand the type of file(s) they are dealing with, whether it's a dataset, document, or service, and the technology or software needed to process or interpret the resource. The purpose is to specify the file format(s) (such as CSV, JSON, NetCDF, or GeoTIFF) or the service format (e.g., WMS, WFS) that the resource is delivered in.
 - **Name** The name of the format used to store or represent the resource. This could refer to the type of file (e.g., CSV, GeoTIFF, XML, JSON), the technology (e.g., WMS for Web Map Service), or the application-specific format (e.g., Shapefile for vector data).
 - **Version** The version of the format used. This helps identify specific versions of a format that may have different features, capabilities, or compatibility issues. Version information is important for ensuring the correct version of a format is used, as formats can evolve over time (e.g., a new version of GeoTIFF might support different compression methods).

REFERENCE SYSTEM

- **Reference system** The reference system(s) typically (but not necessarily) applies to the geographic reference system of the dataset. Multiple reference systems can be listed if a dataset is distributed with different spatial reference systems. This block of elements may also apply to service metadata. A spatial web-service may support several map projections / geographic coordinate reference systems. A reference system is defined by two properties:
 - **Code** The identifier of the reference system. The recommended practice is to use the Spatial Reference Identifier (SRID) number. For example, the SRID of the World Geodetic System (WGS 84) is 4326.
 - **Code space** The code space of the source authority providing the SRID. The best practice is to use the EPSG authority code EPSG (as most of geographic reference systems are registered in it). Codes from other authorities can be used to define ad-hoc projections, for example:
 - ESRI:54012 (Eckert IV equal area projection)
 - EPSG:4326 (World Geodetic System 84 - aka WGS84), the system used for GPS
 - EPSG:3857 (Web Mercator / Pseudo-Mercator) - widely used for map visualization from web map tile providers.

The main reference system registry is EPSG, which provides a "search by name" tool for users who need to find a SRID (global or local/country-specific). Other websites reference geographic systems, but are not authoritative

sources including <http://epsg.io/> and <https://spatialreference.org/>. The advantage of these sites is that they go beyond the EPSG registry, and handle other specific registries given by providers like ESRI.

The following ESRI projections could be relevant, in particular those in support of world equal-area projected maps (maps conserving area proportions):

- ESRI:54012 (Eckert IV)
- ESRI:54009 (Mollweide)
- ESRI:54030 (Robinson)

Spatial Representation

Spatial representation describes how spatial data is represented — that is, the method used to encode the spatial characteristics of the dataset. This element provides structured metadata about the type and structure of spatial data in the resource. It helps users understand how location data is stored and whether the format fits their needs.

- **Vector data** For data consisting of points, lines, and polygons (e.g., shapefiles, geoJSON, feature classes in GIS), spatial representation includes Geometric objects (type and count of geometric features) and topology level (which describes topological complexity, e.g., planar graph).
 - **Topology level** Topology level is the type of topology used in the vector spatial dataset. The ISO 19139 provides a controlled vocabulary. In most cases, vector datasets will be described as geometryOnly which covers common geometry types (points, lines, polygons).
 - **Geometric objects**
 - **Type** The type of geometry handled. A controlled vocabulary is used. In the case of an homogeneous geometry type, a single geometricObject element can be defined. For complex geometries (mixture of various geometry types), one geometricObjects element will be defined for each geometry type.
 - **Count** The number (count) of geometries in the dataset.
- **Grid data** For grid/raster data (e.g., satellite imagery, gridded climate data), spatial representation includes: Cell size and orientation; Transformation parameters (if georeferenced); and Whether the grid is georectified or georeferenceable.
 - **Number of dimensions** Number of dimensions in the grid.
 - **Axis dimension properties** A list of each dimension including, for each dimension, the name, size, and resolution.
 - **Name** The name of the dimension type: the ISO 19139 provides a controlled vocabulary with the following options: row, column, vertical, track, crossTrack, line, sample, and time. These options represent the following:
 - row: ordinate (y) axis
 - column: abscissa (x) axis
 - vertical: vertical (z) axis
 - track: along the direction of motion of the scan point
 - crossTrack: perpendicular to the direction of motion of the scan point
 - line: scan line of a sensor
 - sample: element along a scan line
 - time: duration
 - **Size** The length of the dimension.
 - **Resolution** The dimension resolution: a resolution number associated to a unit of measurement. This is the resolution of the grid cell dimension. For example:
 - for longitude/latitude dimensions, and a grid at 1deg x 5deg, the 'row' dimension will have a resolution of 1 deg and the 'column' dimension will have a resolution of 5 deg.

- for a "vertical" dimension, this will represent the elevation step. For example, the vertical resolution of the mean Ozone concentration between 40m and 50m altitude at a location of longitude x/ latitude y would be 10 m.
- similar: in case of a spatial-temporal grid, the "time" resolution will represent the time lag (e.g., 1 year, 1 month, 1 week, etc.) between two measures.
- **Cell geometry** The type of geometry used for grid cells. Possible values are: point, area, voxel, and stratum. Most "grids" are commonly area-based, but in principle a grid goes beyond this and the grid cells can target a point, an area, or a volume.
 - point: each cell represents a point
 - area: each cell represents an area
 - voxel: each cell represents a volumetric measurement on a regular grid in a three dimensional space
 - stratum: height range for a single point vertical profile
- **Transformation parameter availability**

DATA QUALITY

This section of the metadata document describes the quality of the resource by including specific measures, methods, and procedures used to assess its fitness for use. It helps users assess the trustworthiness of the data for their specific needs.

- **Scope** Describes the scope or extent of the quality evaluation. It indicates which part of the resource (e.g., dataset, attribute, metadata) the quality assessment applies to. Examples: "Dataset level"; "Attribute level"; "Metadata level"; "Feature level". The ISO 19139 recommends the use of a controlled vocabulary.
- **Report** The Report section documents the specific quality measures and the methods used to evaluate the data quality. It provides detailed information on the data quality measures taken and the process by which they were assessed. Each report will include the following key elements: information on the measure (what was measured), a description of the evaluation method (how the quality was evaluated, e.g., direct internal check, external test), and the result (the outcome of the evaluation, e.g., quantitative result, pass/fail). There can be multiple report entries (**element types**), each representing a different dimension of quality.
 - **Element type** Each report details one quality evaluation, the **element type**, such as Completeness, Logical consistency, Positional accuracy, Temporal accuracy, or Thematic accuracy. The Element type must be selected from the following controlled vocabulary:
 - "DQ_CompletenessOmission"
 - "DQ_CompletenessCommission"
 - "DQ_ConceptualConsistency"
 - "DQ_DomainConsistency"
 - "DQ_FormatConsistency"
 - "DQ_TopologicalConsistency"
 - "DQ_PositionalAccuracy"
 - "DQ_ThematicAccuracy"
 - "DQ_TemporalAccuracy"
 - "DQ_QuantitativeAttributeAccuracy"
 - "DQ_UsabilityElement"
 - **Name of measure** The name of the quality measure used to evaluate the dataset's quality. This could refer to specific accuracy measures, completeness, consistency, or other quality indicators. Examples: "Positional accuracy"; "Logical consistency"; "Completeness"; "Temporal accuracy", "Thematic accuracy".

- **Measure identification** An identifier for the specific quality measure, often referencing standards, documents, or predefined measurement systems. Examples: "ISO 19157:2013"; "FGDC-STD-007.3-1998".
- **Measure description** A detailed description of the quality measure used, explaining its purpose, how it was applied, and any specific methodologies or protocols followed. Example: "This measure evaluates the accuracy of the spatial location of each feature compared to the actual location using high-precision GPS data."
- **Evaluation method type** Specifies the type of method used to evaluate the quality of the data, such as statistical analysis, expert review, or automated testing. Examples: "Statistical analysis"; "Expert review"; "Automated testing"; "Sampling" The ISO 19139 recommends the use of a controlled vocabulary with the following options: directInternal, directExternal, indirect.
- **Evaluation method description** A description of the evaluation method, detailing how the quality was assessed and the procedures followed. Example: "Accuracy assessment performed by comparing GPS-derived points with the dataset's coordinates using a root mean square error calculation."
- **Evaluation procedure** The Evaluation procedure documents the specific process used for evaluating the quality measure.
 - **Title** The title of the evaluation procedure, describing the assessment process. Example: "Spatial accuracy evaluation procedure".
 - **Dates** The date related to the evaluation procedure, indicating when the evaluation was conducted or published.
 - **Date** The actual date when the evaluation was performed.
 - **Type** Specifies the type of date, such as creation, publication, or last modified.
 - **Identifier authority** The authority that issued or controls the identifier for the evaluation procedure.
 - **Identifier code** The code or unique identifier used to reference the evaluation procedure. Examples: "ISO-19157:2013"; "FGDC-STD-007.3-1998".
- **Conformance result** The Result of consistency check element documents the findings of any consistency checks performed on the data or resource. These checks are typically conducted to ensure that the data conforms to established standards, rules, or expected values, helping verify its validity and reliability. This element is used to record the outcomes of these checks, which can include the results of tests for consistency, accuracy, logical integrity, or any other form of validation.
 - **Title** The title of the consistency check performed. This could describe the specific test or procedure used to verify the data's quality or conformity. Example: "Positional Accuracy Consistency Check"; "Attribute Consistency Test".
 - **Dates**
 - **Date** The date when the consistency check was performed or completed. This indicates when the check was conducted, which helps provide context for the data's current validity.
 - **Type** The type of date related to the consistency check. This refers to the nature of the date, such as whether it is the date of creation, date of publication, or date of revision. Example: (i) "Creation" (Indicates that the date marks the creation of the check) ; "Revision" (Indicates that the check was revised at this date)
 - **Responsible party** The party responsible for carrying out the consistency check or for ensuring the data's consistency. This typically refers to an individual, organization, or group that conducted the check or is accountable for the quality and validity of the data.
 - **Individual name**
 - **Organisation name**
 - **Email**
 - **Phone**
 - **Address**

- **Delivery point**
- **City**
- **Postal code**
- **Country**
- **Online resource**
 - **Name**
 - **URL**
 - **Description**

- **Resource presentation** The Resource presentation element specifies how a resource is provided, accessed, or presented to users. It indicates the format or method of delivery, such as a document, dataset, map, or service, and how the resource is structured or made available for users to interact with or utilize. This is important for understanding how a resource can be accessed and what type of presentation or output format users can expect when interacting with the data. Example: "PDF document", "CSV file", "Web service", "GeoTIFF", "Shapefile". The element describes the presentation format of the resource, indicating whether it is downloadable, interactive, or available through a service.
- **Explanation** The Explanation element provides additional context or a more detailed description about a specific process, measurement, or result related to the resource. It can include clarification about how a result was obtained, how a test was performed, or why a certain decision was made. This is important for transparency, providing users with the rationale or detailed information behind the data, quality assessments, or any checks that have been carried out. Example: "The resource was assessed using a positional accuracy check, and the dataset passed the consistency check due to correct geospatial coordinates within the acceptable range." This field is used to give further details and clarification on why a result or value was recorded in the metadata.
- **Pass** The Pass element indicates the outcome of a conformance check or validation test conducted on the resource. This is a boolean indicator (True or False) that specifies whether the resource has passed or failed a specific conformance test. If not evaluated, the value may be left blank (null). This helps the user understand the conformance status of the resource in terms of whether it meets certain predefined quality or technical standards. Values: True (the resource has successfully met the required standard or passed the test, indicating conformance) ; False (the resource did not meet the required standard, indicating non-conformance) ; Null/Blank (the resource has not been evaluated, and no conformance status is available). Example: If a dataset has been checked for consistency, a Pass value of "True" would mean that the dataset passed the check, whereas "False" would indicate that it did not meet the required standard. This element is used to track and report the results of quality control or validation tests performed on the resource.
- **Quantitative result** Quantitative results that describe the outcome of a data quality evaluation using numeric measures. These results provide detailed metadata about the type of result, the units used, and the actual values.
 - **Value type** The type of data used to express the result value(s). This indicates the data type (e.g., Real, Integer, Text) of the values listed under value.
 - **Value unit identifier** A unique ID for the unit of measure (usually a URI). This enables referencing a well-defined, controlled unit.
 - **Value unit name** Human-readable name of the unit (e.g., "metre", "seconds"). This makes the unit understandable without resolving the identifier.
 - **Value unit quantity type** Type of quantity measured (e.g., Length, Time, Angle). This indicates the conceptual type of measurement for which the unit is appropriate.
 - **Value** The actual numeric result(s) of the evaluation. This represents the measured outcome of the data quality check.

- **Lineage**

- **Lineage statement** Lineage is "a statement on process history and/or overall quality of the spatial data set. Where appropriate it may include a statement whether the data set has been validated or quality assured, whether it is the

official version (if multiple versions exist), and whether it has legal validity. The value domain of this element is free text." This element is not applicable to geographic services.

RECOMMENDATIONS:

- If a data provider has a procedure for the quality management of their spatial data set (series) then the appropriate ISO data quality elements and measures should be used to evaluate and report (in the metadata) the results. If not, the Lineage metadata element (defined in the Implementing Rules for Metadata) should be used to describe the overall quality of a spatial data set (series).
- The use of acronyms should be avoided. If used, their meaning should be explained.
- **Process steps** The process history may be described by information on the source data used and the main transformation steps that took place in creating the current data set (series). It consists of a list of events or transformations applied during dataset production.
 - **Description** A textual description of the process that was executed (e.g., "resampled from 1m to 10m grid", or "aggregated by region"). This describes what was done during the process step.
 - **Rationale** An explanation of why the process step was undertaken (e.g., "to harmonize resolution with other datasets"). This documents the intent or justification behind the processing.
 - **Date** The date and time at which the process step was completed. This supports tracking the sequence and timing of processing events.
 - **Processor** The party (organization or individual) responsible for executing the process step. This identifies who performed the processing, ensuring accountability and traceability.
 - **Individual name**
 - **Organisation name**
 - **Role**
 - **Email**
 - **Phone**
 - **Address**
 - **Delivery point**
 - **City**
 - **Postal code**
 - **Country**
 - **Online resource**
 - **Name**
 - **URL**
 - **Description**
 - **Source** Specifies the input data used in the process step.
 - **Title** The name of the source dataset or resource. This provides a short reference to the input data.
 - **Description** Details about the content and role of the source dataset. This helps users understand the nature and function of the input data.
 - **Date** Represents one or more dates associated with the source data (e.g., publication date, creation date).
 - **Date** The date itself (e.g., 2024-12-27).
 - **Type** The type of date (e.g., creation, publication, revision).
 - **Identifier authority** The organization responsible for assigning the identifier (e.g., a DOI registration agency). This establishes provenance and authority of the identifier.
 - **Identifier code** A unique string or code used to identify the source dataset (e.g., DOI, catalogue ID).

- **Responsible party** An individual or organization responsible for the source dataset (e.g., data producer or publisher).
 - **Individual name**
 - **Organisation name**
 - **Position**
 - **Role**
 - **Email**
 - **Phone**
 - **Address**
 - **Delivery point**
 - **City**
 - **Postal code**
 - **Country**
 - **Online resource**
 - **Name**
 - **URL**
 - **Description**
 - **Protocol**
 - **Function**
- **Resource presentation** The form in which the source resource is presented (e.g., mapDigital, documentDigital).

CONTENT

- **Coverage description** This section is used to describe the content and structure of a gridded dataset (e.g., imagery, elevation models, multi-band raster data). It describes the attributes or characteristics of the coverage data, which are typically structured in a grid or other regular pattern. It includes general information about the contents and structure of the dataset.
 - **Content type** Indicates the type of information represented in the coverage (e.g., physical measurement, thematic classification, etc.). Examples: image, thematicClassification, physicalMeasurement
 - **Dimension** Represents the axes or bands of the data — typically refers to spectral, temporal, or other data dimensions.
 - **Band** A specific type of range dimension used to describe spectral bands in a raster (e.g., a satellite image or multispectral scan).
 - **Band name** Name or identifier for the spectral band (e.g., "Red", "Band_1").
 - **Band type** A label describing the band function (e.g., "reflectance", "temperature").
 - **Band description** Narrative text describing what the band measures.
 - **Minimum value** The smallest possible or observed value in the band.
 - **Maximum value** The largest possible or observed value in the band.
 - **Band unit identifier** Unique identifier for the unit of measurement (often a URI).
 - **Band unit name** The name of the measurement unit (e.g., "W/m²/sr/μm").
 - **Band unit quantity type** The physical quantity measured (e.g., "radiance", "temperature").
 - **Band peak response** Wavelength or frequency where the band is most sensitive.
 - **Band bits per value** Number of bits used to encode each pixel value in the band.

- **Band tone gradation** Number of gradations (distinct values) that can be represented.
- **Band scale factor** Scale multiplier applied to stored pixel values to get physical values.
- **Band offset** Offset added to scaled pixel values to get physical values.
- **Range dimension** A general description of a data dimension (not necessarily a spectral band).
- **Range dimension name** Label for the dimension (e.g., "Time", "Elevation").
- **Range dimension type** Type or purpose of the dimension.
- **Range dimension description** Explanation of the role or content of the range dimension.

DISTRIBUTION

- **Distribution format** The Distribution format element specifies the format in which the resource (such as data or metadata) is distributed or made available to users. This element helps to describe the technical format of the resource as it is provided for download, transfer, or access through an external service. It identifies the file format or structure in which the resource is presented to the users. The purpose is to describe the format in which a resource is distributed, making it easier for users to know in what format the data or resource is provided. Examples: "CSV" (Comma Separated Values); "GeoTIFF" (for geospatial data); "JSON" (JavaScript Object Notation); "PDF" (Portable Document Format); "Shapefile" (ESRI Shapefile format). This element is critical for users to understand the type of file or format they will receive when they access or download the resource. It provides clarity on how the data or information is structured for compatibility with software tools or systems.
 - **Name** The Name element under Distribution format specifies the name of the format in which the resource is distributed. It gives the specific format type used for distribution (such as "CSV", "GeoTIFF", "XML", etc.). The purpose is to provide the specific name of the format, ensuring that users can understand what type of data they are accessing and whether they can use it with their software tools.
 - **Version** The Version element under Distribution format indicates the version of the format used for distribution. This is important when different versions of a format exist, as software compatibility or features may vary across different versions of the same format. By specifying the version, users are informed about the specific iteration of the format that applies to the resource. The purpose is to specify the exact version of the format in use, ensuring clarity on what version is being distributed and helping users ensure compatibility with their tools or systems. For example: For a "Shapefile", the version might be something like "1.0" or "2.0", depending on the specific release of the Shapefile format.
- **Distributor** The Distributor element in the context of ISO 19139 metadata provides information about the entity or organization responsible for distributing a resource (such as a dataset or service). This element is essential for identifying who is making the resource available to the public or specific users and often includes contact information for accessing the resource. The Distributor element identifies the party (person, organization, or agency) that distributes the resource. This entity is responsible for making the data, service, or information available to users, either by providing access, facilitating downloads, or managing any other type of distribution channel. The purpose is to specify the organization or entity that is responsible for the availability and distribution of the resource. This can include details like the name of the distributor, contact details, and information about how the resource can be obtained.
 - **Individual name**
 - **Organisation name**
 - **Email**
 - **Phone**
 - **Address**
 - **Delivery point**
 - **City**
 - **Postal code**
 - **Country**
 - **Online resource**

- **Name**
- **URL**
- **Description**

FEATURE CATALOGUE

A Feature Catalogue defines the types of features and their attributes used in a dataset. It describes the content and structure of geographic features, enabling consistency, interoperability, and understanding when sharing or using data.

- **Name** The official title or name of the feature catalogue. The purpose is to identify the catalogue referenced or used by the dataset. Example: "Topographic Features Catalogue v2"
- **Scope** Describes the subject or domain covered by the feature catalogue. The purpose is to clarify the thematic or spatial extent, or the application domain (e.g., marine data, cadastral data). Example: "Hydrographic data for coastal management".
- **Fields of application** Describes the subject or domain covered by the feature catalogue. The purpose is to clarify the thematic or spatial extent, or the application domain (e.g., marine data, cadastral data). Example: "Hydrographic data for coastal management".
- **Version number** The identifier for the specific version of the feature catalogue. The purpose is to ensure clarity when multiple versions exist; essential for version control and traceability. Example: "2.1".
- **Version date** The date the version of the feature catalogue was published or became effective. The purpose is to indicate the timeliness of the catalogue. Example: "2023-11-01" (date should be entered in ISO 8601 format).
- **Version date type** Specifies the type of date provided for the version (e.g., creation, publication, revision). This is usually drawn from ISO 19115 date types like: creation; publication; revision.
- **Producer** The organization or authority responsible for creating and maintaining the feature catalogue. The purpose is to provide attribution and a point of contact for updates or clarification.
 - **Individual name**
 - **Organisation name**
 - **Email**
 - **Phone**
 - **Address**
 - **Delivery point**
 - **City**
 - **Postal code**
 - **Country**
 - **Online resource**
 - **Name**
 - **URL**
 - **Description**
- **Functional language** The Functional language element specifies the primary natural language(s) used within the resource being described, such as a dataset, service, document, or feature catalogue. This is not the metadata language (which is recorded separately); it refers specifically to the language used within the content of the resource itself (e.g., labels, feature names, field values, or descriptive text in a dataset or service). The purpose is to help users understand the language context of the resource and to support multilingual data discovery, access, and interoperability. For example, a user searching for resources in Spanish can filter results using this metadata element.

In ISO 19139, this will be represented using a code from the ISO 639-2 three-letter language codes. It may include multiple entries if the resource uses more than one language.

- **Feature type** A Feature type defines a category of real-world geographic entities that share common characteristics (attributes, relationships, and operations). Feature types are the core elements of a feature catalogue, which describes the structure and semantics of data in a geographic dataset.
 - **Type name** The unique name assigned to the feature type. The purpose is to identify the feature class (e.g., Building, Road, Parcel).
 - **Definition** A textual explanation of the meaning of the feature type. The purpose is to describe what the feature represents in the real world. Example: A natural or artificial channel through which water flows.
 - **Code** A unique identifier for the feature type (often used in systems or standards). The purpose is to enable machine-readable identification and mapping across systems.
 - **Is abstract** Indicates whether this feature type is abstract. If true, this feature type cannot have instances; it must be specialized (subtyped). Value: Boolean (true/false)
 - **Aliases** Alternative names used for the feature type. The purpose is to support interoperability and usability in multilingual or multi-standard environments. Example: Stream, Creek (as aliases for Watercourse)
 - **Carrier of characteristics** A carrier of characteristics defines the attributes or properties associated with the feature type (these are sometimes called feature attributes or feature properties).
 - **Member name** The name of the attribute or characteristic. Example: FlowRate, Name, Length
 - **Definition** Description of the meaning or role of the attribute. Example: The volume of water passing a point per unit of time.
 - **Cardinality lower** The minimum number of times this attribute may occur for a single feature instance. Example: 0 (optional), 1 (mandatory).
 - **Cardinality upper** The maximum number of times this attribute may occur. Example: 1, * (unbounded)
 - **Code** A unique identifier for the attribute within the feature catalogue. This corresponds to the actual column name in an attributes table.
 - **Measurement unit** The unit of measure for the attribute, if it is quantitative. Example: m³/s for FlowRate
 - **Value type** The data type of the attribute value. Examples: CharacterString, Integer, Real, Date, Boolean
 - **Listed values** List of controlled value(s) used in the attribute member. Each value corresponds to an object compound by 1) a label, 2) a code (as contained in the dataset), 3) a definition. This element will be used when the feature member relates to reference datasets, such as code lists or registers. e.g., list of countries, land cover types, etc.
 - **Label** The human-readable name of the value. Example: Perennial
 - **Code** A machine-readable identifier of the value. Example: PER
 - **Definition** The description of what the code/label means. Example: A watercourse that flows continuously throughout the year.

TAGS

Tags, especially when organized in tag groups, provide a powerful and flexible solution to enable custom facets (filters) in data catalogs.

- **Tag** A user-defined tag.
- **Tag group** A user-defined group (optional) to which the tag belongs. Grouping tags allows implementation of controlled facets in data catalogs.

Documenting an image

"Although photographs may be more explicit than a long discourse for humans, they don't describe themselves in term of content as texts do. For texts, authors use many clues to indicate what they are talking about: titles, abstract, keywords, etc. which may be used for automatic cataloguing. Searching for photos must rely on manual cataloguing, or relate texts and documents that come with the photos." (Source: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.43.5077&rep=rep1&type=pdf>)

Metadata standards

This chapter explains how to document digital images using the Metadata Editor. Digital images are electronic files typically stored in formats such as JPG, PNG, or TIFF. These may include photographs taken by digital cameras, computer-generated graphics, or scanned images. The Metadata Editor applies metadata standards to ensure that images are discoverable, accessible, and usable. To achieve this, metadata should cover:

- **Content information:** Caption, description, keywords, and other descriptive elements.
 - **Provenance information:** Date, location, author, and source.
 - **Rights and privacy considerations:** Copyright/license details and privacy-related information (e.g., presence of identifiable individuals, especially minors), enabling users to use images legally, ethically, and responsibly.
-

Metadata embedded in digital image files

Some devices, such as digital cameras, automatically generate and embed metadata in image files. This metadata, known as EXIF (Exchangeable Image File Format), records:

- **Date and time** the image was taken.
- **GPS coordinates** (latitude, longitude, and possibly altitude) if geolocation was enabled.
- **Device information:** Manufacturer, model, lens type, focal range, aperture, shutter speed, and flash settings.
- **Unique image identifier** generated by the system.

EXIF metadata can be extracted or viewed using various tools. For instance, R packages such as *ExifTool* and *ExifR* allow extraction and processing of EXIF metadata, while applications like Flickr display EXIF content. However, beyond date, location, and unique identifiers, EXIF metadata provides little information relevant to identifying an image's source or content.

Enhancing metadata with additional descriptive information

Since EXIF metadata lacks detailed content descriptions, curators must supplement it with additional information. Some of this metadata can be manually entered, while other elements can be automatically extracted using machine learning

models or APIs. This information is structured and stored according to metadata standards supported by the Metadata Editor.

The Metadata Editor offers two mutually exclusive options for documenting images:

- **Dublin Core (DCMI)**: A simple and flexible metadata schema.
- **IPTC**: A more detailed and comprehensive schema, by the International Press Telecommunications Council (IPTC).

Both options are supplemented with a few common metadata elements, including cataloging parameters and unique identifiers.

IPTC and Dublin Core standards

Dublin Core standard

The Dublin Core (DCMI) standard consists of 15 core metadata elements, which are supplemented in the Metadata Editor with additional elements, primarily from the ImageObject schema of schema.org. This option provides a simpler, flexible approach to documenting images.

IPTC Standard

The IPTC Photo Metadata Standard (version 2019.1) is the most widely adopted standard for describing images, recognized by news agencies, photographers, libraries, museums, and other industries. It consists of two schemas:

- **IPTC Core**: Covers essential descriptive fields.
- **IPTC Extension**: Provides additional elements for greater detail.

IPTC metadata includes fields for time and geographic coverage, people and objects shown, usage rights, and more. Since the IPTC schema is highly detailed, curators will typically use only a subset of the available fields. Where applicable, the use of controlled vocabularies is recommended for consistency.

Metadata structure and templates

The Metadata Editor uses a schema that contains:

- **Common metadata elements**: Used for cataloging, assigning unique identifiers, and documenting metadata provenance.
- **Two metadata blocks**: One for IPTC and one for Dublin Core. Users can choose one of these options to describe the image.
- **Additional elements**: Common to both options, providing supplementary documentation.

To facilitate image documentation, the Metadata Editor includes two pre-configured metadata templates: one based on Dublin Core, the other based on IPTC. These templates can be customized using the Template Manager to suit the organization's needs.

Documenting an image using the Dublin Core option

This section describes the documentation of an image using the Dublin Core (DCMI) option.

The Dublin Core standard contains 15 core elements, which are generic and versatile enough to be used for documenting different types of resources. Other elements can be added to the specification to increase its relevancy for specific uses. We added a few elements inspired by the [ImageObject](#) schema from schema.org to the 15 elements.

The fifteen elements, with their definition extracted from the [Dublin Core website](#), are the following:

| Element name | Description |
|--------------------------|---|
| <code>identifier</code> | An unambiguous reference to the resource within a given context. |
| <code>type</code> | The nature or genre of the resource. |
| <code>title</code> | A name given to the resource. |
| <code>description</code> | An account of the resource. |
| <code>subject</code> | The topic of the resource. |
| <code>creator</code> | An entity primarily responsible for making the resource. |
| <code>contributor</code> | An entity responsible for making contributions to the resource. |
| <code>publisher</code> | An entity responsible for making the resource available. |
| <code>date</code> | A point or period of time associated with an event in the life cycle of the resource. |
| <code>coverage</code> | The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant. |
| <code>format</code> | The file format, physical medium, or dimensions of the resource. |
| <code>language</code> | A language of the resource. |
| <code>relation</code> | A related resource. |
| <code>rights</code> | Information about rights held in and over the resource. |
| <code>source</code> | A related resource from which the described resource is derived. |

The Metadata Editor does not use the `identifier` element, as the unique identifier is provided by the common element `idno`.

The World Bank schema includes the following additional elements that are not part of the core list of the DCMI:

- `identifiers`

- [caption](#)
- [keywords](#)
- [topics](#)
- [country](#)
- [gps](#) (latitude, longitude, altitude)
- [note](#)

These elements are included in the default template for the DCMI option.

Create a new project

Create a new project by clicking on [CREATE NEW PROJECT](#) in the *My projects* page, and select option *Image* as data type.

Create new project

 Microdata

 Timeseries

 Timeseries database

 Document

 Table

 Image

 Script

 Video

 Geospatial

CLOSE

In the new *project home page*, select a DCMI template (one DCMI template is provided with the application).

Metadata Editor

Required Recommended Empty

Search...

- Home
- Metadata information
- Image identifiers
- DCMI
- License
- Albums
- Tags
- External resources
- Administrative metadata

Market near Ramallah's main mosque



Project owner:
John Doe

Created on:
2025-02-14 11:49:16

Last changed by:
John Doe

Changed on:
2025-02-14 14:06:16

Project IDNO:
2c63f358-E5ef-4bb3-A654-A8968b6ba694

Template

Project template:
IHSN IMAGE DCMI TEMPLATE 1.0 EN -

Administrative metadata templates:

Project validation

Schema validation ⓘ
No validation errors found

Template validation

Collaborators

None

Collections

None

Data and Documentation

Disk usage: 1.09 MB

| DOCUMENTATION | FILES |
|---------------|-------|
| Title | Type |

Edit the thumbnail by selecting an image. It is highly recommended to select the image being documented as thumbnail. The thumbnail will be used in the Metadata Editor itself, and in a NADA catalog is the image is published in NADA. The system will save the thumbnail as a low-resoluton image.

Fill out the *Information on metadata* section

Metadata Editor

Required Recommended Empty

Search...

- Home
- Metadata information
- Information on metadata
 - Document title
 - Document ID
 - Metadata producers
 - Production date
 - Version
- Image identifiers
- DCMI
- License
- Albums
- Tags
- External resources

Market near Ramallah's main mosque

Information on metadata

Document title ⓘ

Document ID ⓘ

Metadata producers ⓘ

| Name | Abbreviation | Affiliation | Role |
|----------|--------------|-------------|------|
| John Doe | | | |

+ ADD ROW

Production date ⓘ
2025-02-14

Version ⓘ

In the list of metadata elements below, the key of each element in the metadata standard is provided between brackets next to the corresponding element's label in the template.

Fill out the DCMI and additional sections (image description)

We provide here some description and recommendations for the key metadata elements in the DCMI template.

IMAGE DESCRIPTION

- **Resource type** (*type*) The Dublin Core schema is flexible and versatile, and can be used to document different types of resources. This element is used to document the type of resource being documented. The DCMI provides a list of suggested categories, including "image" which is the relevant type to be entered here. Some users may want to be more specific in the description of the type of resource, for example distinguishing color from black & white images. This distinction should not be made in this element; another element can be used for such purpose (like tags and tag groups).
- **date** (*date*) The date when the photo was taken / the image was created, preferably entered in ISO 8601 format.
- **Title** (*title*) The title of the photo.
- **Caption** (*caption*) A caption for the photo.
- **Description** (*description*) A brief description of the content depicted in the image. This element will typically provide more detailed information than the title or caption. Note that other elements can be used to provide a more specific and "itemized" description of an image; the element **keywords** for example can be used to list labels associated with an image (possibly generated in an automated manner using machine learning tools).
- **Keywords** (*keywords*) Words or phrases that describe salient aspects of an image content. Can be used for building keyword indexes and for classification and retrieval purposes. A controlled vocabulary can be employed. Keywords should be selected from a standard thesaurus, preferably an international, multilingual thesaurus.
 - **Keyword** (*name*) Keyword (or phrase). Keywords summarize the content or subject matter of the image.
 - **Vocabulary** (*vocabulary*) Controlled vocabulary from which the keyword is extracted, if any.
 - **URL** (*uri*) The URI of the controlled vocabulary used, if any.
- **Topics** (*topics*) The **topics** field indicates the broad substantive topic(s) that the image represents. A topic classification facilitates referencing and searches in electronic survey catalogs. Topics should be selected from a standard controlled vocabulary such as the [Council of European Social Science Data Archives \(CESSDA\) thesaurus](#).
 - **ID** (*id*) The unique identifier of the topic. It can be a sequential number, or the ID of the topic in a controlled vocabulary.
 - **Topic** (*name*) The label of the topic associated with the data.
 - **Parent ID** (*parent_id*) When a hierarchical (nested) controlled vocabulary is used, the **parent_id** field can be used to indicate a higher-level topic to which this topic belongs.
 - **Vocabulary** (*vocabulary*) The name of the controlled vocabulary used, if any.
 - **URL** (*uri*) A link to the controlled vocabulary mentioned in field 'vocabulary'.
- **country** (*country*) The country shown in the image, if applicable. This information is highly relevant and will often be used as a filter (facet) in data catalogs. It is thus a "Recommended" field. An image will only represent part of a country, but we still want to capture this information in the metadata. Note that many organizations have their own policies on the naming and spelling of countries/regions/economies/territories, which data curators will have to comply with. In rare instances, the image may cover more than one country. The element is repeatable; multiple countries can be entered.
 - **Name** (*name*) The name of the country/economy where the photo was taken.
 - **Code** (*code*) The code of the country/economy mentioned in **name**. This will preferably be the ISO country code.

- **Geographic coverage** (*coverage*) In the Dublin Core, the coverage can be either temporal or geographic. In the use of the schema, `coverage` is used to document the geographic coverage of the image. This element complements the `country` element, and allows more specific information to be provided.
- **GPS position** The geographic location where the photo was taken. Some digital cameras equipped with GPS can, when the option is activated, capture and store in the EXIF metadata the exact geographic location where the photo was taken.
 - **GPS latitude** (*latitude*) The latitude of the geographic location where the photo was taken.
 - **GPS longitude** (*longitude*) The longitude of the geographic location where the photo was taken.
 - **GPS altitude** (*altitude*) The altitude of the geographic location where the photo was taken.
- **Format** (*format*) This refers to the image file format. It is typically expressed using a MIME format.
- **Languages** (*languages*) The language(s) in which the image metadata (caption, title) is provided. This is a block of two elements (at least one must be provided for each language).
 - **Name** (*name*) The name of the language.
 - **Code** (*code*) The code of the language. The use of [ISO 639-2](#) (the alpha-3 code in Codes for the representation of names of languages) is recommended. Numeric codes must be entered as strings.
- **Source** (*source*) A related resource from which the described image is derived.
- **Relations** (*relations*) A list of related resources (images or of other type)
 - **Name** (*name*) The name (title) of the related resource.
 - **Type** (*type*) A brief description of the type of relation. A controlled vocabulary could be used.
 - **URL** (*uri*) A link to the related resource being described.
- **Note** (*note*) Any additional information on the image, not captured in one of the other metadata elements.

AUTHORS AND RIGHTS

- **Creator** (*creator*) The name of the person (or organization) who has taken the photo or created the image.
- **Contributor** (*contributor*) The contributor could be a person or organization, possibly a sponsoring organizations.
- **Publisher** (*publisher*) The person or organization who publish the image.
- **Rights** (*rights*) The copyrights for the photograph. License is in another (common) element.

This completes the Dublin Core set of metadata elements. When this information is complete, continue with the other sections (sections common to both the Dublin Core and the IPTC options).

Documenting an image using the IPTC option

This section describes the documentation of an image using the IPTC option.

Create a new project

Create a new project by clicking on [CREATE NEW PROJECT](#) in the *My projects* page, and select option *Image* as data type.

Create new project

- Microdata
- Timeseries
- Timeseries database
- Document
- Table
- Image**
- Script
- Video
- Geospatial

CLOSE

In the new *project home page*, select an IPTC template (one IPTC template is provided with the application).

The screenshot shows the 'Metadata Editor' interface. On the left is a sidebar with navigation links: Required, Recommended, Empty, Home, Metadata information, Image identifiers, IPTC, License, Album, Tags, and External resources. A search bar is also present. The main content area displays a project titled 'Market near Ramallah's main mosque'. It includes a thumbnail image of a pile of tomatoes, basic metadata fields for Project owner (John Doe), Created on (2025-02-14 11:49:16), Last changed by (John Doe), Changed on (2025-02-14 14:06:16), and Project IDNO (2c63f358-E5ef-4bb3-A654-A8968b6ba694). Below this, there are sections for Template (set to IMAGE IHSN SCHEMA (IPTC OPTION) 1.0 EN), Collaborators (None), Collections (None), and Data and Documentation (with tabs for DOCUMENTATION and FILES). A note at the bottom right indicates Disk usage: 1.09 MB.

Edit the thumbnail by selecting an image. It is highly recommended to select the image being documented as thumbnail. The thumbnail will be used in the Metadata Editor itself, and in a NADA catalog is the image is published in NADA. The system will save the thumbnail as a low-resolution image.

Fill out the *Information on metadata* section

The screenshot shows the Metadata Editor interface. The top bar includes a logo, 'Metadata Editor', and links for 'About', 'English', and 'John Doe'. The left sidebar has a tree view with 'Home', 'Metadata information' (selected), 'Image identifiers', 'IPTC', 'License', 'Album', 'Tags', and 'External resources'. Under 'Metadata information', 'Information on metadata' is expanded, showing 'Document title' (checked), 'Document ID', 'Metadata producers' (table with one row: Name 'John Doe'), 'Production date' (2025-02-14), and 'Version'. A 'SAVE' button is at the top right.

Fill out the IPTC metadata and additional sections (image description)

We provide here some description and recommendations for the key metadata elements in the default IPTC metadata template provided in the Metadata Editor. The labels provided between parenthesis are the element names in the IPTC standard. The grouping of elements by section is the one in the template, not in the standard.

TITLE STATEMENT

- **Title** (*tite*) The title is a shorthand reference for the digital image. It provides a short verbal and human readable name which can be a text and/or a numeric reference. It is not the same as the Headline (see below). Some may use the **title** field to store the file name of the image, though the field may be used in many ways. This element should not be used to provide the unique identifier of the image.
- **Headline** (*headline*) A brief publishable summary of the contents of the image. Note that a headline is not the same as a title.
- **Globally unique identifier** (*digitalImageGuid*) A globally unique identifier for the image. This identifier is created and applied by the creator of the digital image at the time of its creation. This value shall not be changed after that time. The identifier can be generated using an algorithm that would guarantee that the created identifier is globally unique. Device that create digital images like digital or video cameras or scanners usually create such an identifier at the time of the creation of the digital data, and add it to the metadata embedded in the image file (e.g., the EXIF metadata). IPTC's requirements for unique ids are as follows:
 - It must be globally unique. Algorithms for this purpose exist.
 - It should identify the camera body.
 - It should identify each individual photo from this camera body.
 - It should identify the date and time of the creation of the picture.
 - It should be secured against tampering.

- **Date created** (*dateCreated*) Designates the date and optionally the time the content of the image was created. For a photo, this will be the date and time the photo was taken. When no information is available on the time, the time is set to 00:00:00. The preferred format for the `dateCreated` element is the truncated DateTime format, for example: 2021-02-22T21:24:06Z

AUTHOR AND CONTRIBUTORS

- **Creator name** (*creatorNames*) Enter details about the creator or creators of this image. The Image Creator must often be attributed in association with any use of the image. The Image Creator, Copyright Owner, Image Supplier and Licenser may be the same or different entities.
- **Creator job title** (*jobtitle*) The job title of the photographer (the person listed in `creatorNames`). The use of this element implies that the photographer information (`creatorNames` is not empty).
- **Job ID** (*jobid*) Number or identifier for the purpose of improved workflow handling (control or tracking). This is a user created identifier related to the job for which the image is supplied. Note: As this identifier references a job of the receiver's workflow it must first be issued by the receiver, then transmitted to the creator or provider of the news object and finally added by the creator to this field.
- **Source** (*Source*) The name of a person or party who has a role in the content supply chain. The `source` can be different from the `creator` and from the entities listed in the Copyright Notice.
- **Supplier** (*supplier*) The supplier of the image (person or organization)
 - **Name** (*name*) The name of the supplier of the image (person or organization).
 - **Identifier** (*identifiers*) The identifier for the most recent supplier of this image. This will not necessarily be the creator or the owner of the image.
- **Supplier ID** (*imageSupplierImageId*) A unique identifier assigned by the image supplier to the image.
- **Credit** (*creditLine*) The credit to person(s) and/or organization(s) required by the supplier of the image to be used when published. This is a free-text field.
- **Caption writer** (*captionWriter*) An identifier, or the name, of the person involved in writing, editing or correcting the description of the image.
- **Contact** (*creatorContactInfo*) The creator's contact information provides all necessary information to get in contact with the creator of this image and comprises a set of elements for proper addressing. Note that if the creator is also the licensor, his or her contact information should be provided in the `licensor` fields.
 - **Country** (*country*) The country name for the address of the person that created this image.
 - **Email (work)** (*emailwork*) The work email address(es) for the creator of the image. Multiple email addresses can be given, in which case they should be separated by a comma.
 - **Region (State/Province)** (*region*) The state or province for the address of the creator of the image.
 - **Phone (work)** (*phonework*) The work phone number(s) for the creator of the image. Use the international format including the country code, such as +1 (123) 456789. Multiple numbers can be given, in which case they should be separated by a comma.
 - **URL** (*weburlwork*) The work web address for the creator of the image. Multiple addresses can be given, in which case they should be separated by a comma.
 - **Address** (*address*) The address of the creator of the image. This may comprise a company name.
 - **City** (*city*) The city for the address of the person that created the image.
 - **Postal code** (*postalcode*) Enter the local postal code for the address of the person who created the image.

CONTENT DESCRIPTION

- **Description** (*description*) A textual description, including captions, of the image. This describes the who, what, and why of what is happening in this image. This might include names of people, and/or their role in the action that is taking place within the image. Example: "The president of the Metadata Association delivers the keynote address".
- **Scene codes** (*sceneCodes*) The *sceneCodes* describe the scene of a photo content. The [IPTC Scene-NewsCodes](#) controlled vocabulary (published under a Creative Commons Attribution (CC BY) 4.0 license) should be used, where a scene is represented as a string of 6 digits.

| code | Label | Description |
|--------|----------------|--|
| 010100 | headshot | A head only view of a person (or animal/s) or persons as in a montage. |
| 010200 | half-length | A torso and head view of a person or persons. |
| 010300 | full-length | A view from head to toe of a person or persons |
| 010400 | profile | A view of a person from the side |
| 010500 | rear view | A view of a person or persons from the rear. |
| 010600 | single | A view of only one person, object or animal. |
| 010700 | couple | A view of two people who are in a personal relationship, for example engaged, married or in a romantic partnership. |
| 010800 | two | A view of two people |
| 010900 | group | A view of more than two people |
| 011000 | general view | An overall view of the subject and its surrounds |
| 011100 | panoramic view | A panoramic or wide angle view of a subject and its surrounds |
| 011200 | aerial view | A view taken from above |
| 011300 | under-water | A photo taken under water |
| 011400 | night scene | A photo taken during darkness |
| 011500 | satellite | A photo taken from a satellite in orbit |
| 011600 | exterior view | A photo that shows the exterior of a building or other object |
| 011700 | interior view | A scene or view of the interior of a building or other object |
| 011800 | close-up | A view of, or part of a person/object taken at close range in order to emphasize detail or accentuate mood. Macro photography. |
| 011900 | action | Subject in motion such as children jumping, horse running |
| 012000 | performing | Subject or subjects on a stage performing to an audience |

| code | Label | Description |
|--------|-------------|---|
| 012100 | posing | Subject or subjects posing such as a "victory" pose or other stance that symbolizes leadership. |
| 012200 | symbolic | A posed picture symbolizing an event - two rings for marriage |
| 012300 | off-beat | An attractive, perhaps fun picture of everyday events - dog with sunglasses, people cooling off in the fountain |
| 012400 | movie scene | Photos taken during the shooting of a movie or TV production. |

- **Scene codes labelled** (*SceneCodesLabelled*) The **Scene codes** element described above only allows for the capture of codes. To improve discoverability (by indexing important keywords), not only the scene codes but also the scene description should be provided. The IPTC standard does not provide an element that allows the scene label and description to be entered. The **Scene codes labelled** is an element that we added to our schema. Ideally, curators will enter the scene codes in the element **Scene codes** to maintain full compatibility with the IPTC, and complement that information by also entering the codes and their description in the **Scene codes labelled** element.
 - **Code** (*code*) The code for the scene of a photo content. The [IPTC Scene-NewsCodes](#) controlled vocabulary (published under a Creative Commons Attribution (CC BY) 4.0 license) should be used, where a scene is represented as a string of 6 digits. See table above.
 - **Label** (*label*) The label of the scene. See table above for examples.
 - **Description** (*description*) A more detailed description of the scene. See table above for examples.
- **Subject codes** (*subjectCodes*) Specifies one or more subjects from the [IPTC Subject-NewsCodes](#) controlled vocabulary to categorize the image. Each Subject is represented as a string of 8 digits. The vocabulary consists of about 1400 terms organized into 3 levels (users can decide to use only the first, or the first two levels; the more detail is provided, the better the discoverability of the image). The first level of the controlled vocabulary is as follows:

| code | Label | Description |
|----------|---------------------------------|---|
| 01000000 | arts, culture and entertainment | Matters pertaining to the advancement and refinement of the human mind, of interests, skills, tastes and emotions |
| 02000000 | crime, law and justice | Establishment and/or statement of the rules of behavior in society, the enforcement of these rules, breaches of the rules and the punishment of offenders. Organizations and bodies involved in these activities. |
| 03000000 | disaster and accident | Man made and natural events resulting in loss of life or injury to living creatures and/or damage to inanimate objects or property. |
| 04000000 | economy, business and finance | All matters concerning the planning, production and exchange of wealth. |
| 05000000 | education | All aspects of furthering knowledge of human individuals from birth to death. |
| 06000000 | environmental | All aspects of protection, damage, and condition of the |

| code | Label | Description |
|----------|------------------------|--|
| | issue | ecosystem of the planet earth and its surroundings. |
| 07000000 | health | All aspects pertaining to the physical and mental welfare of human beings. |
| 08000000 | human interest | Lighter items about individuals, groups, animals or objects. |
| 09000000 | labor | Social aspects, organizations, rules and conditions affecting the employment of human effort for the generation of wealth or provision of services and the economic support of the unemployed. |
| 10000000 | lifestyle and leisure | Activities undertaken for pleasure, relaxation or recreation outside paid employment, including eating and travel. |
| 11000000 | politics | Local, regional, national and international exercise of power, or struggle for power, and the relationships between governing bodies and states. |
| 12000000 | religion and belief | All aspects of human existence involving theology, philosophy, ethics and spirituality. |
| 13000000 | science and technology | All aspects pertaining to human understanding of nature and the physical world and the development and application of this knowledge |
| 14000000 | social issue | Aspects of the behavior of humans affecting the quality of life. |
| 15000000 | sport | Competitive exercise involving physical effort. Organizations and bodies involved in these activities. |
| 16000000 | unrest | conflicts and war Acts of socially or politically motivated protest and/or violence. |
| 17000000 | weather | The study, reporting and prediction of meteorological phenomena. |

As an example of subjects at the three levels, the list below zooms on the subject "education".

| code | Subject | Description |
|----------|-----------------|--|
| 05000000 | education | All aspects of furthering knowledge of human individuals from birth to death |
| 05001000 | Adult education | Education provided for older students outside the usual age groups of 5-25 |

| code | Subject | Description |
|----------|-----------------------|---|
| 05002000 | Further education | Any form of education beyond basic education of several levels |
| 05003000 | parent organization | Groups of parents set up to support schools |
| 05004000 | preschool | Education for children under the national compulsory education age |
| 05005000 | school | A building or institution in which education of various sorts is provided |
| 05005001 | elementary schools | Schools usually of a level from kindergarten through 11 or 12 years of age |
| 05005002 | middle schools | Transitional school between elementary and high school, 12 through 13 years of age |
| 05005003 | high schools | Pre-college/ university level education 14 to 17 or 18 years of age, called freshman, sophomore, junior and senior |
| 05006000 | teachers union | Organization of teachers for collective bargaining and other purposes |
| 05007000 | university | Institutions of higher learning capable of providing doctorate degrees |
| 05008000 | upbringing | Lessons learned from parents and others as one grows up |
| 05009000 | entrance examination | Exams for entering colleges, universities, junior and senior high schools, and all other higher and lower education institutes, including cram schools, which help students prepare for exams for entry to prestigious schools. |
| 05010000 | teaching and learning | Either end of the education equation |
| 05010001 | students | People of any age in a structured environment, not necessarily a classroom, in order to learn something |
| 05010002 | teachers | People with knowledge who can impart that knowledge to others |
| 05010003 | curriculum | The courses offered by a learning institution and the regulation of those courses |
| 05010004 | test/examination | A measurement of student accomplishment |

| code | Subject | Description |
|----------|---------------------|--|
| 05011000 | religious education | Instruction by any faith, in that faith or about other faiths, usually, but not always, conducted in schools run by religious bodies |
| 05011001 | parochial school | A school run by the Roman Catholic faith |
| 05011002 | seminary | A school of any faith specifically designed to train ministers |
| 05011003 | yeshiva | A school for training rabbis |
| 05011004 | madrasa | A school for teaching Islam |

- **Subject codes labelled** (`subjectCodesLabelled`) The `Subject codes` element described above only allows for the capture of codes. To improve discoverability (by indexing important keywords), not only the subject codes but also the subject description should be provided. The IPTC standard does not provide an element that allows the subject label and description to be entered. The `subjectCodesLabelled` is an element that we added to our schema. Ideally, curators will enter the subject codes in the element `subjectCodes` to maintain full compatibility with the IPTC, and complement that information by also entering the codes and their description in the `Subject codes labelled` element.
 - **Code** (`code`) Specifies one or more subjects from the [IPTC Subject-NewsCodes](#) controlled vocabulary to categorize the image. Each Subject is represented as a string of 8 digits. The vocabulary consists of about 1400 terms organized into 3 levels (users can decide to use only the first, or the first two levels; the more detail is provided, the better the discoverability of the image). See examples in the table above.
 - **Label** (`label`) The label of the subject. See table above for examples.
 - **Description** (`description`) A more detailed description of the subject. See table above for examples.
- **Keywords** (`keywords`) Keywords (terms or phrases) to express the subject of the image. Keywords do not have to be taken from a controlled vocabulary.
- **Topics** (`aboutCvTerms`) One or more topics, themes or entities the content is about, each one expressed by a term from a controlled vocabulary.
 - **Vocabulary ID** (`cvId`) The globally unique identifier of the Controlled Vocabulary the term is from.
 - **Term Label** (`cvTermName`) The natural language name of the term from a Controlled Vocabulary.
 - **Term ID** (`cvTermId`) The globally unique identifier of the term from a Controlled Vocabulary.
 - **Details** (`cvTermRefinedAbout`) Refined 'about' relationship of the CV-Term. The refined 'about' relationship of the term with the content. Optionally enter a refinement of the 'about' relationship of the term with the content of the image. This must be a globally unique identifier from a Controlled Vocabulary.
- **Event name** (`eventName`) The name or a brief description of the event where the image was taken. If this is a sub-event of a larger event, mention both in the description. For example: "Opening statement, 1st International Conference on Metadata Standards, New York, November 2021".
- **Location shown** (`locationsShown`) This block of elements is used to document the location shown in the image. This information should be provided with as much detail as possible. It contains elements that can be used to provide a "nested" description of the location, from a high geographic level (world region) down to a very specific location (city and sub-location within a city).
 - **Name** (`name`) The full name of the location.
 - **Identifier** (`identifiers`) A globally unique identifier of the location shown.

- **World region** (*worldRegion*) The name of a world region. This element is at the first (top) level of the top-down geographical hierarchy.
- **Country name** (*countryName*) The name of a country of a location. This element is at the second level of a top-down geographical hierarchy.
- **Country code** (*countryCode*) The ISO code of the country mentioned in [countryName](#).
- **Sub-region** (*provinceState*) The name of a sub-region of the country - for example a province or a state name. This element is at the third level of a top-down geographical hierarchy.
- **City** (*city*) The name of the city. This element is at the fourth level of a top-down geographical hierarchy.
- **Sub-location** (*sublocation*) The sublocation name could either be the name of a sublocation to a city or the name of a well known location or (natural) monument outside a city. This element is at the fifth (lowest) level of a top-down geographical hierarchy.
- **GPS altitude** (*gpsAltitude*) The altitude in meters of a WGS84 based position of this location.
- **GPS latitude** (*gpsLatitude*) Latitude of a WGS84 based position of this location (in some cases, this information may be contained in the EXIF metadata).
- **GPS longitude** (*gpsLongitude*) Longitude of a WGS84 based position of this location (in some cases, this information may be contained in the EXIF metadata).
- **Genres** (*genres*) Artistic, style, journalistic, product or other genre(s) of the image (expressed by a term from any Controlled Vocabulary)
 - **Vocabulary ID** (*cvId*) The globally unique identifier of the Controlled Vocabulary the term is from.
 - **term label** (*cvTermName*) The natural language name of the term from a Controlled Vocabulary.
 - **Term ID** (*cvTermId*) The globally unique identifier of the term from a Controlled Vocabulary.
 - **About** (*cvTermRefinedAbout*) Optionally enter a refinement of the 'about' relationship of the term with the content of the image. This must be a globally unique identifier from a Controlled Vocabulary. May be used to refine the generic about relationship.
- **Intellectual genre** (*intellectualGenre*) A term to describe the nature of the image in terms of its intellectual or journalistic characteristics (for example "actuality", "interview", "background", "feature", "summary", "wrapup" for journalistic genres, or "daybook", "obituary", "press release", "transcript" for news category related genres. It is advised to use terms from a controlled vocabulary such as the [NewsCodes Scheme](#) published by the IPTC under a Creative Commons Attribution (CC BY) 4.0 license.

| Genre | Description |
|---------------------|--|
| Actuality | Recording of an event |
| Advertiser Supplied | Content is supplied by an organization or individual that has paid the news provider for its placement |
| Advice | Letters and answers about readers' personal problems |
| Advisory | Recommendation on editorial or technical matters by a provider to its customers |
| On This Day | List of data, including birthdays of famous people and items of historical significance, for a given day |
| Analysis | Data and conclusions drawn by a journalist who has conducted in depth research for a story |

| Genre | Description |
|--------------------|---|
| Archival material | Material selected from the originator's archive that has been previously distributed |
| Background | Scene setting and explanation for an event being reported |
| Behind the Story | The content describes how a story was reported and offers context on the reporting |
| Biography | Facts and background about a person |
| Birth Announcement | News of newly born children |
| Current Events | Content about events taking place at the time of the report |
| Curtain Raiser | Information about the staging and outcome of an immediately upcoming event |
| Daybook | Items filed on a regular basis that are lists of upcoming events with time and place, designed to inform others of events for planning purposes. |
| Exclusive | Information content, in any form, that is unique to a specific information provider. |
| Fact Check | The news item looks into the truth or falsehood of another reported news item or assertion (for example a statement on social media by a public figure) |
| Feature | The object content is about a particular event or individual that may not be significant to the current breaking news. |
| Fixture | The object contains data that occurs often and predictably. |
| Forecast | The object contains opinion as to the outcome of a future event. |
| From the Scene | The object contains a report from the scene of an event. |
| Help us to Report | The news item is a call for readers to provide information that may help journalists to investigate a potential news story |
| History | The object content is based on previous rather than current events. |
| Horoscope | Astrological forecasts |
| Interview | The object contains a report of a dialogue with a news source that gives it significant voice (includes Q and A). |
| Listing of facts | Detailed listing of facts related to a topic or a story |
| Music | The object contains music alone. |

| Genre | Description |
|-------------------------------------|---|
| Obituary | The object contains a narrative about an individual's life and achievements for publication after his or her death. |
| Opinion | The object contains an editorial comment that reflects the views of the author. |
| Polls and Surveys | The object contains numeric or other information produced as a result of questionnaires or interviews. |
| Press Release | The object contains promotional material or information provided to a news organisation. |
| Press-Digest | The object contains an editorial comment by another medium completely or in parts without significant journalistic changes. |
| Profile | The object contains a description of the life or activity of a news subject (often a living individual). |
| Program | A news item giving lists of intended events and time to be covered by the news provider. Each program covers a day, a week, a month or a year. The covered period is referenced as a keyword. |
| Question and Answer Session | The object contains the interviewer and subject questions and answers. |
| Quote | The object contains a one or two sentence verbatim in direct quote. |
| Raw Sound | The object contains unedited sounds. |
| Response to a Question | The object contains a reply to a question. |
| Results Listings and Statistics | The object contains alphanumeric data suitable for presentation in tabular form. |
| Retrospective | The object contains material that looks back on a specific (generally long) period of time such as a season, quarter, year or decade. |
| Review | The object contains a critique of a creative activity or service (for example a book, a film or a restaurant). |
| Satire | Uses exaggeration, irony, or humor to make a point; not intended to be understood as factual |
| Scener | The object contains a description of the event circumstances. |
| Side bar and supporting information | Related story that provides additional context or insight into a news event |

| Genre | Description |
|-------------------------|--|
| Special Report | In-depth examination of a single subject requiring extensive research and usually presented at great length, either as a single item or as a series of items |
| Sponsored | Content is produced on behalf of an organization or individual that has paid the news provider for production and may approve content publication |
| Summary | Single item synopsis of a number of generally unrelated news stories |
| Supported | Content is produced with financial support from an organization or individual, yet not approved by the underwriter before or after publication |
| Synopsis | The object contains a condensed version of a single news item. |
| Text only | The object contains a transcription of text. |
| Transcript and Verbatim | A word for word report of a discussion or briefing |
| Update | The object contains an intraday snapshot (as for electronic services) of a single news subject. |
| Voice | Content is only voice |
| Wrap | Complete summary of an event |
| Wrapup | Recap of a running story |

TERMS OF USE AND RIGHTS

- **Copyright notice** (`copyrightNotice`) Contains any necessary copyright notice for claiming the intellectual property for this photograph and should identify the current owner of the copyright for the photograph. Other entities like the creator of the photograph may be added in the corresponding field. Notes on usage rights should be provided in "Rights usage terms". Example: ©2008 Jane Doe. If the copyright ownership must be expressed in a more controlled manner, use the fields "Copyright Owner", "Copyright Owner ID", "Copyright Owner Name" described below instead of the `copyrightNotice` element.
- **Copyright owners** (`copyrightOwners`) Owner or owners of the copyright in the licensed image, described in a structured format (as an alternative to the element `copyrightNotice` described above). This block serves the same purpose of identifying the rights holder/s for the image. The Copyright Owner, Image Creator and Licensor may be the same or different entities.
 - **Name** (`name`) The name of the owner of the copyright in the licensed image.
 - **Role** (`role`) The role the entity.
 - **Identifier** (`identifiers`) The identifier of the owner of the copyright in the licensed image.
- **Usage terms** (`usageTerms`) The licensing parameters of the image expressed in free-text. Enter instructions on how this image can legally be used. The PLUS fields of the IPTC Extension can be used in parallel to express the licensed usage in more controlled terms.

- **Rights expression** (*embedEncRightsExpr*) An embedded rights expression using a rights expression language which is encoded as a string. (Embedded Encoded Rights Expression (EERE) structure)
 - **Encoded rights expression** (*encRightsExpr*) Rights Expression Language ID. An identifier of the rights expression language used by the rights expression.
 - **Encoding type** (*rightsExprEncType*) The encoding type of the rights expression, identified by an IANA Media Type.
 - **Rights expression language ID** (*rightsExprLangId*) An embedded rights expression using any rights expression language. @@@@ <https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata#embedded-encoded-rights-expression-eere-structure>
- **Linked rights expression** (*linkedEncRightsExpr*) Link to Encoded Rights Expression.
 - **Link to encoded rights expression** (*linkedRightsExpr*) The link to a web resource representing an encoded rights expression.
 - **Encoding type** (*rightsExprEncType*) The encoding type of the rights expression, identified by an IANA Media Type.
 - **Rights expression language ID** (*rightsExprLangId*) The identifier of the rights expression language used by the rights expression.
- **Web statement rights** (*webstatementRights*) URL referencing a web resource providing a statement of the copyright ownership and usage rights of the image.
- **Instructions to users** (*instructions*) Any of a number of instructions from the provider or creator to the receiver of the image which might include any of the following: embargoes and other restrictions not covered by the "Rights Usage Terms" field; information regarding the original means of capture (scanning notes, colourspace info) or other specific text information that the user may need for accurate reproduction; additional permissions required when publishing; credits for publishing if they exceed the IIM length of the credit field.
- **Property release status** (*propertyReleaseStatus*) This summarizes the availability and scope of property releases authorizing usage of the properties appearing in the photograph. One value should be selected from a controlled vocabulary. It is recommended to apply the value PR-UPR very carefully and to check the wording of the property release thoroughly before applying it.
 - **Vocabulary ID** (*cvId*) The globally unique identifier of the Controlled Vocabulary the term is from.
 - **Term label** (*cvTermName*) The natural language name of the term from a Controlled Vocabulary.
 - **Term ID** (*cvTermId*) The globally unique identifier of the term from a Controlled Vocabulary.
 - **About** (*cvTermRefinedAbout*) Refined 'about' relationship of the CV-Term. The refined 'about' relationship of the term with the content. Optionally enter a refinement of the 'about' relationship of the term with the content of the image. This must be a globally unique identifier from a Controlled Vocabulary.
- **Property release documents** (*propertyReleaseDocuments*)
- **Contact information** (*creatorContactInfo*) Contact information for licensing and rights.
 - **Country** (*country*) The country name for the address of the person that created this image.
 - **Email (work)** (*emailwork*) The work email address(es) for the creator of the image. Multiple email addresses can be given, in which case they should be separated by a comma.
 - **Region (State/Province)** (*region*) The state or province for the address of the creator of the image.
 - **Phone (work)** (*phonework*) The work phone number(s) for the creator of the image. Use the international format including the country code, such as +1 (123) 456789. Multiple numbers can be given, in which case they should be separated by a comma.
 - **URL** (*weburlwork*) The work web address for the creator of the image. Multiple addresses can be given, in which case they should be separated by a comma.

- **Address** (*address*) The address of the creator of the image. This may comprise a company name.
- **City** (*city*) The city for the address of the person that created the image.
- **Postal code** (*postalCode*) Enter the local postal code for the address of the person who created the image.

ART WORK, OBJECTS, PRODUCTS IN IMAGE

- **Art work or object** (*artworkOrObjects*) This block provides a set of metadata elements to be used to describe the object or artwork shown in the image.
 - **Title** (*title*) A human readable name of the object or artwork shown in the image.
 - **Content description** (*contentDescription*) A textual description of the content depicted in the object or artwork.
 - **Physical description** (*physicalDescription*) A textual description of the physical characteristics of the artwork or object, without reference to the content depicted. This would be used to describe the object type, materials, techniques, and measurements.
 - **Creator name** (*creatorNames*) The name of the person(s) (possibly an organization) who created the object or artwork shown in the image.
 - **Creator identifier** (*creatorIdentifiers*) One or multiple globally unique identifier(s) for the artist who created the artwork or object shown in the image. This could be an identifier issued by an online registry of persons or companies. Make sure to enter these identifiers in the exact same sequence as the names entered in the field *creatorNames*.
 - **Contribution description** (*contributionDescription*) A description of any contributions made to the artwork or object. It should include the type, date and location of contribution, and details about the contributor.
 - **Style period** (*stylePeriod*) The style, historical or artistic period, movement, group, or school whose characteristics are represented in the artwork or object. It is advised to take the terms from a Controlled Vocabulary.
 - **Date created** (*dateCreated*) The date and optionally the time the artwork or object shown in the image was created.
 - **Circa date created** (*circaDateCreated*) The approximate date or range of dates associated with the creation and production of an artwork or object or its components.
 - **Source** (*source*) The name of the organization or body holding and registering the artwork or object in this image for inventory purposes.
 - **Source inventory number** (*sourceInventoryNr*) The inventory number issued by the organization or body holding and registering the artwork or object in the image.
 - **Source inventory URL** (*sourceInventoryUrl*) A reference URL for the metadata record of the inventory maintained by the Source.
 - **Current copyright owner name** (*currentCopyrightOwnerName*) The name of the current owner of the copyright of the artwork or object.
 - **Current copyright owner identifier** (*currentCopyrightOwnerId*) A globally unique identifier for the current copyright owner e.g. issued by an online registry of persons or companies.
 - **Copyright notice** (*copyrightNotice*) Any necessary copyright notice for claiming the intellectual property for artwork or an object in the image and should identify the current owner of the copyright of this work with associated intellectual property rights.
 - **Current licensor name** (*currentLicensorName*) Name of the current licensor of the artwork or object.
 - **Current licensor ID** (*currentLicensorIdentifier*) A globally unique identifier for the current licensor e.g. issued by an online registry of persons or companies.
- **Products shown** (*productsShown*) Details about a product shown in the image.
 - **Name** (*name*) The name of the product.
 - **Description** (*description*) A textual description of the product.

- **GTIN** (gtin) The [Global Trade Item Number \(GTIN\)](#) of the product (GTIN-8 to GTIN-14 codes can be used).

PERSONS AND MODELS IN IMAGE

- **Persons in image (list)** (personInImageNames) This repeatable block of elements is used to provide information on the person(s) shown in the image.
- **Persons shown (itemized)** (personsShown) Details about person(s) shown in the image. It is not required to list all, just those details which can be recognized.
 - **Name** (name) The name of a person shown in the image.
 - **Description** (description) A textual description of the person. For example, you may include actions taken, emotional expressions shown and more.
 - **Identifiers** (identifiers) Globally Unique identifiers of the person, such as those from [WikiData](#).
 - **Characteristics** (characteristics) A property or trait of the person, provided as a term selected from a Controlled Vocabulary.
 - **Term label** (cvId) The globally unique identifier of the Controlled Vocabulary the term is from.
 - **Term ID** (cvTermName) The natural language name of the term from a Controlled Vocabulary.
 - **Vocabulary ID** (cvTermId) The globally unique identifier of the term from a Controlled Vocabulary.
 - **About** (cvTermRefinedAbout) The refined 'about' relationship of the term with the content. Optionally enter a refinement of the 'about' relationship of the term with the content of the image. This must be a globally unique identifier from a Controlled Vocabulary.
- **Model ages** (modelAges) Age of the human model(s) at the time the image was taken. Be aware of any legal implications of providing ages for young models. Ages below 18 years should not be included.
- **Additional model information** (additionalModelInfo) Information about other facets of the model(s).
- **Minor model age disclosure** (minorModelAgeDisclosure) The age of the youngest model pictured in the image, at the time the image was created. This information is not intended to be displayed publicly; it is intended to be used as a filter for inclusion/exclusion of images in catalogs and dissemination processes.
- **Model release documents** (modelReleaseDocuments) Identifier associated with each Model Release.
- **Model release status** (modelReleaseStatus) Summarizes the availability and scope of model releases authorizing usage of the likenesses of persons appearing in the photograph.
 - **Term label** (cvId) The globally unique identifier of the Controlled Vocabulary the term is from.
 - **Term ID** (cvTermName) The natural language name of the term from a Controlled Vocabulary.
 - **Vocabulary ID** (cvTermId) The globally unique identifier of the term from a Controlled Vocabulary.
 - **About** (cvTermRefinedAbout) The refined 'about' relationship of the term with the content. Optionally enter a refinement of the 'about' relationship of the term with the content of the image. This must be a globally unique identifier from a Controlled Vocabulary. May be used to refine the generic about relationship.

ORGANIZATIONS IN IMAGE

- **Name of the organization** (organisationInImageCodes) The code, extracted from a controlled vocabulary, used to identify the organization or company featured in the image. For example a stock ticker symbol may be used. Enter an identifier for the controlled vocabulary, then a colon, and finally the code from the vocabulary assigned to the organization (e.g. nasdaq:companyA)
- **Code of the organization** (organisationInImageNames) The name of the organization or company which is featured in the image.

TECHNICAL INFORMATION

- **Image rating** (*imageRating*) Rating of the image by its user or supplier. The value shall be -1 or in the range 0 to 5. -1 indicates "rejected" and 0 "unrated". If an explicit value is not provided, the default value is 0 will be assumed.
- **Digital source type** (*digitalSourceType*) The type of the source of this digital image. One value should be selected from the IPTC controlled vocabulary (published under a Creative Commons Attribution (CC BY) 4.0 license license) that contains the following values:

| Type | Source | Description |
|----------------|--|---|
| digitalCapture | Original digital capture of a real life scene | The digital image is the original and only instance and was taken by a digital camera |
| negativeFilm | Digitized from a negative on film | The digital image was digitized from a negative on film on any other transparent medium |
| positiveFilm | Digitized from a positive on film | The digital image was digitized from a positive on a transparency or any other transparent medium |
| print | Digitized from a print on non-transparent medium | The digital image was digitized from an image printed on a non-transparent medium |
| softwareImage | Created by software | The digital image was created by computer software |

- **Maximum available height** (*maxAvailHeight*) The maximum available height in pixels of the original photo from which this photo has been derived by downsizing.
- **Maximum available width** (*maxAvailWidth*) The maximum available width in pixels of the original photo from which this photo has been derived by downsizing.
- **Registry entries** (*registryEntries*) A structured element used to provide cataloguing information (i.e. an entry in a registry). It includes the unique identifier for the image issued by the registry and the registry's organization identifier.
 - **Role** (*role*) An identifier of the reason and/or purpose for this Registry Entry.
 - **Asset identifier** (*assetIdentifier*) A unique identifier created by the registry and applied by the creator of the digital image. This value shall not be changed after being applied. This identifier is linked to a corresponding Registry Organization Identifier. Enter the unique identifier created by a registry and applied by the creator of the digital image. This value shall not be changed after being applied. This identifier may be globally unique by itself, but it must be unique for the issuing registry. An input to this field should be made mandatory.
 - **Registry identifier** (*registryIdentifier*) An identifier for the registry/organization which issued the corresponding Registry Image Id.

This completes the IPTC set of metadata elements. When this information is complete, continue with the other sections (sections common to both the Dublin Core and the IPTC options).

Fill out the License section

This section is common to both the DCMI and IPTC options.

- **License** (*license*) Enter the name and URL of the license under which the image is published (if any).
 - **Name** (*name*) The name of a person shown in the image.
 - **Description** (*description*) A textual description of the person. For example, you may include actions taken, emotional expressions shown and more.

Fill out the *Albums* section

This section is common to both the DCMI and IPTC options.

- **Album** (*album*) Enter information on the name, description, owner, and URL of the online album(s) in which the image is published.
 - **Name** (*name*) The name (label) of the album.
 - **Description** (*description*) A brief description of the album.
 - **Owner** (*owner*) The owner or custodian of the album.
 - **URL** (*URL*) A link (URL) to the album.

Fill out the *Tags* section

This section is common to both the DCMI and IPTC options.

See section **Documenting data - General instructions**.

Add the external resources

This section is common to both the DCMI and IPTC options.

External resources are all materials (and links) that relate to the image. This will typically include the image file itself, possibly in multiple resolutions. If provided in only one resolution, it is recommended to provide the highest resolution available.

Click on **External resources** in the navigation tree, then on **CREATE RESOURCE**. Enter the relevant information on the resource (at least a title), then provide either a filename (the file will then be uploaded on the server that hosts the Metadata Editor) or a URL to the resource.

The screenshot shows the Metadata Editor interface. On the left, there's a sidebar with a search bar and a tree view of resources under categories like Home, Metadata information, Image identifiers, IPTC, License, Album, Tags, and External resources. Under External resources, 'World Bank Flickr Album' is selected. The main content area is titled 'Market near Ramallah's main mosque' and shows an 'Edit resource' form. The form includes fields for Resource type (set to Web Site), Resource format, Title (World Bank Flickr Album), Author, Date (set to 2025-02-14), and Country. At the top right of the form are 'SAVE' and 'CANCEL' buttons.

External resources that have already been created for another project can also be imported. To do that, they must first be exported as JSON or RDF from the other project. The click on **IMPORT** in the External resources page, and select the file.

External resources will be part of the project ZIP package (when the ZIP package is generated - See the main menu).

See also section *Documentation - General instructions*.

Add information on provenance

This section is common to both the DCMI and IPTC options.

The **Provenance** container is used to document how, from where, and when the image was acquired. It is used to ensure traceability. See section *Documenting data - General instructions* for more information.

Save and export metadata (DCMI or IPTC)

Publish metadata (DCMI or IPTC)

Augmenting image metadata using AI

To make images discoverable, metadata that describe the content depicted in an image, the source of the image and the rights and licensing associated with it, are essential but not provided in the EXIF. Additional metadata must be provided.

Some of these metadata will have to be generated by image authors and/or curators, other can be generated in a much automated manner using machine learning models and tools. Image processing algorithms that make it possible to augmented metadata include algorithms of face detection, person identification, automated labeling, text extraction, and others. Before describing the proposed metadata schema in the following sections, we present here some example of tools that make such metadata enhancement easy and affordable.

The example we provide below makes use of the [Google Vision API](#) to generate image metadata. Google Vision is one out of multiple tools that can be used for that purpose such as [Amazon Rekognition](#), or [Microsoft Azure Computer Vision](#). This example makes use of a photo selected from the [World Bank Flickr album](#).



The image comes with a brief description that identifies the photographer, the location (name of the country and town, not GPS location), and the content of the image. The description of the image includes important keywords that, when indexed in a catalog, will support discoverability of the image. This information, to be manually entered, is valuable and must be part of the curated image metadata.


World Bank Photo Collecti...
+ Follow

Partly flooded streets

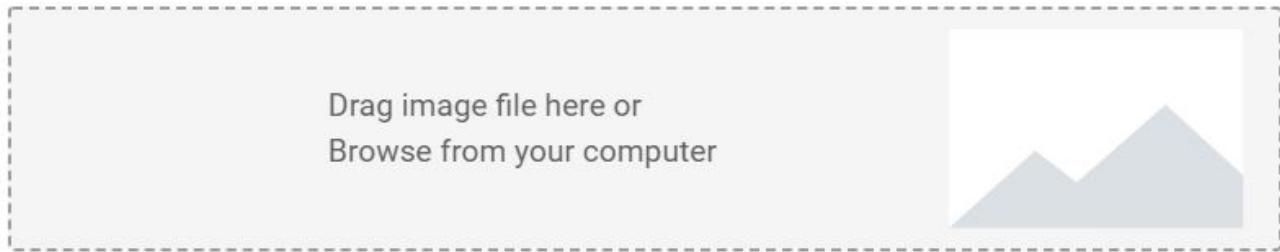
PRO

Partly flooded streets. The floods cause health and security hazards for the residents of low lying areas. Gamarra, Colombia. Photo: Scott Wallace / World Bank

Photo ID: SW-COE-004

But we can add useful additional information in an automated manner and at low cost using machine learning models. In the example below, we use the (free) on-line ["Try it" tool](#) of the Google Vision application.

Try the API



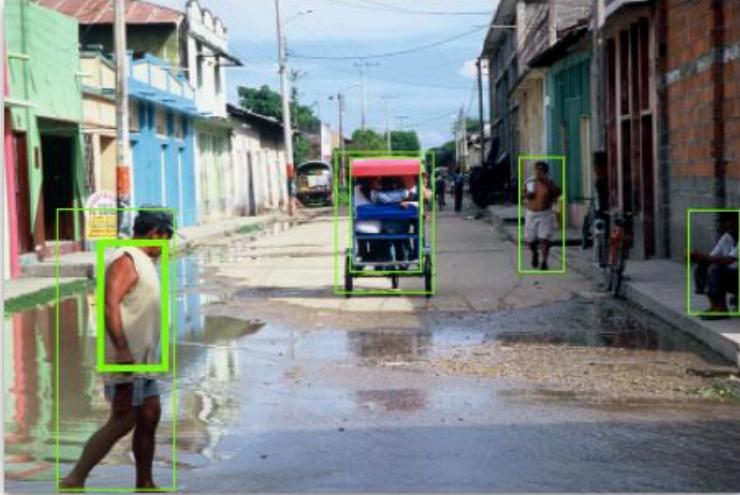
The Google Vision API returns and displays the results of the image processing in multiple tabs. The same content is available programmatically in JSON format. The content of this JSON file can be mapped to elements of the metadata schema, for automatic addition to the image metadata.

The first tab is the result of **faces** detection. Each detected face has a bounding box and metadata such as the derived emotion of the person. The bounding box can be used to automatically flag images that have one or multiple "significant size" face(s) and may have to be excluded from the published images for privacy protection reasons.

| Faces | Objects | Labels | Text | Properties | Safe Search |
|--|---------|--------|------|---|-------------|
|  1543136297_8c73e81c14_c.jpg | | | | Face 1 Joy █ Very Unlikely Sorrow █ Very Unlikely Anger █ Very Unlikely Surprise █ Very Unlikely Exposed ██ Possible Blurred ███ Likely Headwear █ Very Unlikely Roll: 2° Tilt: -6° Pan: -8° Confidence <div style="width: 22%; background-color: green; height: 10px;"></div> 22% | |

The second tab reports on detected **objects**.

Faces Objects Labels Text Properties Safe Search



1543136297_8c73e81c14_c.jpg

| | |
|--------|-----|
| Person | 90% |
| Person | 89% |
| Person | 82% |
| Person | 78% |
| Cart | 75% |
| Top | 67% |

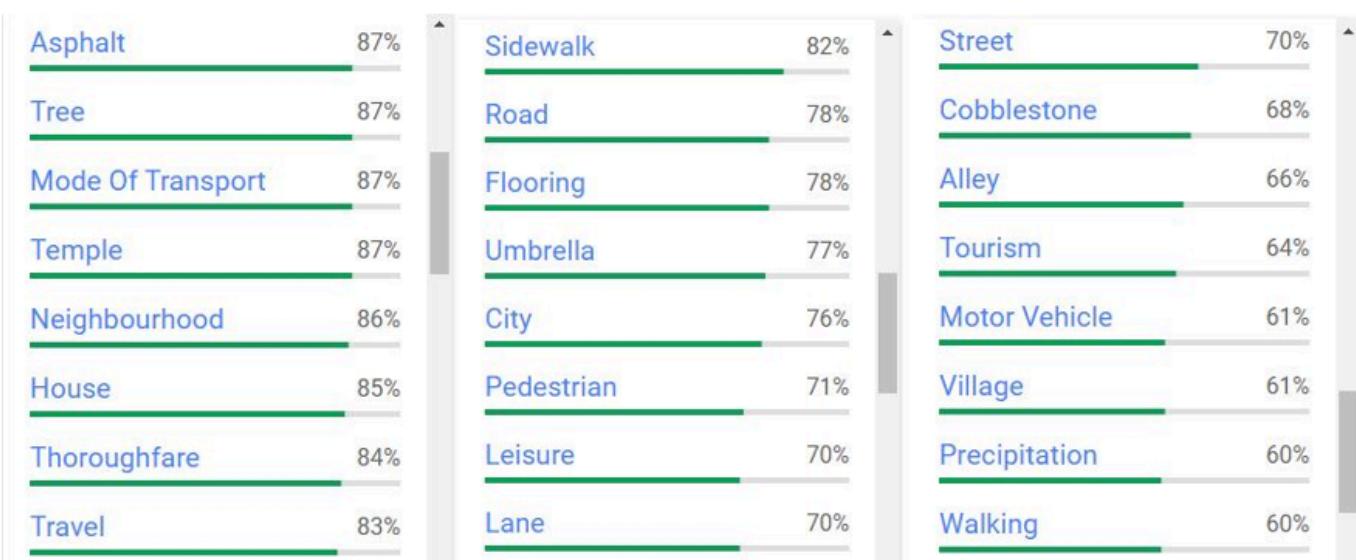
The third tab suggests **Labels** that could be attached to the image, provided with a degree of confidence. A threshold can be set to automatically add (or not) each proposed label as a keyword in the image metadata.

Faces Objects Labels Text Properties Safe Search



1543136297_8c73e81c14_c.jpg

| | |
|----------------|-----|
| Tire | 96% |
| Wheel | 95% |
| Building | 94% |
| Water | 93% |
| Vehicle | 91% |
| Infrastructure | 90% |
| Road Surface | 89% |
| Sky | 89% |



The fourth tab shows the **text** detected in the image. The quality of text detection and recognition depends on the resolution of the image and on the size and orientation of the text in the image. In our example, the algorithm fails to read (most of) the small, rotated and truncated text.



The tool managed to recognize some, but not all characters. In this case, this would be considered as not useful information to be added to the image metadata.

Faces Objects Labels Text Properties Safe Search



1543136297_8c73e81c14_c.jpg

+Page 1

+Block 1

HUN

+Block 2

+Paragraph 1

VITAE

+Block 3

+Paragraph 1

TE ESPEE

+Block 4

We are not interested in the **properties** tab which does not provide information that can be used for discoverability of images based on their content or source.

The last tab, **Safe search**, could be used as warnings if you plan to make the image publicly accessible.

Faces Objects Labels Text Properties Safe Search



1543136297_8c73e81c14_c.jpg

| | | |
|----------|---|---------------|
| Adult | <div style="width: 100%;"><div style="width: 100%;"> </div></div> | Possible |
| Spoof | <div style="width: 10%;"> </div> | Very Unlikely |
| Medical | <div style="width: 10%;"> </div> | Very Unlikely |
| Violence | <div style="width: 10%;"> </div> | Unlikely |
| Racy | <div style="width: 100%;"> </div> | Possible |

Likeliness values are Unknown, Very Unlikely, Unlikely, Possible, Likely, and Very Likely

This "Try it" tool demonstrates the capabilities of the application which, for automating the processing of a collection of images, would be accessed programmatically using R, Python or another programming language. Accessing the application's API requires a key. The cost of image labeling, face detection, and other image processing is low. For information on pricing, consult the website of the API providers.

Documenting a video

The metadata standard

The metadata schema implemented in the Metadata Editor to document video files is a combination of elements extracted from the [Dublin Core Metadata Initiative](#) (DCMI) and from the [VideoObject \(from schema.org\)](#) schemas.

The Dublin Core is a generic and versatile standard, which we also use (in an augmented form) for the documentation of *Documents* and *Images*. It contains 15 core elements, to which we added a selection of elements from *VideoObject*. We also included the elements `keywords`, `topics`, `tags`, `datacite`, `provenance` and `additional` that are found in all standards supported by the Metadata Editor.

Preparing for the documentation and publishing of the video

Generate a transcription

Videos typically come with limited metadata. To make them more discoverable, a transcription of the video content can be generated, stored, and indexed in the catalog. The metadata schema we propose includes an element `transcription` that can store transcriptions (and possibly their automatically-generated translations) in the video metadata. Word embedding models and topic models can be applied to the transcriptions to further augment the metadata. This will significantly increase the discoverability of the resource, and offer the possibility to apply semantic searchability on video metadata.

Machine learning speech-to-text solutions are available (although not for all languages) to automatically generate transcriptions at a low cost. This includes commercial applications like [Whisper by openAI](#), [Microsoft Azure](#), or [Amazon Transcribe](#). Open source Python solutions also exist.

Transcriptions of videos published on Youtube are available on-line (see the example provided in the description of the `Embed URL` element below).

Publish your video online

If you plan to publish the video metadata in a NADA catalog, with a possibility for the catalog users to view the video directly in NADA (i.e., if you want to embed the video in the NADA cataloguing page), the video must be published in a video streaming site like YouTube. NADA can embed a published video, but is not a video streaming application.

Documenting the video

Create a new project

The first step in documenting a video in the Metadata Editor is to create a new project. You do that by clicking on **CREATE NEW PROJECT** in the *My projects* page, then selecting *Video* as data type when prompted. This will open a new, untitled *Project* page.

In that page, select the template you want to use to document the video. A default template is proposed; no action is needed if you want to use the default template. Otherwise, switch to another template by clicking on the template title in the *Templates* frame. Note that you can at any time change the template used for the documentation of a project. The selected template will determine what you see in the navigation tree and in the metadata entry pages, but switching from one template to another will not impact the metadata that has already been entered; no information will be deleted from the metadata.

Enter information on metadata

The metadata information set is used to document the video metadata (not the video itself). This provides information useful for archiving purposes. This set is optional. It is recommended however to enter at least the identification and affiliation of the metadata producer, and the date of creation of the metadata. One reason for this is that metadata can be shared and harvested across catalogs/organizations, so metadata produced by one organization can be found in other data centers.

Enter the video description

The **Video description** section contains all elements that will be used to describe the video and its content. These are the elements that will be indexed and made searchable when published in a data catalog. We provide below a brief description and recommendations for the metadata elements included in the default metadata template in the Metadata Editor.

In the list of metadata elements below, the *key* of each element in the metadata standard is provided between brackets next to the corresponding element's label in the template.

TITLE STATEMENT

- **Primary ID** (*idno*) An identification number that is used to uniquely identify a video in a catalog. It will also help users of the data cite the video properly. The best option is to obtain a [Digital Object Identifier \(DOI\)](#) for the video, as it will ensure that the ID is unique globally. Alternatively, it can be an identifier constructed by an organization using a consistent scheme. Note that the schema allows you to provide more than one identifier for a video (see [Other identifiers](#) below). This element maps to the "identifier" element in the Dublin Core.
- **Title** (*title*) The title of the video. This element maps to the element *caption* in VideoObject.
- **Other identifiers** (*identifiers*) This element is used to enter video identifiers other than the **idno** element described above). It can for example be a Digital Object Identifier (DOI). Note that the identifier entered in **idno** can be repeated here, allowing to attach a "type" attribute to it.
 - **Type** (*type*) The type of unique identifier, e.g., "DOI".
 - **Identifier** (*value*) The identifier.
- **Alternate title** (*alt_title*) An alias for the video title. This element maps to the element *alternateName* in VideoObject.

AUTHORS AND CONTRIBUTORS

- **Creator** (*creator*) Organization or person who created/authored the video.

- **Production company** (*production_company*) The production company or studio responsible for the item. This element maps to the element `productionCompany` in VideoObject.
- **Recorded at** (*recorded_at*) This element maps to the element `recordedAt` in VideoObject schema. It identifies the event where the video was recorded (e.g., a conference, or a demonstration).
- **Publisher** (*Publisher*) The name of the publisher of the video.
- **Translators** (*Translators*) Organization or person who adapted the video to different languages. This element maps to the element `translator` in VideoObject.
 - **First name** (*first_name*) The first name of the translator.
 - **Initial** (*initial*) The initials of the translator.
 - **Last name** (*last_name*) The last name of the translator.
 - **Affiliation** (*affiliation*) The affiliation of the translator.
- **Sponsors** (*sponsors*) This element is used to list the funders/sponsors of the video. If different funding agencies financed different stages of the production process, use the "role" attribute to distinguish them.
 - **Name** (*name*) The name of the sponsor (person or organization)
 - **Abbreviation** (*abbr*) The abbreviation (acronym) of the sponsor.
 - **Grant** (*grant*) The grant (or contract) number.
 - **Role** (*role*) The specific role of the sponsor.
- **Other contributors** (*contributors*) Identifies the person(s) and/or organization(s) who contributed to the production of the video. The `role` attribute allows defining what the specific contribution of the identified person or organization was.
 - **Name** (*name*) The name of the contributor (person or organization).
 - **Affiliation** (*affiliation*) The affiliation of the contributor.
 - **Abbreviation** (*abbr*) The abbreviation for the institution which has been listed as the affiliation of the contributor.
 - **Role** (*role*) The specific role of the contributor. This could for example be "Cameraman", "Sound engineer", etc.
 - **URI** (*uri*) A URI (link to a website, or email address) for the contributor.
- **Credits** (*credit_text*) This element can be used to credit the person(s) and/or organization(s) associated with a published video. This element corresponds to the "creditText" element of VideoObject.
- **Contacts** (*contacts*) Users of the video may need further clarification and information. This section may include the name-affiliation-email-URI of one or multiple contact persons. This block of elements will identify contact persons who can be used as resource persons regarding problems or questions raised by the user community. The URI attribute should be used to indicate a URN or URL for the homepage of the contact individual. The email attribute is used to indicate an email address for the contact individual. It is recommended to avoid putting the actual name of individuals. The information provided here should be valid for the long term. It is therefore preferable to identify contact persons by a title. The same applies for the email field. Ideally, a "generic" email address should be provided. It is easy to configure a mail server in such a way that all messages sent to the generic email address would be automatically forwarded to some staff members.
 - **Name** (*name*) Name of a person or unit (such as a data help desk). It will usually be better to provide a title/function than the actual name of the person. Keep in mind that people do not stay forever in their position.
 - **Role** (*role*) The specific role of `name`, in regards to supporting users. This element is used when multiple names are provided, to help users identify the most appropriate person or unit to contact.
 - **Affiliation** (*affiliation*) Affiliation of the person/unit.
 - **Email** (*email*) E-mail address of the person.

- **Telephone** (telephone) A phone number that can be called to obtain information or provide feedback on the table. This should never be a personal phone number; a corporate number (typically of a data help desk) should be provided.
- **URL** (uri) A link to a website where contact information for **name** can be found.

DATES AND VERSION

- **Date created** (date_created) The date the video was created. It is recommended to enter the date in the ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY). The date the video is created refers to the date that the video was produced and considered ready for dissemination.
- **Date published** (date_published) The date the video was published. It is recommended to use the ISO 8601 format (YYYY-MM-DD or YYYY-MM or YYYY).
- **Version** (version) The version of the video refers to the published version of the video.
- **Status** (status) The status of the video in terms of its stage in a lifecycle. A controlled vocabulary should be used. Example terms include { **Incomplete**, **Draft**, **Published**, **Obsolete** }. Some organizations define a set of terms for the stages of their publication lifecycle. This element maps to the element *creativeWorkStatus* in VideoObject.

CONTENT

- **Description** (description) A brief description of the video, typically about a paragraph long (around 150 to 250 words). This element maps to the element *abstract* in VideoObject.
- **Genre** (genre) The genre of the video, broadcast channel or group. This is a VideoObject element. A controlled vocabulary can be used.
- **Audience** (audience) A brief description of the intended audience of the video, i.e. the group for whom it was created.
- **Keywords** (keywords) A list of keywords that provide information on the core content of the video. Keywords provide a convenient solution to improve the discoverability of the video, as it allows terms and phrases not found elsewhere in the video metadata to be indexed and to make the video discoverable by text-based search engines. A controlled vocabulary will preferably be used (although not required), such as the [UNESCO Thesaurus](#). The list can combine keywords from multiple controlled vocabularies, and user-defined keywords.
 - **Keyword** (name) The keyword itself.
 - **Vocabulary** (vocabulary) The controlled vocabulary (including version number or date) from which the keyword is extracted, if any.
 - **URL** (uri) The URL of the controlled vocabulary from which the keyword is extracted, if any.
- **Topics** (topics) Information on the topics covered in the video. A controlled vocabulary will preferably be used, for example the [CESSDA Topics classification](#), a typology of topics available in 11 languages; or the [Journal of Economic Literature \(JEL\) Classification System](#), or the [World Bank topics classification](#). Note that you may use more than one controlled vocabulary. This element is a block of five fields:
 - **ID** (id) The identifier of the topic, taken from a controlled vocabulary.
 - **Topic** (name) The name (label) of the topic, preferably taken from a controlled vocabulary.
 - **Parent ID** (parent_id) The parent identifier of the topic (identifier of the item one level up in the hierarchy), if a hierarchical controlled vocabulary is used.
 - **Vocabulary** (vocabulary) The name (including version number) of the controlled vocabulary used, if any.
 - **URL** (uri) The URL to the controlled vocabulary used, if any.
- **Persons** (persons) A list of persons who appear in the video.
 - **Name** (name) The name of the person.

- **Role** (*role*) The role of the person mentioned in `name`.
- **Main entity** (*main_entity*) Indicates the primary entity described in the video. This element maps to the element `mainEntity` in VideoObject.
- **Transcript** (*transcript*) The transcript of the video content, provided as a text. Note that if the text is very long, an alternative is to save it in a separate text file and to make it available in a data catalog as an external resource.
 - **Language code** (*language_code*) The code of the language of the transcript, preferably the ISO code.
 - **Language name** (*language_name*) The name of the language of the transcript.
 - **Text** (*text*) The transcript itself. Adding the transcript in the metadata will make the video much more discoverable, as the content of the transcription can be indexed in catalogs.
- **Album** (*album*) When a video is published in a catalog containing many other videos, it may be desirable to organize them by album. Albums are collections of videos organized by theme, period, location, or other criteria. One video can belong to more than one album. Albums are "virtual collections".
 - **Name** (*name*) The name (label) of the album.
 - **Description** (*description*) A brief description of the album.
 - **Owner** (*owner*) The owner of the album.
 - **URL** (*uri*) A link (URL) to the album.
- **Languages** (*language*) Most videos will only be provided in one language. This is however a repeatable field, to allow for more than one language to be listed. For the language code, ISO codes will preferably be used. The language refers to the language in which the video is published. This is a block of two elements (at least one must be provided for each language):
 - **Name** (*name*) The name of the language.
 - **Code** (*code*) The code of the language. The use of [ISO 639-2](#) (the alpha-3 code in Codes for the representation of names of languages) is recommended. Numeric codes must be entered as strings.
- **Is based on** (*is_based_on*) A resource from which this video is derived or from which it is a modification or adaption. This element maps to the element `isBasedOn` in VideoObject.
- **Is part of** (*is_part_of*) Indicates another video that this video is part of. This element maps to the element `isPartOf` in VideoObject.
- **Relations** (*relations*) Defines, as a free text field, the relation between the video being documented and other resources. This is a Dublin Core element.

GEOGRAPHIC AND TIME COVERAGE

- **Country** (*country*) The list of countries (or regions) covered by the video, if applicable. This refers to the content of the video, not to the country where the video was released. This is a repeatable block of two elements:
 - **Name** (*name*) The country/region name. Note that many organizations have their own policies on the naming of countries/regions/economies/territories, which data curators will have to comply with.
 - **Code** (*code*) The country/region code (entered as a string, even for numeric codes). It is recommended to use a standard list of countries and regions, such as the ISO country list ([ISO 3166](#)).
- **Spatial coverage** (*spatial_coverage*) Indicates the place(s) which are depicted or described in the video. This element maps to the element `contentLocation` in VideoObject. This element complements the `ref_country` element. It can be used to qualify the geographic coverage of the video, in the form of a free text.
- **Bounding box** (*bbox*) This element is used to define one or multiple bounding box(es), which are the (rectangular) fundamental geometric description of the geographic coverage of the video. A bounding box is defined by west and east longitudes and north and south latitudes, and includes the largest geographic extent of the video's geographic

coverage. The bounding box provides the geographic coordinates of the top left (north/west) and bottom-right (south/east) corners of a rectangular area. This element can be used in catalogs as the first pass of a coordinate-based search.

- **West** (west) West longitude of the box
- **East** (east) East longitude of the box
- **South** (south) South latitude of the box
- **North** (north) North latitude of the box
- **Reference time** (content_reference_time) The specific time described by the video, for works that emphasize a particular moment within an event. This element maps to the element `contentReferenceTime` in VideoObject.
- **Temporal coverage** (temporal_coverage) Indicates the period that the video applies to, i.e. that it describes, either as a DateTime or as a textual string indicating a time period in ISO 8601 time interval format. This element maps to the element `temporalCoverage` in VideoObject.

ACCESS AND RIGHTS

- **Videoproducer** (video_provider) The person or organization who provides the video. This element maps to the element `provider` in VideoObject.
- **Video URL** (video_url) URL of the video. This element maps to the element `url` in VideoObject.
- **Embed URL** (embed_url) A URL pointing to a player for a specific video. For example, "<https://www.youtube.com/embed/7Aif1xjstws>". To be embedded, a video must be hosted on a video sharing platform like Youtube (www.youtube.com). To obtain the "embed link" from youtube, click on the "Share" button, then "Embed". In the result box, select the content of the element `src =`.

```
{width=100%}
```

- **Repository** (repository) The name of the repository (organization)
- **Rights** (rights) A textual description of the rights associated to the video. If a copyright is available, the three following elements will be used instead of this element.
- **Copyright holder** (copyright_holder) The party holding the legal copyright to the video. This element corresponds to the "copyrightHolder" element of VideoObject.
- **Copyright notice** (copyright_notice) Text of a notice appropriate for describing the copyright aspects of the video, ideally indicating the owner of the copyright. This element corresponds to the "copyrightNotice" element of VideoObject.
- **Copyright year** (copyright_year) The year during which the claimed copyright for the video was first asserted. This element corresponds to the "copyrightYear" element of VideoObject.
- **Citation** (citation) This element provides a required or recommended citation of the audio file.

TECHNICAL INFORMATION

- **Media** (media) A description of the media on which the recording is stored (other than the online file format); e.g., "CD-ROM".
- **Duration** (duration) The duration of the item (movie, audio recording, event, etc.) in ISO 8601 format. ISO 8601 durations are expressed using the following format, where (n) is replaced by the value for each of the date and time elements that follows the (n). For example: (3)H means 3 hours.

Duration in ISO 8601 format is in the form: **P(n)Y(n)M(n)DT(n)H(n)M(n)S** where:

- P is the **Period designator** and is always placed at the beginning of the duration
 - (n)Y represents the number of years
 - (n)M represents the number of months
 - (n)W represents the number of weeks
 - (n)D represents the number of days
- T is the **Time designator** and always precedes the time components
 - (n)H represents the number of hours
 - (n)M represents the number of minutes
 - (n)S represents the number of seconds

For example, **P1Y2M20DT3H30M8S** represents a duration of one year, two months, twenty days, three hours, thirty minutes, and eight seconds.

Date and time elements including their designator may be omitted if their value is zero, and lower-order elements may also be omitted for reduced precision. For example, "P23DT23H" and "P4Y" are both acceptable duration representations.

As M can represent both Month and Minutes, the time designator T is used. For example, "P1M" is a one-month duration and "PT1M" is a one-minute duration.

This information on the ISO 8601 was adapted from [wikipedia](#) where more detailed information can be found.

- **Encoding format** (*encoding_format*) The video file format, typically expressed using a MIME format. This element corresponds to the "encodingFormat" element of VideoObject and maps to the element *format* of the Dublin Core.

DataCite

See section **Documenting - General instructions**.

Tags

See section **Documenting - General instructions**.

Provenance

The **Provenance** container is used to document how and when the video was acquired, if the video and its metadata were extracted from an external catalog. It is used to ensure traceability. See section **Documenting - General instructions**.

External resources

External resources are all materials (and links) that relate to the video, if any. These materials and links are added in the External resources container. Select *External resources* in the navigation tree, then on **CREATE RESOURCE**. Enter the relevant information on the resource (at least a title), then provide either a filename (the file will then be uploaded on the server that hosts the Metadata Editor) or a URL to the resource.

External resources that have already been created for another project can also be imported. To do that, they must first be exported as JSON or RDF from the other project. Then click on **IMPORT** in the *External resources* page, and select the file.

Documenting research projects and scripts

Rationale

Documenting, cataloguing and disseminating **data** has the potential to increase the volume and diversity of data analysis. There is also much value in documenting, cataloguing and disseminating **data processing and analysis scripts**.

Technological solutions such as GitHub, [Jupyter Notebooks or Jupiter Lab](#) facilitate the preservation and sharing of code, and enable collaborative work around data analysis. Coding style guides like the [Google style guides](#) and the [Guide to Reproducible Code in Ecology and Evolution](#) by the British Ecological Society, contribute to foster the usability, adaptability, and reproducibility of code. But these tools and guidelines do not fully address the issue of cataloguing and discoverability of the data processing and analysis programs and scripts. We propose --as a complement to collaboration tools and style guides-- a metadata schema to document data analysis projects and scripts. The production of structured metadata will contribute not only to discoverability, but also to the reproducibility, replicability, and auditability of data analytics.

There are multiple reasons to make reproducibility, replicability, and auditability of data analytics a component of a data dissemination system. This will:

- Improve the **quality of research and analysis**. Public scrutiny enables contestability and independent quality control of the output of research and analysis; these are strong incentives for additional rigor in data analysis.
- Allow the **re-purposing or expansion of analysis** by the research community, thereby increasing the relevance, utility and value of both the data and of the analytical work.
- Strengthen the **reputation and credibility** of the analysis.
- Provide students and peers with rich **training materials**.
- In some cases, satisfy a **requirement** imposed by peer reviewed journals or financial sponsors of research activities. For example, the [Data and Policy Code of the American Economic Association](#) (accessed on June 29, 2020), states that *It is the policy of the American Economic Association to publish papers only if the data and code used in the analysis are clearly and precisely documented, and access to the data and code is clearly and precisely documented and is non-exclusive to the authors. Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to acceptance, information about the data, programs, and other details of the computations sufficient to permit replication, as well as information about access to data and programs.*
- Contribute to **assuring the fairness of policy advice and interventions** resulting from data analysis. Data analysis may be used to identify or target the beneficiaries of policies and programs, or may contribute otherwise to the design and implementation of development policies and projects. By doing so, they also contribute to identifying populations to be excluded from these interventions. Errors and biases may be introduced in analysis by accidental or intentional human errors, by the algorithms themselves, or they can result from flaws in the data. The analysis that informs such projects and policies should therefore be made auditable and contestable, i.e. documented and published.

The metadata standard

To make data processing and analysis scripts more discoverable and usable, we propose a metadata schema inspired by the schemas available to document datasets. The proposed schema contains two main blocks of metadata elements: the *document description* intended to document the metadata themselves (the term *document* refers to the file that will

contain the metadata), and the *project description* used to document the research or analytical work and the related scripts. We also include in the schema the `tags`, `provenance`, and `additional` elements common to all schemas.

Preparing for the documentation

Before you start documenting the project, it is highly recommended to organize all relevant resources in an optimal manner. This includes following best practice in the writing of scripts (and ideally adopting a relevant style guide), of naming the files containing your programs and scripts, or organizing a GitHub repository if you publish content on GitHub, and more.

Resources related to best practice in code writing and reproducibility include (among others):

- **World Bank Reproducible Research Catalog:** An online catalog of reproducible research by the World Bank, which makes use of the metadata standard documented in the chapter (<https://reproducibility.worldbank.org/>)
- **Reproducible Research:** A wiki page by the World Bank DIME department (https://dimewiki.worldbank.org/Reproducible_Research)
- **World Bank Reproducible Research Repository Resources:** (<https://worldbank.github.io/wb-reproducible-research-repository/>)
- **PEP 8 – Style Guide for Python Code** (for Python users) (<https://peps.python.org/pep-0008/>)
- **Tidyverse style guide** by the tidyverse team (for R users) (<https://style.tidyverse.org/>)

Documenting the research project

Create a new project

The first step in documenting a research project and its scripts is to create a new project. You do that by clicking on `CREATE NEW PROJECT` in the *My projects* page. Select *Script* as data type. This will open a new, untitled *Project Home* page.

The screenshot shows the Metadata Editor application. At the top, there's a dark header bar with the 'Metadata Editor' logo, 'ABOUT', and a user profile for 'JOHN DOE'. Below the header is a navigation bar with three tabs: 'PROJECTS' (which is active and highlighted in blue), 'COLLECTIONS', and 'TEMPLATES'. The main content area is titled 'My projects' and displays a search bar with placeholder text 'Search...'. Below the search bar, it says 'Showing 1 - 7 of 7 projects'. On the far right of the main area, there are buttons for 'CREATE NEW PROJECT' (blue), 'IMPORT' (light blue), and navigation arrows. To the left of the main content, there's a sidebar with a 'Type' dropdown menu containing the following options:

- Microdata
- Timeseries
- Timeseries (database)
- Script
- Geospatial
- Document
- Table
- Image
- Video

Create new project

-  Microdata
-  Timeseries
-  Timeseries database
-  Document
-  Table
-  Image
-  Script
-  Video
-  Geospatial

CLOSE

In that page, edit the thumbnail and replace it with an image of your choice (recommended, not required).

Then select (in the *Template* frame) the project template you want to use to document the project. A default template is proposed; no action is needed if you want to use the default template. Otherwise, switch to another template by clicking on the template name in the **Templates** frame. Note that you can at any time change the template used for the documentation of a project. The selected template will determine what you see in the navigation tree and in the metadata entry pages, but switching from one template to another will not impact the metadata that has already been entered; no information will be deleted from the metadata.

The screenshot shows the Metadata Editor interface. At the top, there are three status indicators: 'Required' (green checkmark), 'Recommended' (yellow circle), and 'Empty' (grey circle). Below them is a search bar with placeholder text 'Search...'. On the left, a navigation tree includes 'Home', 'Metadata information' (selected), 'Project description', 'Tags', and 'External resources'. The main content area is titled 'Untitled'. It features a thumbnail image of a person working on a laptop. To the right of the image are several metadata fields: 'Project owner: John Doe', 'Created on: 2025-02-20 16:22:55'; 'Last changed by: John Doe', 'Changed on: 2025-02-20 16:22:55'; and 'Project IDNO: 05644962-11c0-40ef-8524-1ee045e91960'. Below this is a 'Template' section showing 'Project template: IHSN SCRIPT 1.0 TEMPLATE V01 EN - 1.0' and an 'Administrative metadata templates' dropdown. A 'Project validation' section contains 'Schema validation' (No validation errors found) and 'Template validation' (No validation errors found). To the right are sections for 'Collaborators' (None), 'Collections' (None), and 'Data and Documentation' (Documentation tab selected, showing Disk usage: 21.17 KB). A 'FILES' tab is also present in the documentation section.

Enter metadata

Metadata information

The Metadata information section in the navigation tree (in the Project page) contains elements intended to document the metadata being generated, i.e., metadata about the metadata. All content in this section is optional; it is however recommended practice to document the metadata as precisely as possible. This information will not be useful to data users, but it will be to catalog administrators. When metadata is shared across catalogs, the information entered in the [Information on metadata](#) provides transparency and clarity on the origin of the metadata.

Information on metadata

Document title ⓘ

Document ID ⓘ

Metadata producers ⓘ

| Name | Abbreviation | Affiliation | Role |
|---------|--------------|-------------|---|
| [Empty] | [Empty] | [Empty] | [Empty] X |

+ ADD ROW

Production date ⓘ

Version ⓘ

Project description

Title statement

Primary ID ⓘ

Other identifiers ⓘ

| Type | Identifier |
|---------|---|
| [Empty] | [Empty] X |

+ ADD ROW

Title ⓘ

Subtitle ⓘ

Alternate title ⓘ

Translated title ⓘ

Project website ⓘ

We provide below a description of the metadata elements contained in the [Project description](#) section of the default metadata template provided in the Metadata Editor. Other templates may show a different selection, different labels, or present the elements in a different sequence. When you document a dataset, it is not expected that all these elements will be filled. Fill all required elements, all recommended elements when content can be made available, and fill as many as the other elements (the required and recommended elements will be those marked as *required* or *recommended* in the metadata standard or template).

In the list of metadata elements below, the key of each element in the metadata standard is provided between brackets next to the corresponding element's label in the template.

TITLE STATEMENT

- **Primary ID** (*idno*) A unique identifier to the project. Define and use a consistent scheme to use. Avoid including spaces in the ID. The ID number of a research project is a unique number that is used to identify a particular project. This ID number is a vital reference. A research project can be the formal cause of a survey, scripts, tables and knowledge products. Do not include spaces in the *idno* element. Use a system that guarantees uniqueness of the ID (DOI, own reference number).
- **Other identifiers** (*identifiers*) This repeatable element is used to enter identifiers (IDs) other than the *idno* entered in the *title_statement*. It can for example be a Digital Object Identifier (DOI). Note that the identifier entered in *idno* can (and in some cases should) be repeated here. The element *idno* does not provide a *type* parameter; repeating it in this section makes it possible to add that information.
 - **Type** (*type*) The type of unique ID, e.g. "DOI".
 - **Identifier** (*identifier*) The identifier itself.
- **Title** (*title*) The title is the official name of the project as it may be stated in reports, papers or other documents. The title will in most cases be identical to the Document Title (see above). The title may correspond to the title of an academic paper, of a project impact evaluation, etc. Pay attention to capitalization in the title.
- **Subtitle** (*sub_title*) Subtitle is optional and rarely used. A short subtitle for the project. Often the sub title is used to qualify the title or rephrase the title.
- **Alternate title** (*alternate_title*) An alternate title of the project. This would be any alternate title that would help discover the research project. In countries with more than one official language, a translation of the title may be provided. Likewise, the translated title may simply be a translation into English from a country's own language.
- **Translated title** (*translated_title*) A translated version of the title (this will be used for example when a catalog documents all entries in English, but wants to preserve the title of a project in its original language when the original language is not English).
- **Project website** (*project_website*) URL of the project website.

AUTHORS AND CONTRIBUTORS

- **Authoring entity** (*authoring_entity*) This section will identify the person(s) and/or organization(s) in charge of the intellectual content of the research project, and specify their respective role.
 - **Name** (*name*) Name of the person or organization responsible for the research project.
 - **Role** (*role*) Specific role of the person or organization mentioned in *name*.
 - **Affiliation** (*affiliation*) Agency or organization affiliation of the author/primary investigator mentioned in *name*.
 - **Abbreviation** (*abbreviation*) Abbreviation used to identify the agency stated under *affiliation*.
 - **Email** (*email*) Depending on the agency policies, a researcher may provide a personal email or an agency email to field inquiries related to the project.
 - **Author ID** (*author_id*) A block of two elements used to provide unique identifiers of the authors, as provided by different registers of researchers. For example, this can be an ORCID number (ORCID is a non-profit organization supported by a global community of member organizations, including research institutions, publishers, sponsors, professional associations, service providers, and other stakeholders in the research ecosystem.)
 - **Type** (*type*) The type of ID; for example, "ORCID".
 - **ID** (*id*) A unique identification number/code for the authoring entity, entered as a string variable.

- **Contributors** (*contributors*) This section is provided to record other contributors to the research project and provide recognition for the roles they provided.
 - **Name** (*name*) Name of the person, corporate body, or agency contributing to the intellectual content of the project (other than the PI). If a person, invert first and last name and use commas.
 - **Role** (*role*) Title of the person (if any) responsible for the work's substantive and intellectual content.
 - **Affiliation** (*affiliation*) Agency or organization affiliation of the contributor.
 - **Abbreviation** (*abbreviation*) Abbreviation used to identify the agency stated under **affiliation**.
 - **Email** (*email*) Depending on the agency policies, a researcher may provide a personal email or an agency email to field inquiries related to the project.
 - **URL** (*url*) The URL that provides information on the contributor or its affiliate
- **Sponsors** (*sponsors*) The source(s) of funds for production of the work. If different funding agencies sponsored different stages of the production process, use the 'role' attribute to distinguish them.
 - **Name** (*name*) Name of the funding agency/sponsor.
 - **Abbreviation** (*abbreviation*) Abbreviation of the funding/sponsoring agency.
 - **Role** (*role*) Specific role of the funding/sponsoring agency.
 - **Grant No** (*grant_no*) Grant or award number.
- **Curators** (*curators*) A list of persons and/or organizations in charge of curating the resources associated with the project.
 - **Name** (*name*) The name of the person or organization.
 - **Role** (*role*) The specific role of the person or organization in the curation of the project resources.
 - **Affiliation** (*affiliation*) The affiliation of the person or organization.
 - **Abbreviation** (*abbreviation*) An acronym of the organization, if an organization was entered in **name**.
 - **Email** (*email*) The email address of the person or organization. The use of personal email addresses must be avoided.
 - **URL** (*url*) A link to the website of the person or organization.
- **Acknowledgments** (*acknowledgments*) This repeatable block of elements is used to provide an itemized list of persons and organizations whose contribution to the project must be acknowledged. Note that specific metadata elements are available for listing financial sponsors and main contributors to the study. An alternative to this field is the **acknowledgment_statement** field (see below) which can be used to provide the acknowledgment in the form of an unstructured text.
 - **Name** (*name*) The name of the person or agency being recognized for supporting the project.
 - **Affiliation** (*affiliation*) The affiliation of the person or agency being acknowledged.
 - **Role** (*role*) A brief description of the role of the person or agency that is being recognized or acknowledged for supporting the project.
- **Acknowledgement statement** (*acknowledgement_statement*) This field is used to provide acknowledgments in the form of an unstructured text. An alternative to this field is the *acknowledgments* field which provides a solution to itemize the acknowledgments.

REPRODUCIBILITY STATUS

- **Status** (*type*) Information on the reproducibility status of the project, using a controlled vocabulary
- **Reproducibility note** (*note*) Any additional information on the status of reproducibility.

VERSION STATEMENT

- **Project completion date** (*production_date*) The date in ISO 8601 format (YYYY-MM-DD) the project was completed (this refers to the version that is being documented and released.)
- **Version** (*version*) A label describing the version. For example, "Version 1.2" [*String*]
- **Version date** (*version_date*) Date (in ISO 8601 format, YYYY-MM-DD) the version was released [*String*]
- **Responsibility** (*version_resp*) Person(s) or organization(s) responsible for this version. [*String*]
- **Note on version** (*version_notes*) Additional information on the version if any; it is good practice to describe what distinguishes this version from the previous one(s). The version must be entered as a string, even when composed only of numbers.

SCOPE AND COVERAGE

- **Abstract** (*abstract*) The abstract should provide a clear summary of the purposes, objectives and content of the project. An abstract can make reference to the various outputs associated with the research project. Example: "Food price inflation is an important metric to inform economic policy but traditional sources of consumer prices are often produced with delay during crises and only at an aggregate level. This may poorly reflect the actual price trends in rural or poverty-stricken areas, where large populations reside in fragile situations. This data set includes food price estimates and is intended to help gain insight in price developments beyond what can be formally measured by traditional methods. The estimates are generated using a machine-learning approach that imputes ongoing subnational price surveys, often with accuracy similar to direct measurement of prices. The data set provides new opportunities to investigate local price dynamics in areas where populations are sensitive to localized price shocks and where traditional data are not available.",
- **Geographic areas** (*geographic_units*) The geographic areas covered by the project. When the project relates to one or more countries, or part of one or more countries, it is important to provide the country name. This means that for a project related to a specific province or town of a country, the country name will be entered in addition to the province or town (as separate entries in this repeatable block of elements). Note that the area does not have to be an administrative area; it can for example be an ocean.
 - **Name** (*name*) The name of the geographic area.
 - **Code** (*code*) The code of the geographic area. For countries, it is recommended to use the [ISO 3166](#) country codes and names.
 - **Type** (*type*) The type of geographic area.
- **Keywords** (*keywords*) A list of keywords that provide information on the core scope and objectives of the research project. Keywords provide a convenient solution to improve the discoverability of the research, as it allows terms and phrases not found elsewhere in the metadata to be indexed and to make a project discoverable by text-based search engines. A controlled vocabulary will preferably be used (although not required), such as the [UNESCO Thesaurus](#). The list provided here can combine keywords from multiple controlled vocabularies, and user-defined keywords.
 - **Keyword** (*name*) The keyword itself.
 - **Vocabulary** (*vocabulary*) The controlled vocabulary (including version number or date) from which the keyword is extracted, if any.
 - **URL** (*uri*) The URL of the controlled vocabulary from which the keyword is extracted, if any.
- **Themes** (*themes*) A list of themes covered by the research project. A controlled vocabulary will preferably be used. Note that **themes** will rarely be used as the elements **topics** and **disciplines** are more appropriate for most uses. This is a block of five fields:
 - **ID** (*id*) The ID of the theme, taken from a controlled vocabulary.
 - **Name** (*name*) The name (label) of the theme, preferably taken from a controlled vocabulary.
 - **Parent ID** (*parent_id*) The parent ID of the theme (ID of the item one level up in the hierarchy), if a hierarchical controlled vocabulary is used.
 - **Vocabulary** (*vocabulary*) The name (including version number) of the controlled vocabulary used, if any.

- **URL** (*uri*) The URL to the controlled vocabulary used, if any.
- **Topics** (*topics*) Information on the topics covered in the research project. A controlled vocabulary will preferably be used, for example the [CESSDA Topics classification](#), a typology of topics available in 11 languages; or the [Journal of Economic Literature \(JEL\) Classification System](#), or the [World Bank topics classification](#). Note that you may use more than one controlled vocabulary. This element is a block of five fields:
 - **ID** (*id*) The identifier of the topic, taken from a controlled vocabulary.
 - **Name** (*name*) The name (label) of the topic, preferably taken from a controlled vocabulary.
 - **Parent ID** (*parent_id*) The parent identifier of the topic (identifier of the item one level up in the hierarchy), if a hierarchical controlled vocabulary is used.
 - **Vocabulary** (*vocabulary*) The name (including version number) of the controlled vocabulary used, if any.
 - **URL** (*uri*) The URL to the controlled vocabulary used, if any.
- **Disciplines** (*disciplines*) Information on the academic disciplines related to the content of the research project. A controlled vocabulary will preferably be used, for example the one provided by the list of academic fields in [Wikipedia](#). This is a block of five elements:
 - **ID** (*id*) The identifier of the discipline, taken from a controlled vocabulary.
 - **Name** (*name*) The name (label) of the discipline, preferably taken from a controlled vocabulary.
 - **Parent ID** (*parent_id*) The parent identifier of the discipline (identifier of the item one level up in the hierarchy), if a hierarchical controlled vocabulary is used.
 - **Vocabulary** (*vocabulary*) The name (including version number) of the controlled vocabulary used, if any.
 - **URL** (*uri*) The URL to the controlled vocabulary used, if any.
- **Related projects** (*related_projects*) The objective of this block is to provide links (URLs) to other, related projects which can be documented and disseminated in the same catalog or any other location on the internet.
 - **Name** (*name*) The name (title) of the related project.
 - **URL** (*uri*) A link (URL) to the related project web page.
 - **Note** (*note*) A brief description or other relevant information on the related project.

PROCESSES AND OUTPUT

- **Process** (*process*) This element is used to document the life cycle of the research project, from its design and inception to its conclusion. This can include phases of fundraising, IRB, concept note review, data acquisition, analysis, publishing of a working paper, peer review, publishing in journal, presentation to conferences, publishing, evaluation, reporting to sponsors, etc. It is recommended to provide these steps in a chronological order.
 - **Name** (*name*) This is a header for the phase of the process.
 - **Start date** (*date_start*) Date the phase started (preferably in ISO 8601 format, YYYY-MM-DD)
 - **End date** (*date_end*) Date the phase ended (preferably in ISO 8601 format, YYYY-MM-DD)
 - **Description** (*description*) A brief description of the phase.
- **Institutional review board** (*review_board*) Information on whether and when the project was submitted, reviewed, and approved by an institutional review board (or independent ethics committee, ethical review board (ERB), research ethics board, or equivalent).
- **Approval process** (*approval_process*) The *approval_process* is a group of six elements used to describe the formal approval process(es) (if any) that the project had to go through. This may for example include an approval by an Ethics Board to collect new data, followed by an internal review process to endorse the results.
 - **Phase name** (*approval_phase*) A label that describes the approval phase.

- **Authority** (*approval_authority*) Identification of the person(s) or organization(s) whose approval was required or sought.
- **Submission date** (*submission_date*) The date, entered in ISO 8601 format (YYYY-MM-DD), when the project (or a component of it) was submitted for approval.
- **Reviewer** (*reviewer*) Identification of the reviewer(s).
- **Review status** (*review_status*) Status of approval.
- **Approval date** (*approval_date*) Date the approval was formally received, preferably entered in ISO 8601 format (YYYY-MM-DD).
- **Reviews and comments** (*reviews_comments*) Many research projects will be subject to a review process, which may happen at different stages of the project implementation (from design to review of the final output). This block is intended to document the comments received by reviewers during this process. It is a repeatable block of metadata elements, which can be used to document comments with a fine granularity.
 - **Date** (*comment_date*) The date the comment was provided, in ISO 8601 format (YYYY-MM-DD or YYYY-MM).
 - **Name** (*comment_by*) The name of the person or organization that provided the comment.
 - **Comment** (*comment_description*) The comment itself, in its original formulation or in a summary version.
 - **Response** (*comment_response*) The response provided by the research team/person to the comment, in its original formulation or in a summary version.
- **Output** (*output*) This element will describe and reference all substantial/intended products of the research project, which may include publications, reports, websites, datasets, interactive applications, presentations, visualizations, and others. An output may also be referred to as a "deliverable". The **output** is a repeatable block of seven elements, used to document all output of the research project:
 - **Type** (*type*) Type of output. The type of output relates to the media which is used to convey or communicate the intended results, findings or conclusions of the research project. This field may be controlled by a controlled vocabulary. The kind of content could be "Working paper", "Database", etc.
 - **Title** (*title*) Formal title of the output. Depending upon the kind of output, the title will vary in formality.
 - **Authors** (*authors*) Authors of the output; if multiple, they will be listed in one same text field.
 - **Description** (*description*) Brief description of the output (NOT an abstract)
 - **Abstract** (*abstract*) If the output consists of a document, the abstract will be entered here.
 - **URL** (*uri*) A link where the output or information on the output can be found.
 - **DOI** (*doi*) Digital Object Identifier (DOI) of the output, if available.
- **Language** (*language*) A block of two elements describing the language(s) of the project. At least one of the two elements must be provided for each listed language. The use of [ISO 639-2](#) (the alpha-3 code in Codes for the representation of names of languages) is recommended.
 - **Name** (*name*) The name of the language.
 - **Code** (*code*) The code of the language. Numeric codes must be entered as strings.
- **Errata** (*errata*) This field is used to list and describe errata.
 - **Date** (*date*) Date (in ISO 8601 format, YYYY-MM-DD) the erratum was released.
 - **Description** (*description*) Description of the error(s) and measures taken to address it/them.

DATA

- **Data statement** (*data_statement*) An overall statement on the data used in the project. A separate field is provided to list and document the origin and key characteristics of the datasets.

- **Datasets** (*datasets*) This field is used to provide an itemized list of datasets used in the project. The data are not documented here (specific metadata are available for documenting data of different types, like the DDI for microdata, the ISO 19139 for geographic datasets, etc.)
 - **Name** (*name*) The dataset name (title)
 - **Dataset ID** (*idno*) The unique identifier of the dataset
 - **Note** (*note*) A brief description of the dataset.
 - **Access policy** (*access_type*) The access policy applied to the dataset.
 - **License** (*license*) The access license that applies to the dataset.
 - **License URL** (*license_uri*) The URL of a web page where more information on the license can be obtained.
 - **Data URL** (*uri*) The URI where the dataset (or a detailed description of it) can be obtained.

METHODS, SOFTWARE AND SCRIPTS

- **Methods** (*methods*) A list of analytic, statistical, econometric, machine learning methods used in the project. The objective is to allow users to find projects based on a search on methods applied, e.g. answer a query like "*poverty prediction using random forest*".
 - **Name** (*name*) A short name for the method being described.
 - **Note** (*note*) Any additional information on the method.
- **Software** (*software*) This field is used to list the software and the specialized packages and libraries/packages that were used to implement the project and that are required to reproduce the scripts. The libraries that are loaded by the scripts (e.g., by the R *require* or *library* command) are included (not all their own dependencies, which will be assumed to be installed automatically).
 - **Name** (*name*) The name of the software.
 - **Version** (*version*) The version of the software.
 - **Library** (*library*) A list of libraries/packages required to run the scripts. Note that the specific version of each package is not documented here; it is expected to be found in the script or in the reproduction instructions.
- **Scripts** (*scripts*) This field is used to describe the scripts written by the project authors. All scripts are expected to have been written using software listed in the field *software*.
 - **File_name** (*file_name*) Name of the script file (for R users, this will typically include files with extension [.R], for Stata users it will be files with extension [.do], for Python users ...). But this can also include other files related and required to run the scripts (for example lookup CSV files, etc.) This does not include the data files, which are described in a specific field.
 - **Zip package** (*zip_package*) If the script files have been saved as or in a compressed file (zip, rar, or equivalent), we provide here the name of the zip file containing the script.
 - **Title** (*title*) A title (label) given to the script file
 - **Authors** (*authors*) This is a repeatable block that allows entering a list of authors and co-authors of a script
 - **Name** (*name*) Name of the author (person or organization) of the script
 - **Affiliation** (*affiliation*) The affiliation of the author.
 - **Role** (*role*) Specific role of the person or organization in the production of the script.
 - **Date** (*date*) Date the script was produced, in ISO 8601 format (YYYY-MM-DD)
 - **Format** (*format*) File format
 - **Software** (*software*) Software used to run the script

- **Description** (*description*) Brief description of the script
- **Methods** (*methods*) Statistical/analytic methods included in the script
- **Dependencies** (*dependencies*) Any dependencies (packages/libraries) that the script relies on. This field is not needed if dependencies were described in the **library** element.
- **Instructions** (*instructions*) Instructions for running the script. Information on the sequence in which the scripts must be run is critical.
- **Repository** (*source_code_repo*) Repository (e.g. GitHub repo) where the script has been published.
- **Notes** (*notes*) Any additional information on the script.
- **License** (*license*) License, if any, under which the script is published.
 - **Name** (*name*) Name (label) of the license
 - **URL** (*uri*) License URI
- **Repository** (*repository_uri*) In the process of producing the outputs of the research project, a researcher may want to share their source code for transparency and replicability. This repository provides information for finding the repository where the source code is kept.
 - **Name** (*name*) Name of the repository where code is hosted.
 - **Type** (*type*) Repository type e.g. GitHub, Bitbucket, etc.
 - **URL** (*uri*) URI of the project source code/script repository
- **Reproducibility**
 - **Technology environment** (*technology_environment*) This field is used to provide a description (as detailed as possible) of the computational environment under which the scripts were implemented and are expected to be reproducible. A substantial challenge in reproducing analyses is installing and configuring the web of dependencies of specific versions of various analytical tools. Virtual machines (a computer inside a computer) enable you to efficiently share your entire computational environment with all the dependencies intact.
[\(https://ropensci.github.io/reproducibility-guide/sections/introduction/\)](https://ropensci.github.io/reproducibility-guide/sections/introduction/)
 - **Technology requirements** (*technology_requirements*) Software/hardware or other technology requirements needed to run the scripts and replicate the outputs
 - **Reproduction instructions** (*reproduction_instructions*) Instructions to secondary analysts who may want to reproduce the scripts.

ACCESS AND RIGHTS

- **Disclaimer** (*disclaimer*) Disclaimers limit the responsibility or liability of the publishing organization or researchers associated with the research project. Disclaimers assure that any research in the public domain produced by an organization has limited repercussions to the publishing organization. A disclaimer is intended to prevent liability from any effects occurring as a result of the acts or omissions in the research.
- **Confidentiality** (*confidentiality*) A confidentiality statement binds the publisher to ethical considerations regarding the subjects of the research. In most cases, the individual identity of an individual that is the subject of research can not be released and special effort is required to assure the preservation of privacy.
- **Citation requirement** (*citation_requirement*) The citation requirement is specific to the output and is a preferred shorthand or means to refer to the publication or published good.
- **License** (*license*) Information on the license(s) attached to the research project resources, which defines their terms of use.
 - **Name** (*name*) The name of the license.

- **URL** (*uri*) The URL of the license, where detailed information on the license can be obtained.
- **Copyright** (*copyright*) Information on the copyright, if any, that applies to the research project metadata.

CONTACTS

- **Contacts** (*contacts*) The contacts element provides the public interface for questions associated with the research project. There could be various contacts provided depending upon the organization. It is important to assure that the proper contacts are provided to channel public inquiry.
 - **Name** (*name*) The name of the contact person that should be contacted depending on the role defined below.
 - **Role** (*role*) Role of the contact person. A research project may have contact persons depending on the output or some of the technical input. Some complex projects may have various data collection processes that have different processing channels and contacts. This section should provide for a key primary public interface that can refer the public inquiry or provide a collection of entry points.
 - **Affiliation** (*affiliation*) The organization or affiliation of the contact person. This is usually the organization that the contact person represents.
 - **Email** (*email*) Email address of the responsible person, institution, or division in charge of the research project or output.
 - **Phone** (*telephone*) Phone number of the responsible institution or division of the research project or output.
 - **URL** (*uri*) The URI of the agency or organization of the contact organization. This may be the same as the web page of the project or may be a permanent contact name at an institutional level and not project related. Eventually a project web site may be removed but there may still be need to have a contact. In this case, it is recommended to have a contact that is permanent.

DataCite

See section [Documenting - General instructions](#).

Tags

See section [Documenting - General instructions](#).

Provenance

The **Provenance** container is used to document how and when the project was acquired, if the project was extracted from an external catalog. It is used to ensure traceability. See section [Documenting - General instructions](#).

External resources

External resources are all materials (and links) that relate to the indicator. This will include the programs and scripts, documents, and other digital resources and links. These materials and links are added in the External resources container. Select *External resources* in the navigation tree, then on [CREATE RESOURCE](#). Enter the relevant information on the resource (at least a title), then provide either a filename (the file will then be uploaded on the server that hosts the Metadata Editor) or a URL to the resource.

External resources that have already been created for another project can also be imported. To do that, they must first be exported as JSON or RDF from the other project. Then click on [IMPORT](#) in the *External resources* page, and select the file.

Publish metadata and data to NADA

Metadata generated by the Metadata Editor (or by another tool that would generate metadata compliant with the same standards) serves as input that can (and should) be published in one or multiple (meta)data dissemination platforms. The World Bank has developed the NADA cataloguing application, an open source application fully compatible with the Metadata Editor.

Metadata exported from the Metadata Editor can be imported from NADA, using the NADA administrator interface. Another option, described below, is to publish the metadata (and data, for microdata) directly to NADA from the Metadata Editor.

Publishing metadata to NADA requires a NADA API key with permissions to publish in NADA. NADA API keys are generated from NADA (login to NADA, open your profile, and click on [Generate API key](#)).

⚠ API keys must always be kept strictly confidential. Your permissions and roles are embedded in the key; by sharing your API key, you would share your permissions. If you accidentally share or publish your key (for example in a script), you should immediately DELETE the key and issue a new one (in your *Profile* page in NADA).

[Home](#) / [Profile](#)

Profile

| | |
|---------|----------------------|
| | Edit |
| Name | |
| Email | |
| Company | |
| Phone | |
| Country | |

API keys

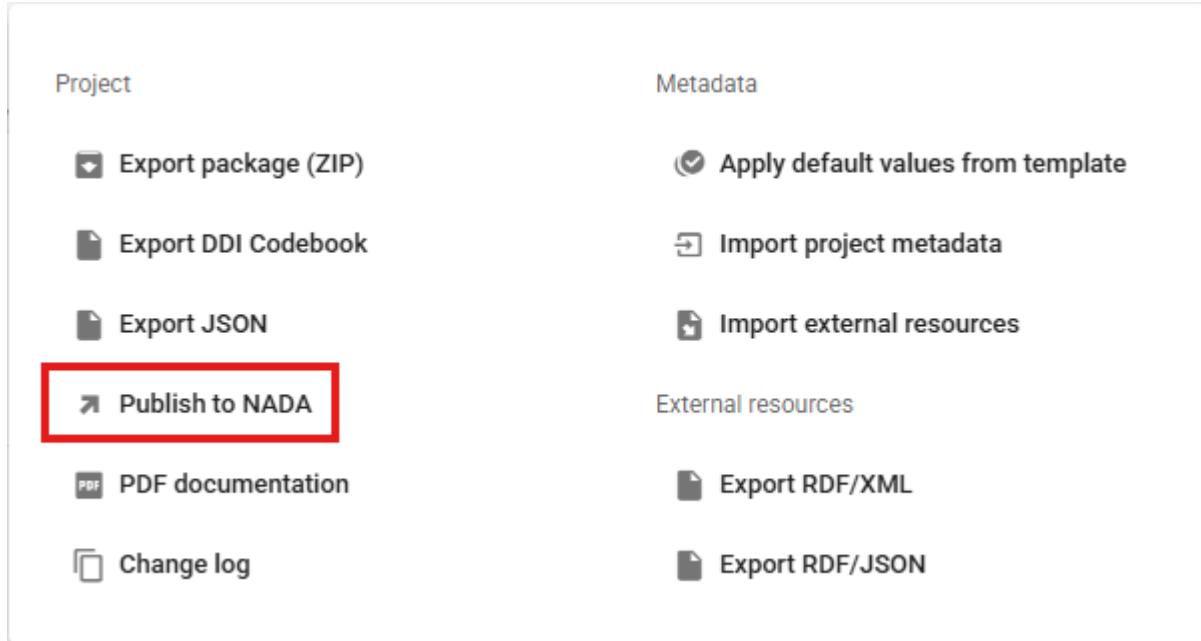
[Generate API key](#)

| | |
|--|------------------------|
| | Delete |
| | Delete |

If you have a NADA API key that grants you permission to publish your metadata (and possibly data) in NADA, you will have to first configure a data catalog (the one in which you plan to publish the metadata) in the Metadata Editor. Then you can publish metadata from any project to the NADA catalog.

Configuring a NADA catalog

In the project page of a project you own, co-own, or collaborate on, click on Publish to NADA in the project main menu.



In the **Catalog** frame, you will have the option to *Select* a NADA catalog among those you will have configured. If you have not configured any catalog yet, or to add a new one, click on **Configure new catalog**.

The screenshot shows the 'Publish to NADA' configuration dialog with the following interface:

- Publish to NADA** (Section title)
- Publish project directly to a NADA catalog** (Text)
- Catalog** (Link to 'Configure new catalog')
- Select-** (Dropdown menu placeholder)

The catalog configuration page will be displayed, where you must enter a short name for your catalog, its URL, and your NADA API key (which must be kept strictly confidential).

The screenshot shows two side-by-side configuration pages:

- Catalogs** (Left):

| Title | Catalog URL |
|--------------|-------------|
| Demo catalog | [Redacted] |
| NADA demo | [Redacted] |
- Configure new catalog** (Right):
 - Catalog title**: My new catalog
 - Catalog URL**: <https://nadacatalog.mysite.org>
 - API Key**: [Redacted]
 - SUBMIT** button

Publishing metadata in NADA

To publish metadata in NADA, click on "Publish to NADA" in the Project page menu. In the **Publish to NADA** page, select a configured NADA catalog then enter the **Project options**. These options are as follows:

- **Overwrite if already exists?** The options are Yes (overwrite) or No (do nothing). Answer the question on whether you want to overwrite the project if it has previously been published in the NADA catalog. A project is identified by its **Primary ID**. If a project already exists in NADA with the same primary ID, it will be overwritten if you answer **Yes**.
- **Publish.** A project can be pushed to a NADA catalog and made immediately accessible to users of the data catalog (option **Publish**), or it can be pushed but under a **Draft** status, in which case only administrators of the NADA catalog will be able to view it (and will have to publish it themselves by changing the project status in the NADA administrator interface).
- **Data access.** You may publish your data in NADA, and set the access policy for the dataset. See the next section.
- **Collection.** If the selected NADA catalog contains collections, this option allows you to select the collection in which you want to publish your metadata. A list of collections (if any) available in the selected NADA catalog will be displayed as a drop-down menu. The project will then be "owned" by this collection. Note that a same project can be shown in multiple collections in a NADA catalog; only one collection will "own" (and administer) the project, the others will only "borrow" it. Publishing a project in multiple collections is done in NADA, not in the Metadata editor.

Publish to NADA

Publish project directly to a NADA catalog

Catalog [Configure new catalog](#)

NADA demo - <https://nada-demo.ihsn.org>

```
{ "id": "16", "title": "NADA demo", "url": "https://nada-demo.ihsn.org", "user_id": "11" }
```

Project options

| Option | Value |
|---|--------------------------|
| Overwrite if already exists? | Yes |
| Publish | Publish |
| Data access  | Direct access - [direct] |
| Collection  | N/A |

In the **External resources** frame, select the files you want to upload to NADA. the files will be accessible to the visitors of the NADA catalog. Indicate whether you want to overwrite the files if they are already found in the NADA catalog.

In **Options**, select what information you want to publish:

- The project metadata
- The project thumbnail
- The external resources

Then click **PUBLISH**. If your metadata is valid (contains no validation errors), and your API key provides you with the necessary privileges on the NADA catalog, the metadata and related materials will now be published in the catalog.

External resources
Select external resources to publish

Overwrite resources

3 resources found 3 selected

| Title | Type |
|--|-----------------------------------|
| Survey questionnaire synthetic_survey_questionnaire.xlsx | Document, Questionnaire [doc/qst] |
| Survey information synthetic_survey_info.xlsx | Document, Technical [doc/tec] |
| Full dataset in Stata 17 format WLD_2023_SYNTH-SVY-EN_v01_M.zip | Microdata File [dat/micro] |

Options

- Publish project
- Publish thumbnail
- External resources (3)

PUBLISH

Publishing data in NADA

If you want to make your data files available to users in your NADA catalog, you must document the data files as **external resources** in your project (see the section on External resources). If the data are not intended to be openly accessible, and always for microdata, make sure you set the resource type as **microdata**.

A specific data access policy will be applied to all external resources identified as **microdata**. The **Data access** option provided in the **Publish to NADA** page is where you control how visitors of the NADA catalog will be able to access the data files.

Publish

| | |
|-------------|-----|
| Draft | |
| Data access | N/A |
| Collection | N/A |

External resources
Select external resources to publish

Overwrite resources

Data access (dropdown menu open)

- N/A
- N/A** (selected)
- Direct access - [direct]
- Public use files - [public]
- Licensed data files - [licensed]
- Data accessible only in data enclave - [data_enclave]
- Data available from external repository - [remote]
- Data not available - [data_na]
- Open access - [open]

It offers the following options:

- **Direct access.** Users will be able to download the data after accepting a disclaimer statement. No registration is needed.
- **Public use files.** Registered users will be able to download the data after accepting a disclaimer statement.
- **Licensed data files.** Registered users will have to submit a request for accessing the data, which will be processed by the NADA administrators. Data will be available for download after the user has been approved.
- **Data accessible only in data enclave.** Data are not accessible from the NADA catalog. They are only accessible on-site. When this option is selected, the microdata file(s) should not be published to NADA.
- **Data accessible from external repository.** Data are not accessible from the NADA catalog but from an external website. When this option is selected, the microdata file(s) should not be published to NADA. The URL to the external catalog must be provided.

| | |
|---|--|
| Data access  | <p>Data available from external repository - [remote]</p> <p>Link to remote repository</p> <input type="text"/> |
|---|--|

- **Data not available.** Data are not accessible from the NADA catalog or from any other source. When this option is selected, the microdata file(s) should not be published to NADA.
- **Open access.** Data are accessible under an open license (the most permissive option).

Introduction to the Metadata Editor API

The **Metadata Editor API** is a RESTful web service that allows users to interact programmatically with the Metadata Editor. It enables automation of key tasks such as uploading, transforming, validating, and exporting metadata across supported formats.

The API adheres to REST principles and supports standard HTTP methods (`GET`, `POST`, `PUT`, `DELETE`) for resource operations. All responses are returned in JSON format, and endpoints are secured using API keys tied to user permissions.

Key Features

- Upload and retrieve metadata files
 - Convert between metadata formats (e.g., CSV to DDI XML)
 - Validate metadata against predefined schemas
 - Manage projects, datasets, and schema mappings
-

Programming Language Support

The Metadata Editor API can be used with any language that supports HTTP requests. Official libraries are available for the following languages:

Python

A Python client library is provided to simplify interaction with the API using familiar data structures like `pandas.DataFrame`.

See: [Python package →](#)

R

An R package is also available, enabling integration into R-based data workflows and analysis scripts.

See: [R package →](#)

Refer to the following chapters for detailed usage instructions, including authentication, endpoint references, and code examples in both Python and R.

Warnings and Recommendations

Access to the Metadata Editor API requires a valid **API key**.

Key Ownership and Permissions

- Each API key is uniquely tied to a **registered user account**.
- The key carries the **same permissions and roles** as the user within the Metadata Editor interface.
- Any action permitted via the UI is also permitted through the API — and vice versa.

Security Guidelines

- **Keep your API key secret.** Do not share it or expose it in public repositories, scripts, or notebooks.
- Treat your API key like a **password**. Hardcoding it in plaintext files is strongly discouraged.
- If you believe your API key has been compromised:
 - **Revoke it immediately** and generate a new one.
 - **Notify your system administrator** so usage logs can be reviewed for unauthorized activity.

Additional Security Recommendations

- Use environment variables or secure credential stores (e.g., `.env` files, secret managers) to manage keys in production environments.

By following these precautions, you help ensure the integrity and security of your metadata workflows.

Getting Started

To begin using the Metadata Editor API, you'll need to generate an API key and use it to authenticate your requests.

Step 1: Generate an API Key

1. Log in to the Metadata Editor through the web interface.
2. Navigate to **User profile** page.
3. Click on "**Generate API Key**".
4. Copy and securely store your key.
5. Use this key in your API requests as an `X-API-KEY` header:

`X-API-Key: YOUR_API_KEY_HERE`

`http`

Quick examples

Python example using `requests`

```
import requests

API_KEY = "your_api_key_here"
headers = {"X-API-Key": API_KEY}

response = requests.get("https://your-metadata-editor.org/api/projects", headers=headers)

print(response.json())
```

Python example using `requests`

```
import requests

API_KEY = "your_api_key_here"
headers = {"X-API-Key": API_KEY}

response = requests.get("https://your-metadata-editor.org/api/projects", headers=headers)

print(response.json())
```

R example using `httr`

```
library(httr)

api_key <- "your_api_key_here"
url <- "https://your-metadata-editor.org/api/projects"

res <- GET(url, add_headers(`X-API-Key` = api_key))
content(res, "parsed")
```

Developers

The source code of this application is available on GitHub and is published under the MIT License. We warmly welcome contributions from the developer community and others who are interested in improving the application.

If you would like to contribute code, please start by forking the repository, creating a feature branch, and submitting a pull request. We encourage you to review any available contribution guidelines in the repository before submitting your changes.

Even if you are not able to contribute code directly, your ideas and suggestions are valuable. Please feel free to open an issue to report bugs or suggest enhancements. You can also participate in or start discussions to share broader feedback or propose new features.

We ask all contributors to follow international best practices for open source collaboration. This includes:

- Communicating respectfully and constructively;
- Writing clear, maintainable, and well-documented code;
- Following the project's coding and documentation standards;
- Testing contributions thoroughly before submitting;
- Being open to feedback and iterative improvement.

By contributing, you help make the software better for everyone. Thank you for your interest and support!

Translation of the software and templates

Translating the Metadata Editor software application

The Metadata Editor is designed following internationalization best practices and can be easily translated into other languages using the built-in translation tool.

Translations generate a set of PHP files (.php extension), which must be uploaded to the server where the Metadata Editor is installed. Instructions on where to save these files and how to activate a new language are provided in the following sections.

Translation files can be shared with other organizations using the Metadata Editor. You are encouraged to share your translations and inform the Metadata Editor maintenance team at the World Bank of their availability. Validated translations provided as open materials may be published in the Metadata Editor GitHub repository.

Supported languages

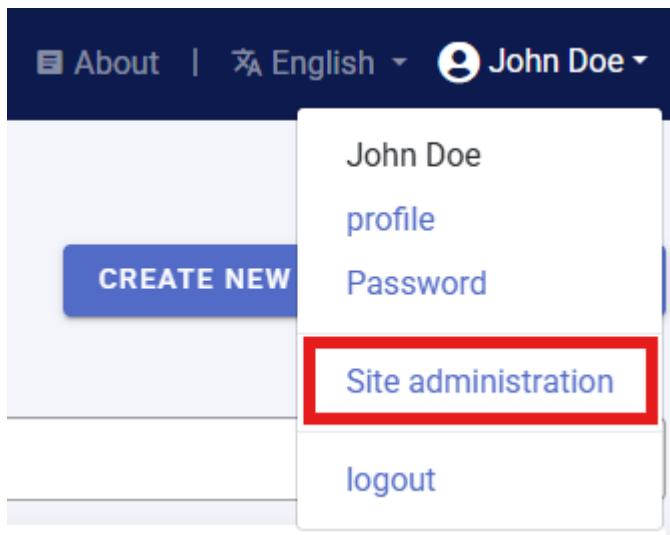
The Metadata Editor supports translation into any UTF-8 encoded language, including Arabic, Chinese, and Russian.

- For right-to-left (RTL) languages like Arabic, the user interface (UI) must be adjusted to ensure correct display.
- Translations only affect the labels and text, not the overall UI layout.

Note: The application also allows translation from a language to itself (e.g., from English to English). This option can be used to modify the display of labels or text to align with specific terminology used by your organization.

Adding a new language

If the application is not available in your language, you can create a translation through the user interface. This requires System Administrator privileges. If you have been assigned this role, a *Site Administration* option will appear in the menu when you click on your name.



To add a new language, a folder must be created in the application/languages directory on the server where the Metadata Editor is installed. For example, to create a Spanish translation, create a folder named Spanish.



No UI is provided for creating this folder; it must be done manually.

Once the folder is created, it will appear in the *Translate* section of the UI.

Translate

Template language set to: BASE

| Language | actions |
|----------|-----------------|
| english | Edit download |
| french | Edit download |
| spanish | Edit download |

Creating or editing a translation

To create or edit a translation:

1. Click on [Settings](#).
2. Select [Translate](#) from the Site Administration menu.
3. Select the language from the translations page.
4. Click [Edit](#).
5. Items without a translation will be highlighted in red. Enter your translation in the text box. Note that if the translation is significantly longer than the original text, it may not display properly in limited-space areas such as buttons and menus.

Translate

Base template language: BASE

| Select language to translate |
|--|
| spanish <input type="button" value="▼"/> |
| breadcrumbs |
| configurations |
| general |
| install |
| permissions |
| rest controller |
| template manager |
| user groups |
| users |

File: spanish / permissions /Volumes/webdev/editor/application/language/spanish/permissions_lang.php

| Key | Translation |
|------------------|------------------|
| permission | Permission |
| permissions | Permissions |
| administrator | Administrator |
| permissions_list | Permissions List |
| edit_permission | Edit Permission |

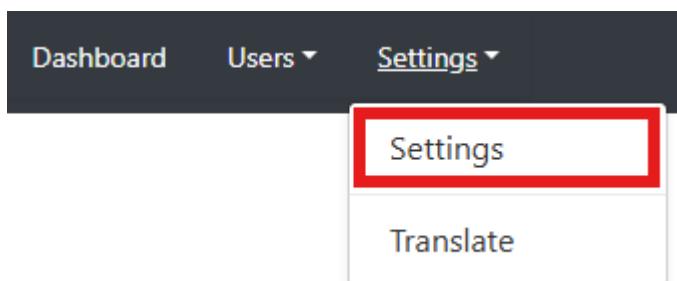
After translating each section, your work.

Importing a translation

Activating a translation

To activate a new translation as the default language:

1. Open [Settings](#) from the Site Administration menu.



2. Select [Language](#).

site_configurations

— General site settings

Language

Language

english 

— Use HTML Editor for HTML editing?

— Survey catalog settings

— Site login

— Google Analytics

— SMTP settings

3. Choose the desired language from the list of available translations.

This will set the new language as the default for your instance of the Metadata Editor.

Enabling multiple languages

The Metadata Editor supports enabling multiple languages, allowing users to switch between them. To enable multiple languages:

1. Edit the the *supported_languages* line of the application/config/config.php file as follows.

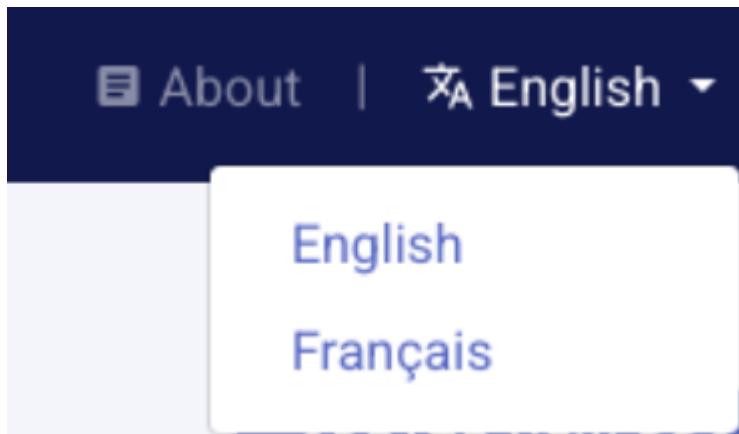
```

71  /*
72  |-----
73  | Default Language
74  |-----
75  |
76  | This determines which set of language files should be used. Make sure
77  | there is an available translation if you intend to use something other
78  | than english.
79  |
80  */
81 $config['language'] = "english";
82
83 //List of supported languages. type language name in lower case
84 $config['supported_languages']=array("english","french");

```

2. Modify the supported_languages line to include the desired languages.

Once configured, a dropdown menu will appear in the site navigation menu, allowing users to select their preferred language.



Translating metadata templates

When translating the Metadata Editor, you may also want to translate the metadata templates, as they also define the User Interface (UI) seen by data curators.

To translate templates:

1. Open the *Template Manager* (see the section on *Designing Templates*).
2. [Duplicate](#) the template you want to translate, and edit the description of the copy (including the name and language information).
3. Translate all labels, instructions, and controlled vocabularies.
4. Set a translated template as the default for your language.

By translating both the UI and the templates, you ensure a seamless experience for users across different languages.

Feedback and contact

We welcome feedback, suggestions, and reports of issues related to the use of the Metadata Editor. The primary channel for communication is the GitHub repository (<https://github.com/worldbank/metadata-editor>), where you can:

- Submit an **issue** to report bugs, request features, or suggest improvements.
- Participate in **Discussions** to ask questions, share use cases, or provide general feedback.

Please note that we are not able to provide individual technical support or financial assistance for the use of the Metadata Editor outside the official activities of our organization.

Thank you for your interest and contributions.

Useful resources and links

Metadata standards

- DDI Alliance: <https://ddialliance.org/>
- DDI Codebook metadata standard: <https://ddialliance.org/getting-started-ddi-c>
- Dublin Core: <https://www.dublincore.org/>
- ISO 19139: <https://www.iso.org/standard/67253.html>
 - INSPIRE directive (European Union): https://knowledge-base.inspire.ec.europa.eu/index_en
 - GEMINI (UK): A description of the elements included in the GEMINI template.
<https://agiorguk.github.io/gemini/1062-gemini-datasets-and-data-series.html>
- IPTC Photo Metadata Standard: <https://iptc.org/standards/photo-metadata/iptc-standard/>
- schema.org: <https://schema.org/>
- Croissant: <https://github.com/mlcommons/croissant>
- Data Catalog Vocabulary (DCAT): <https://www.w3.org/TR/vocab-dcat-3/>

Related tools

- NADA cataloguing application: an open source cataloguing application compatible with all metadata standards supported by the Metadata Editor
- R package **MetadataEditR**: a R package, available on GitHub, to support the use of the Metadata Editor API and to automate tasks
- Python Library: PyMetadataEditor: a Python library, available on GitHub, to support the use of the Metadata Editor API and to automate tasks
- AI tools for metadata evaluation and enhancement: a collection of AI tools developed by the World Bank Office of the Chief Statistician

Training and advocacy materials (forthcoming)

- eLearning course: Introduction to metadata and to the Metadata Editor
- Hands-on training program on data documentation and the Metadata Editor
 - Recommended profile of trainers and trainees
 - Recommended agenda for a hands-on training on data documentation using the Metadata Editor

- Presentations and training materials