Reference Guide on Metadata Standards for Development Data

Version 1.0.0 (2026-02-20)

A guide by the Office of the World Bank Group Chief Statistician and Chief Data and Technology Officer

Office of the Chief Statistician and Data Technology Office

License. This document was created by the World Bank Group and is published under a CC BY 4.0 Attribution license.

Disclaimer. While every effort has been made to ensure the reliability of the content of this guide, the World Bank Group makes no warranties or representations regarding its completeness, suitability, or applicability to any specific use or context. Users outside the World Bank Group are solely responsible for how they interpret, apply, or modify the information contained in the guide.

Use of artificial intelligence (AI). AI tools, including ChatGPT 4o by OpenAI and Microsoft Copilot's Auto Rewrite tool, were used in the development of this guide. The initial content was drafted by the author; AI was used to suggest, enhance, and edit text for clarity and consistency. All AI-generated content was reviewed for accuracy and relevance.

Maintenance. This guide is maintained on GitHub as Markdown, supporting version control, collaborative editing, and automatic conversion to HTML and PDF.

github.com/worldbank/guide-metadata-standards

Foreword

The World Bank Group ("the Bank") possesses an extensive body of knowledge related to social and economic development. A significant portion of this knowledge is embedded within **development data**, which plays a crucial role in informing, monitoring, and evaluating the Bank's operations, as well as in supporting research on development issues. Recognizing the importance of its data assets to the broader development community, the Bank launched its Open Data Initiative in 2010.

Sharing and utilizing development data carries significant responsibilities, with a primary focus on maintaining **high standards of data quality**. In October 2024, the Bank introduced a *Development Data Quality Policy* as a framework to uphold excellence in its data products, processes, systems, and tools. Equally vital is the need to **promote effective data use** among a broad and diverse audience. This means data should be easily discoverable, accessible, interpretable, and usable not only by people but also by machines, including AI-driven applications. Meeting these goals demands providing detailed, structured, high-quality, AI-ready metadata in convenient formats for all stakeholders.

This Guide serves as a reference for the creation and maintenance of metadata for development data within the Bank, ensuring compliance with metadata standards and international best practices. It forms part of the *Development Data Quality Policy* implementation guidance. It is primarily intended for Bank staff and business units involved in data production, curation, and dissemination, as well as developers responsible for tools and systems used to manage or disseminate development data. The Guide may also be applicable to external organizations engaged in the production, curation, and dissemination of data.

The Guide is supported by a set of tools developed by the Bank to promote the implementation and utilization of standard-compliant metadata. This set includes a dedicated metadata editor application, AI-powered utilities for metadata assessment and enhancement, and data cataloging systems. These resources are made openly available to users, except in cases where legal or contractual limitations restrict access.

This document may be periodically updated to reflect changes in metadata standards and advancements in technology.

Haishan Fu World Bank Group Chief Statistician	Roman P. Kovalenko World Bank Group Chief Data Technology Officer

Acknowledgments

The Guide was written by **Olivier Dupriez** (Deputy Chief Statistician, Office of the Chief Statistician and Development Data Group) with the assistance of **Robert Malloy** (Manager, WBG Data, Information & AI Governance) and **Arvind Srinivasan** (Manager, Information and Technology Solutions, Data & Analytical Solutions).

Mehmood Asghar (Senior Data Scientist, Development Data Group - Data Governance) co-led the development of metadata schemas described in the Guide and served as the lead developer of the Bank's Metadata Editor software application. **Aivin Solatorio** (Senior Data Scientist, Development Data Group, Data Governance) offered guidance on AI solutions for metadata creation and assessment.

Ivan Donev (Data Platforms Lead, Information and Technology Solutions - Data & Analytical Solutions), **Jeanne Kramer-Smyth** (Senior IT Officer, Information Governance & Archives), **Monisha Menon** (Data Governance & Privacy Specialist, Information and Technology Solutions, Data & Analytical Solutions), and **Jeannette Marie Smith** (IT Officer, Information and Technology Solutions - Knowledge Management and Learning) provided feedback and input.

Jorn J. Wegter (Senior Counsel, Institutional Administration), who co-led the development of the Development Data Quality Policy, provided advisory support to the team.

Francesco Ciriaci (Privacy Officer, Office of the Chief Data Privacy Officer) and **Ahmed Khalifa** (Privacy Analyst, Office of the Chief Data Privacy Officer) reviewed the document on behalf of Data Privacy Office to ensure consistency with the Bank's Policy on Data Privacy.

Haishan Fu (World Bank Group Chief Statistician) and **Roman P. Kovalenko** (World Bank Group Chief Data Technology Officer) provided overarching guidance and enabling support of this guide's production.

Acronyms

Acronym	Definition
AI	Artificial Intelligence
API	Application Programming Interface
CSV	Comma-Separated Values
DCAT	Data Catalog Vocabulary
DCMI	Dublin Core Metadata Initiative
DDI	Data Documentation Initiative
DOI	Digital Object Identifier
DSD	Data Structure Definition
EXIF	Exchangeable Image File Format
FGDC	Federal Geographic Data Committee (USA)
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
IMF	International Monetary Fund
IPTC	International Press Telecommunications Council
ISO	International Organization for Standardization
IT	Information and Technology
ITS	Information and Technology Solutions
JPEG/JPG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
LLM	Large Language Model
MARC21	Machine-Readable Cataloging
MSD	Metadata Structure Definition
OAI-PMH	Open Archives Initiative - Protocol for Metadata Harvesting

Acronym	Definition
OECD	Organisation for Economic Co-operation and Development
PDF	Portable Document Format
PNG	Portable Network Graphics
RDA	Research Data Alliance
RDF	Resource Description Framework
SDMX	Statistical Data and Metadata Exchange
SEO	Search Engine Optimization
SQL	Structured Query Language
TIFF	Tagged Image File Format
WDI	World Development Indicators
XML	Extensible Markup Language
WB	World Bank
WBG	World Bank Group

Introduction

Launched in 2010, the World Bank's Open Data Initiative made the Bank's statistics and indicators freely accessible to researchers, policymakers, journalists, and other stakeholders. In 2024, the introduction of the *Knowledge Compact for Action* marked the debut of an updated *Data Bank*, committed to optimize data utilization through increased investments in data systems and advanced analytics.[1] In this context, the Bank issued a *Development Data Quality Policy* in October 2024, which **mandates the adoption and implementation of metadata standards** for the curation and dissemination of Bank development data.

> **Bank Development Data** is defined in the Bank's *Development Data Quality Policy* as "[a]ny Data Product that (i) is not Third Party Data, and (ii) is produced or disseminated by the Bank or by the Bank in collaboration with a third party for the purpose of facilitating development or otherwise relating to any development activity of the Bank (which for the avoidance of doubt excludes Bank Corporate Data)."

> **Development data** encompass a wide variety of information, including statistical indicators, microdata from surveys, censuses and other sources, as well as geographic datasets. They may also include unstructured data—such as text, audio recordings, videos, and images—when these are used as media for publishing development data or as inputs for artificial intelligence applications, either to train models or to extract development-relevant information from the underlying resources.

The adoption of metadata standards aims to guarantee the availability of comprehensive and structured metadata, providing critical information about the context, quality, structure, and other characteristics of the data. Metadata that conforms to these standards fosters responsible and efficient management, dissemination, and use of data, enhancing discoverability, clarity, interpretability, and trustworthiness. Metadata also plays a crucial role in the Bank's efforts to develop cost-effective and interoperable data management and dissemination systems and tools, and to make its data products AI-ready.

This reference guide serves as an internal resource for the World Bank Group (WBG). It outlines and justifies the Bank's metadata strategy for development data. It specifies the principles, procedures, tools, roles, and responsibilities necessary to comply with metadata standards. While primarily designed for internal use, this guide—and the tools created to support its implementation—also serves as an advocacy instrument for promoting the adoption of metadata standards among partner agencies and client countries.

The guide is divided into three chapters. The first chapter explains the Bank's approach to managing metadata. The second chapter introduces structured metadata and relevant standards, covering their definitions, formats, benefits, and how they are created, checked, and used. It also covers key concepts like metadata templates, controlled vocabularies, quality control, governance, sharing, archiving, and the role of AI in improving metadata production. The third chapter addresses how the Bank chooses and applies metadata standards, outlining the processes for selection, development, maintenance, and versioning. It also includes administrative and relationship metadata, plus coordination with global standards. Annexes offer advice for system developers, data and metadata procurement, and provide a list of useful resources.

---

1. The Knowledge Compact for Action: From Ideas to Development Impact., World Bank Group, April 2024. ↩

1. Metadata Strategy for Development Data

The adoption of metadata standards represents a fundamental aspect of the Bank's Policy on Development Data Quality. This decision is designed to improve the usability, credibility, and discoverability of development data, while fostering interoperability among the Bank's diverse data management and dissemination systems. With its metadata strategy, the Bank seeks to implement a structured approach for developing, maintaining, and applying standardized metadata throughout its data operations. This framework is designed to promote consistency, enhance accessibility, and facilitate the integration of data products across various departments and platforms.

Metadata Strategy for Development Data

Goal and Objectives

1. The World Bank Group's (WBG) *Development Data Quality Policy* mandates the adoption of metadata standards for development data, aiming to establish a uniform metadata management framework. The metadata strategy for development data aligns metadata production, management, and dissemination with the Bank's commitment to data quality. It ensures that development data generated, used, and disseminated by the Bank is discoverable, trustworthy, usable, and AI-ready. The goal of the strategy is to empower stakeholders to unlock the full potential of the Bank's development data for informed decision-making, analysis, and research.

2. The objectives of the metadata strategy are:

    a. Ensure metadata associated with development data curated by the Bank complies with a set of metadata standards identified by the WBG Chief Statistician, aligning with international best practices to facilitate data interoperability and integration within the Bank and the global data ecosystem.

    b. Develop, deploy, and maintain robust systems for cataloging, indexing, and publishing development data and metadata to enhance data discoverability and accessibility for internal and external stakeholders.

    c. Enable data users to understand, responsibly utilize, and effectively apply development data across various domains.

    d. Support integration of Bank development data with advanced knowledge discovery systems and AI models by ensuring metadata is comprehensive and semantically rich, machine-readable, and machine-understandable.

    e. Define criteria and implement processes to regularly assess and update metadata quality.

    f. Engage development data producers, curators, and end-users in the continuous improvement of metadata practices.

    g. Foster innovation by ensuring development data is reusable and leveraged for research, collaboration, and new insights across sectors.

Planning and Readiness

3. The WBG Chief Statistician selects, develops, maintains, and documents the metadata standards used by the Bank for documenting and cataloguing development data. An advisory team, established by the WBG Chief Statistician and comprising staff from the Development Data Group, Information and Technology Solutions, and other relevant Bank departments, provides guidance to the Chief Statistician regarding the selection and maintenance of metadata standards.

4. The Bank collaborates with international partners, standards setters, and national organizations to periodically update the standards.

5. The Bank maintains a metadata editor software application for the production and management of metadata for development data. The application's programming interface (API) and complementary tools are developed and used to automate, when applicable, the production, maintenance, quality assurance, use, and sharing of metadata.

6. The toolset developed by the Bank for the curation of metadata for development data is made openly accessible, when technically and legally feasible, to support the adoption of metadata standards outside the Bank.

7. The Office of the WBG Chief Statistician oversees the creation, dissemination, and maintenance of metadata templates designed to support the consistent application of the metadata standards.

Metadata Creation and Quality Assurance

8. Business units responsible for the creation or maintenance of development data are responsible for the creation and versioning of metadata in compliance with Bank standards. Metadata creation is embedded throughout the data lifecycle to ensure accuracy, relevance, and alignment with data quality objectives. A data curation team within the Development Data Group provides advisory and technical support.

9. For third-party development data acquisitions, the responsible business unit must ensure that all required metadata accompanies the data. When the Bank purchases data, it is recommended to utilize the Bank's metadata templates to guide vendors regarding the expected content and format of the metadata. If the provided metadata does not align with the Bank's metadata standards, it is the responsibility of the acquiring business unit to generate metadata that meets compliance requirements.

10. Administrators of data management and dissemination systems are responsible for creating the administrative and usage metadata, leveraging automation through the Bank's metadata editor API.

11. Business units are accountable for evaluating metadata quality and ensuring that metadata is updated when data undergoes changes. The Office of the Chief Statistician offers guidance and resources to facilitate thorough assessments of metadata.

12. The Bank develops AI-driven solutions to facilitate the efficient production, assessment, and enhancement of metadata compliant with its standards.

13. The Bank promotes metadata standards and best practices across partnerships and collaborations with external organizations and seeks to support their implementation in client countries in the context of its statistical and data-related technical and financial support.

14. Metadata for World Bank Group data products is generated in English. Metadata obtained from external sources and metadata for country-specific data products created with a partner in the country are produced in the primary national languages. For datasets acquired from third parties and used by the Bank, metadata in languages other than English, French, or Spanish is translated into English when resources allow.

Storage, Versioning, Preservation, and Dissemination

15. The Bank's metadata editor application serves as the primary metadata repository for development data, with access controls managed by the data curators and regular backups managed by the Bank's Information and Technology Solutions (ITS) in accordance with the Bank's data security policy and practice.

16. Metadata must be maintained in alignment with data updates. A semantic versioning system is recommended to ensure traceability and consistency.

17. Metadata may be decommissioned when no longer relevant. Descriptive and reference metadata are, however, archived indefinitely unless contractual or ethical obligations require their deletion.

18. Metadata is stored and shared in open formats to ensure long-term accessibility and usability.

19. Personal data should not be present in metadata, except for information on authorship of documents, datasets, or other resources. Personal data found in metadata is treated in compliance with the Bank's Policy on Personal Data Privacy.

20. The Bank disseminates metadata alongside published data in user-friendly and machine-readable formats, including via application programming interfaces (APIs), under an open license by default.

2. Metadata: Definitions, Benefits, Creation, and Use

This chapter introduces structured metadata and relevant standards, covering their definitions, formats, benefits, and how they are created, checked, and used.

Chapter outline

- 2.1 Metadata
- 2.2 Structured Metadata
- 2.3 Metadata Standards
- 2.4 Metadata Templates
- 2.5 Choosing the Right Level of Metadata Production
- 2.6 Generating Metadata Compliant with Standards
- 2.7 Metadata Quality Control
- 2.8 Metadata Governance
- 2.9 Metadata Dissemination and Use
- 2.10 Archiving and Preserving Metadata

2.1 Metadata

2.1.1 Definition

Metadata refers to information that describes, explains, locates, or otherwise makes it easier to retrieve, interpret, use, or manage data.[1] It may also include information that describes how data should be published, managed, and displayed. Metadata must be made accessible in forms and formats that are convenient for both users and machines.

It is essential that metadata accurately matches the dataset it represents and gets updated anytime the data changes. This makes keeping versions of data and metadata files important, with semantic versioning being the suggested approach.

2.1.2 Types

Metadata encompasses several categories, each fulfilling a distinct role. Certain types of metadata facilitate data users in locating, evaluating, interpreting, and utilizing datasets efficiently, whereas others are intended to assist administrators of data repositories in the management and distribution of data.

Metadata relevant to data users

- **Descriptive metadata** identifies and describes a dataset, typically including elements such as the dataset title, abstract or summary, relevant keywords and topics, information on producers and publishers, geographic and temporal coverage, and versioning details. Descriptive metadata encompasses **cataloguing metadata**, a subset used specifically to register and manage datasets within catalogs or repositories, which is vital for ensuring datasets are both findable and citable.

- **Reference metadata** offers contextual and methodological details about the dataset. Common components include conceptual definitions, sources of data, methodologies employed, sampling and processing procedures, quality assessments, and known limitations.

- **Structural metadata** specifies the format and organization of the dataset, as well as relationships among its elements, facilitating effective analysis and integration into workflows. For microdata, structural metadata may consist of a *data dictionary* detailing files and variables; for indicators, it involves a *data structure definition* describing the data model and structure; and for geographic datasets, it takes the form of a *feature catalog*.

Collectively, descriptive, reference, and structural metadata empower data users to evaluate a dataset's suitability for their needs, comprehend its scope and coverage, and utilize it appropriately. These metadata types are indispensable for data producers and curators, as they underpin the usability, reliability, and discoverability of datasets. The creation and maintenance of such metadata is mandated by the Development Data Quality Policy and constitutes a primary focus of the standards detailed in this guide.

Metadata intended for catalog and systems administrators

Metadata relevant to catalog and systems administrators is tailored to the specific IT environments in which catalogs and systems function. As such, no universal standard applies to these metadata types; however, it remains essential that they are structured and formatted consistently alongside descriptive, reference, and structural metadata.

- **Administrative metadata** encompasses information supporting dataset management, security, and governance by catalog and system administrators. This includes details regarding dataset origin, acquisition date, and access rights, such as licensing terms and permissions. Administrative metadata defines conditions for publishing or accessing datasets, specifying authorized users, permitted purposes, and applicable constraints. Additionally, it captures operational and technical requirements necessary for effective dataset administration—including file formats, storage needs, and system dependencies. Proper documentation of administrative metadata ensures secure, efficient, and compliant management in accordance with institutional policies and procedures. Its context-specific nature means the structure and content will vary among organizations, and no standard is mandated.

- **Presentation and portrayal metadata** establish protocols for the visual representation of data, promoting clarity, consistency, and accurate interpretation within dissemination platforms and data catalogs. As optional elements—since many systems default to preset presentation and portrayal parameters—presentation metadata (sometimes referred to as *display or visualization metadata*) describes general rules for non-spatial data displays, such as chart types, color schemes, symbols, axis labels, legends, and formatting conventions. These components facilitate adherence to organizational standards and visualization style guides and consistent rendering across platforms. Portrayal metadata, conversely, focuses on the visualization of geographic or spatial datasets by providing specific cartographic instructions, including symbology, feature styling, color palettes, scale-dependent rendering, and related elements. Collectively, presentation and portrayal metadata ensure statistical and spatial data are presented in a clear, standardized, and comprehensible manner.

- **Usage and impact metadata** serve distinct functions in understanding dataset engagement and influence. Usage metadata tracks how datasets are accessed and utilized, gathering metrics such as download counts, API usage, page views, and geographic patterns of user activity. Impact metadata documents instances where datasets contribute to research, policy formation, or decision-making, reflected in citations, references, or indicators of user engagement and reach. Combined, these metadata dimensions offer valuable insights into dataset relevance and effectiveness. For institutions such as the Bank, this supports trend analysis, user experience optimization, and enhances visibility and credibility in the global data environment. They further contribute to catalog ranking and resource discovery strategies. Usage data should be collected via appropriately configured web analytics tools for real-time reporting, and while integration into records is optional, automated dynamic updates are recommended for accuracy. Impact metadata generally requires ongoing monitoring of external dissemination channels, such as publication databases.

- **Relationship metadata** articulate links between datasets and associated resources, playing a critical role in documenting data lineage and dependency. This metadata identifies source datasets, derivative outputs, and connections to supplementary materials (e.g., publications, methodological documents, code repositories). For example, relationship metadata may associate a working paper with its referenced original dataset, derived analytical datasets, and reproducible scripts used in processing. Maintaining these connections enables catalog administrators and repository managers to develop integrated, interoperable systems that foster discoverability, traceability, and data reuse. Relationship metadata facilitating data interpretation and supporting reproducibility should be made accessible to end users.

---

1. Definition adapted from *Understanding Metadata*, National Information Standards Organization, 2004. ↩

2.2 Structured Metadata

2.2.1 Definition

To ensure maximum utility, metadata should be generated and maintained in a structured format. **Structured metadata** comprises an organized set of elements, each representing a distinct characteristic of a dataset—such as the dataset title, geographic coverage, temporal scope, producer information, or details for each variable including names and labels. For metadata to be effectively leveraged by computer applications, every element should include the following attributes:

- **Key**: The unique identifier assigned to each metadata element.
- **Type**: Elements may take various forms such as text (string), numbers (numeric), arrays, or Booleans (logical values).
- **Label**: A succinct descriptor for the element.
- **Description**: A detailed description of the metadata element, which can serve as guidance or instructions for data curators.
- *Repeatable status:** Indicates whether an element can support multiple entries.
- **Required status**: Clarifies if the element is "required" (mandatory), "optional," or possibly "recommended."
- **Controlled vocabularies**: These provide predetermined lists of permissible values to promote consistency and uniformity within metadata. Controlled vocabularies apply exclusively to selected metadata elements.

The types of metadata used to describe a dataset vary based on the type of data. This guide looks at several primary data types: **microdata** (individual-level data from surveys, censuses, administrative records, or sensors), **indicators**, **geographic datasets**, **documents**, **images**, **videos**, **audio recordings**, and **scripts**. Every type has unique metadata requirements and a specific metadata structure, though some overlap exists. For instance, microdata metadata often includes information about sample design or data collection methods, which are important for survey and census data but not for other kinds. On the other hand, geographic datasets need details like coordinate systems and spatial resolution, which aren't relevant elsewhere. By defining metadata specific to each type of data, users can better understand, interpret, and make use of the datasets.

The range of metadata elements applicable for documenting any dataset can be extensive. To promote clarity and enhance usability, these elements are organized into logical categories and subcategories, either for practical purposes or because they reflect different facets of the dataset, such as descriptive, structural, or administrative metadata. For instance, when documenting a survey dataset, metadata elements are typically arranged into groups that delineate the survey's attributes (*study description*, including title, objective, scope, coverage, timeframe, and sampling methodology), the data files (such as file architecture and format), and the individual variables (including names, definitions, value labels, and coding schemes).

2.2.2 File Formats

Structured metadata is generally maintained in a metadata file produced in either JSON (JavaScript Object Notation) or XML (Extensible Markup Language) format. The selected format should be open, adaptable, and specifically designed to support usability and interoperability across various platforms. It must accommodate both machine and database

processing while enabling seamless transformation into human-readable formats such as HTML, PDF, or formatted text. Furthermore, structured metadata should be accessible and functional via APIs to foster integration and reuse. For metadata associated with development data, the Bank prefers the JSON format due to its enhanced usability, lightweight design, and robust interoperability. JSON offers the following advantages:

- It is lightweight and human-readable, as JSON uses a plain text format that is straightforward to review and modify.
- Its hierarchical structure supports flexible representation of complex metadata schemas through key-value pairs.
- JSON is extensively supported across programming languages, ensuring broad compatibility.
- It can be easily parsed by software applications, promoting automation as well as efficient publishing and indexing within databases; moreover, JSON is readily ingested by both SQL and NoSQL databases.
- The format facilitates uncomplicated conversion into other file types, such as HTML or PDF.

The following is an example of core metadata describing a book, in JSON format.

```json
{
  "document_description": {
    "title_statement": {
      "title": "The Analysis of Household Surveys",
      "subtitle": "A Microeconometric Approach to Development Policy",
      "idno": "10.1596/978-1-4648-1331-3"
    },
    "date_published": "2010/07/01",
    "authors": [
      {
        "first_name": "Angus",
        "last_name": "Deaton"
      }
    ],
    "type": "Book",
    "languages": [
      {
        "name": "English",
        "code": "EN"
      }
    ],
    "abstract": "This book reviews the analysis of household survey data, including the construction of h
  }
}
```

2.2.3 Benefits

Effective data management and quality assurance depend heavily on high-quality metadata, and aligns with the FAIR principles: Findability, Accessibility, Interoperability, and Reusability. [1] Well-organized and detailed metadata allows users to handle data responsibly and accurately. Additionally, it strengthens systems for discovering and sharing data. By enriching metadata, traditional keyword searches become more effective, and more advanced search methods like semantic and hybrid search are enabled—turning data catalogs into smart data recommendation tools.

The increasing adoption of generative AI technologies underscores the critical role of high-quality metadata. As methods for accessing and utilizing data evolve, users are increasingly interacting directly with AI-enabled systems, including internet search engines, chatbots, and advanced data platforms. In order to produce reliable outcomes, metadata must provide AI systems with comprehensive context, incorporating elements such as definitions, limitations, sources, and methodologies. Metadata should be semantically robust, machine-readable, and machine-understandable to ensure accurate interpretation and effective utilization by AI technologies. Furthermore, AI can play a pivotal role in enhancing metadata quality through processes such as metadata augmentation and evaluation.

Structured metadata serves a critical role in numerous aspects of data management and dissemination:

1. **Enhancing Data Quality:** Metadata fosters superior data quality by ensuring accuracy, consistency, and interpretability. Through clear definitions and documentation of collection and analysis methods, it supports validation, facilitates effective auditing, and enables the resolution of errors.

2. **Promoting Metadata Completeness:** Well-defined metadata structures function as comprehensive checklists encompassing required, recommended, and optional elements for dataset documentation. Compliance with these standards guarantees systematic inclusion of essential information, thereby minimizing omissions.

3. **Strengthening Metadata Quality Assurance Processes:** Standardized metadata contributes to more effective quality assurance by establishing uniform criteria for evaluating consistency, and accuracy in data descriptions.

4. **Facilitating Data Discoverability:** The use of standardized terminology and structures within metadata enhances discoverability, optimizing data indexing across catalogs and dissemination platforms. Detailed metadata additionally supports the advanced capabilities of artificial intelligence systems, and the creation and optimization of sophisticated search functionalities.

5. **Supporting Data Harmonization and Integration:** High-quality metadata is indispensable for addressing inconsistencies and enabling harmonization efforts—both ex-ante (during data design) and ex-post (after data creation).[2] It provides foundational frameworks and resources to standardize concepts, methods, and formats, as well as retrospective reconciliation and integration, thus improving utility for longitudinal, comparative, and multi-source analyses.

6. **Improving Data Usability and Reusability:** Comprehensive metadata facilitates understanding, sharing, and repurposing of datasets by elucidating data relationships, hierarchies, and context. Detailed documentation reduces dependency on original data teams, streamlining secondary usage and reducing the burden of data dissemination.

7. **Enabling Effective Data Governance:** Metadata underpins robust governance by recording provenance, lineage, and modifications, which supports auditability and ensures compliance with organizational standards and policies.

8. **Increasing Credibility and Transparency:** Thorough metadata bolsters credibility and transparency, providing stakeholders with clear insights into data sources, methodologies, and quality for informed reliability assessments.

9. **Supporting Reproducibility and Verifiability:** Extensive metadata documentation fosters reproducibility and verifiability by detailing the origins, processing, and structural attributes of datasets.

10. **Advancing Efficiency and Cost-Effectiveness:** By organizing and simplifying access to datasets, metadata expedites information retrieval and analysis, contributing to operational efficiencies.

11. **Empowering AI and Automation Initiatives:** Structured metadata facilitates the efficient processing and utilization of data by AI systems and supports algorithm training through provision of clearly defined datasets.

12. **Preserving Longevity and Institutional Memory:** Metadata safeguards the long-term usability and value of datasets by documenting their context, structure, and purpose, thereby enabling future and legacy applications.

---

1. https://www.go-fair.org/fair-principles/ ↩
2. See Gordon Priest, 2010, The Struggle for Integration and Harmonization of Social Statistics in a Statistical Agency ↩

2.3 Metadata Standards

2.3.1 Definition

Metadata structures are composed of sets of elements that describe a dataset, often encompassing numerous attributes. Anticipating all the necessary aspects for thorough documentation or designing an optimally complete structure is a significant challenge for any single data curator. Consequently, collaboration among practitioners and subject matter experts is critical for defining metadata structures that address shared documentation requirements, leading to the creation of *metadata standards*.

A **metadata standard**, also referred to as a **metadata schema**, [1] is a systematically organized collection of clearly defined metadata elements intended for dataset documentation. It provides rules and guidelines to promote consistency and uniform implementation, and its adoption is supported by a community of practice. Metadata standards generally encompass descriptive, reference, and structural metadata components.

Each primary category of development data—including microdata, indicators, geographic datasets, and others—utilizes a corresponding metadata standard. This guide covers several data types: microdata; indicators or time series and their databases; geographic datasets; documents; images; video; audio recordings; [2] and scripts for data processing, editing, analysis, or other data transformations. Metadata standards have been developed independently for each type of data. In principle, common elements across different standards—such as those that denote the title or identifier of a resource—should be defined consistently. However, such consistency is not always achieved, given that some metadata standards have evolved independently.

Metadata standards also permit the use of custom metadata elements to meet the particular requirements and use cases of an organization. These extensions must be implemented in a manner that preserves full compliance with the established standard, typically through namespaces or similar mechanisms that clearly distinguish custom elements from standard elements. This strategy provides both flexibility and adaptability, while ensuring interoperability and adherence to recognized metadata specifications.

2.3.2 Benefits

The implementation of metadata standards represents an effective strategy for enhancing the completeness, usability, discoverability, and interoperability of metadata, thereby improving data accessibility and facilitating integration across various systems. Beyond the principal advantages of structured metadata discussed in section 2.2.3, metadata standards offer the following key benefits:

1. **Enhancing interoperability across data management and dissemination systems**. Structured metadata, particularly when coupled with Application Programming Interfaces (APIs), minimizes the risk of siloed operations within data management and dissemination systems. This approach enables automated metadata harvesting and seamless information exchange among disparate platforms, fostering connectivity between assets housed in different environments. Such integration advances data accessibility, consistency, and cohesion throughout the data ecosystem.

2. **Facilitating collaboration, partnerships, and knowledge sharing**. Metadata establishes a shared framework that promotes collaboration among teams and partner organizations by providing essential context for joint analyses and supporting the development of interoperable tools and systems. This common understanding strengthens the formation of an integrated and cooperative data community.

3. **Improving efficiency and cost-effectiveness in technical support.** The adoption of standardized metadata practices and tools enhances the efficiency and cost-effectiveness of technical support rendered to data curators. Standardization simplifies training, expedites knowledge transfer, and supports contributions from external organizations to technical assistance initiatives. It also fosters experience-sharing and collaborative engagement among institutions, thereby reinforcing capacity-building efforts.

---

1. A *metadata schema* is a structured framework that defines the specific elements, attributes, and rules for describing data or information. It provides a blueprint for how metadata should be organized and presented. A *metadata standard*, on the other hand, is a specific type of schema that is developed, maintained, and governed by an authoritative body. Metadata standards can be established at various levels, such as within a single organization, across a community of practice, or at an international level, depending on the scope and the governance arrangement that oversees its use and updates. In the context of this Guide, metadata schemas adopted by the Bank are considered as metadata standards. ↩️
2. Documents, images, videos, and audio recordings are considered as development data to the extent that they may contain development data, and/or serve as input to AI applications for development data. ↩️

2.4 Metadata Templates

2.4.1 Definition and Purpose

Metadata standards contain numerous components to accommodate a wide range of use cases, providing more detail than is necessary for the documentation of a particular dataset or for a given organization's needs. To improve usability and ensure contextual relevance, **metadata templates** are developed. These templates represent carefully selected subsets of elements from the broader standard, crafted for specific applications, while ensuring ongoing compliance with the original standard. Metadata templates allow for customization in several dimensions:

- ***Sub-setting***: Templates extract relevant metadata elements from the broader standard, enabling users to focus solely on those most pertinent to their specific context.
- ***Groupings***: Templates allow for the creation of custom clusters of metadata elements, facilitating the design of intuitive structures and entry tools that align with user preferences and operational requirements.
- ***Label***: Standard-provided labels may be replaced with organization-specific terminology in accordance with institutional lexicon or protocols.
- ***Status***: Although few elements are deemed *required* within the standards themselves, templates may designate additional elements as *mandatory* for quality assurance purposes. However, it is recommended to refrain from designating an excessive number as mandatory; elements vital but not essential should be classified as *recommended*.
- ***Description/Instructions***: Each metadata element is accompanied by a default description, which organizations may tailor for greater relevance and clarity. Customized descriptions and instructions support consistent and context-sensitive metadata creation and can be compiled into practical guidelines for data curators.
- ***Controlled vocabularies***: Certain elements can reference controlled vocabularies, which must conform to standardized code lists established by the organization.
- ***Default values***: Predefined default values can be assigned to select elements to streamline metadata entry.
- ***Validation rules***: Custom validation rules—such as regular expressions or value ranges—can be specified to enforce consistency and integrity across metadata records.
- ***Flexibility***: Templates may introduce additional metadata elements not found in the base standard, addressing unique organizational requirements.
- ***Translation***: Since labels, groupings, and instructions are editable, templates can be translated into any language, supporting multilingual environments.

These customization mechanisms enable the adaptation of standards to specialized use cases without compromising compliance.

Templates are created using metadata editor applications and stored as JSON files, facilitating easy distribution and reuse organization-wide, thereby promoting standardization and interoperability.

To maintain uniformity in metadata practices across the institution, template development must adhere to a structured governance process. Ad hoc modifications or template creation by individual data curators are discouraged. Instead, templates should be strategically designed and managed by an appointed administrator or the data documentation team.

While multiple templates per standard may exist to address diverse needs, limiting their overall number is advisable to foster consistency, reduce complexity, and ease ongoing maintenance.

2.4.2 Controlled Vocabularies

A **controlled vocabulary**, or **code list**, refers to a predefined, systematically organized collection of terms (with corresponding codes) consistently utilized to populate designated metadata elements. Examples include widely adopted national and international classifications, such as ISO 3166 for country codes and names, or the International Standard Industrial Classification (ISIC). Organization-specific vocabularies may also be employed, such as keyword taxonomies or custom code lists tailored for particular domains or purposes.

Controlled vocabularies are selectively applied to certain elements within a metadata standard to promote uniformity and precision. Their use ensures that identical concepts are consistently represented by the same term across multiple records, thereby minimizing ambiguity and improving searchability and interoperability. Controlled vocabularies are integral to effective data discovery. By assigning standardized terms to metadata elements, data catalogs can offer these terms as filtering options (facets), enabling users to refine their searches based on specific attributes or categories.

2.4.3 Development and Documentation of Metadata Templates

Metadata templates offer both flexibility and efficiency in designing metadata, helping to ensure organizational consistency when adhering to established metadata standards. To promote coherence, it is recommended that each data type be assigned a single standardized template; however, in certain specialized scenarios, multiple templates per data type may be justified, though these should be kept to a minimum to avoid fragmentation and inefficiency. At the Bank, the Office of the Chief Statistician, in collaboration with various departments, has developed a curated set of templates accessible to all staff via the Metadata Editor.[1] While specialized templates may be created for particular requirements, their use will be restricted to authorized user groups.

Templates are distinguished by essential attributes such as name, custodian, version, language, and description. Users developing or updating templates must provide a precise and informative name, version information (including date, version number, and a summary of modifications), language designation, and custodian details to guarantee comprehensive documentation and traceability.

Metadata templates are saved in JSON format to facilitate seamless sharing among teams or with external partners. Templates may be exported as PDF or HTML documents to serve as reference material for data curators.

2.4.4 Templates as SDMX Metadata and Data Structure Definitions

The metadata standards and templates used by the Bank for the documentation of indicators are designed to be compatible with SDMX-based applications.

The set of ***descriptive and reference metadata*** elements included in a template for documenting indicators, together with their respective code lists, aligns with the concept of a ***Metadata Structure Definition*** (MSD) as outlined in the Statistical Data and Metadata Exchange (SDMX) standard. An MSD establishes the organization and content of both

reference and descriptive metadata by specifying metadata concepts, their arrangement, the code lists used, and how they relate to the corresponding data. SDMX does not prescribe any pre-defined MSDs, allowing flexibility for customization.

Similarly, the list of ***structural metadata*** elements contained in a template, along with their associated code lists, corresponds to a ***Data Structure Definition*** (DSD) under the SDMX standard.

---

1. The Metadata Editor is a multi-standard metadata editor for development data that runs on the World Bank's intranet, enabling collaborative production and maintenance of metadata across multiple data types. Developed by the World Bank Group and released as open source, it is intended for other organizations to install on their own intranets to promote wider adoption of metadata standards. The application is provided with a set of recommended templates. ↩️

2.5 Choosing the Right Level of Metadata Production

Metadata can be created at various levels depending on the dataset's nature and its intended use.

- For **microdata** like surveys or census data, metadata is generated at the dataset level. It covers details about each specific survey or census, including files, variables, and methods. When surveys are part of a series, such as monthly labor force reports, each dataset has separate metadata that reflects both shared series elements and unique dataset features.

- **Geographic datasets** require metadata that may account for multiple layers or images within a dataset. The unit of data being documented is the dataset, but metadata specific to each data file included in the dataset can be generated.

    - For ***vector data*** metadata is compiled at the dataset level using the ISO 19115-1 standard, complemented with information at the data file level, compiled using the ISO 19110 feature catalog standard.

    - For ***raster data***, where a dataset may contain multiple images (e.g., different time periods, spectral bands, or tiles), the metadata is typically generated at the dataset level using the ISO 19115-2 metadata standard.

    - Some datasets may include both ***vector and raster data***. Both types will be documented together but distinguished in the same record, using respectively the ISO 19115-1 and ISO 19115-2, with ISO 19110 supporting the documentation of vector features.

- For **indicators**, metadata granularity becomes more complex. Indicators often reside within larger databases and have several time series. For example, the Bank's World Development Indicators (WDI) database includes over 1,400 indicators for approximately 200 countries, resulting in more than 280,000 time series. Metadata can be available at several levels:

- ***Database level***: Summarizes the entire database, outlining its scope, purpose, and key attributes.

- ***Indicator level***: Defines methodologies, sources, and definitions applicable across all countries and periods.

- ***Time series level***: Adds context for a country or region within an indicator, including any adjustments specific to that time series.

- ***Value level***: Contains notes for individual data points, detailing sources, collection methods, or caveats.

> The primary emphasis is placed on indicator-level and dataset-level metadata. Time series or period-specific notes can be included within the indicator metadata. Metadata related to observation values should be recorded directly as annotations within the dataset.

- **Data processing and analysis scripts** have metadata created at the research project or process workflow level. Research projects document script roles, objectives, authors, dates, and context; data processing workflows group related scripts under one metadata record, describing the overall process and technical dependencies.

- **Documents**, such as papers, books, or reports, are generally documented individually. Multi-volume works may have collective metadata, but the recommended practice is to document each volume separately, especially if there are differences in authorship, publication date, or content.

- **Images** receive individual metadata records covering content, creator, date, format, and technical details. All versions of an image count as a single item, with one metadata file. Albums or collections to which an image belongs are referenced in each image's metadata.

- **Videos** are documented individually, with metadata for title, creator, creation date, duration, format, and resolution. Multiple formats or resolutions are logged as versions within a same metadata record. Series or collections to which a video belongs are referenced in the individual video metadata.

- **Audio recordings** get metadata either as single files or as part of a group or collection. When documented as a group, each recording is documented within the collection metadata. Different formats or resolutions of a same recording are noted as versions of the same audio file within the same metadata record.

2.6 Generating Metadata Compliant with Standards

Generating metadata in accordance with the Bank's metadata standards requires the creation of a structured, machine-readable file based on a prescribed set of metadata elements. Metadata is maintained in JSON format, with the flexibility to export to XML, CSV, or additional formats as necessary.

The Bank's metadata standards accommodate multiple implementation methods. A dedicated metadata editor application is available for the creation, management, and storage of metadata, featuring full API support to facilitate programmatic import and export where automation enhances efficiency. Additionally, metadata can be produced through other specialized tools, provided that compliance with established standards is maintained. The ensuing sections detail these three approaches to metadata production.

2.6.1 Generating Metadata Using a Specialized Metadata Editor

The Bank utilizes a dedicated metadata editor application, which is compatible with all metadata standards outlined in this Guide. [1] Accessible to all Bank staff members via the Bank's intranet, the application features an extensive permission and role-management system to facilitate collaborative workflows and robust metadata governance.

This purpose-built solution provides comprehensive capabilities for the creation and management of metadata related to development data assets:

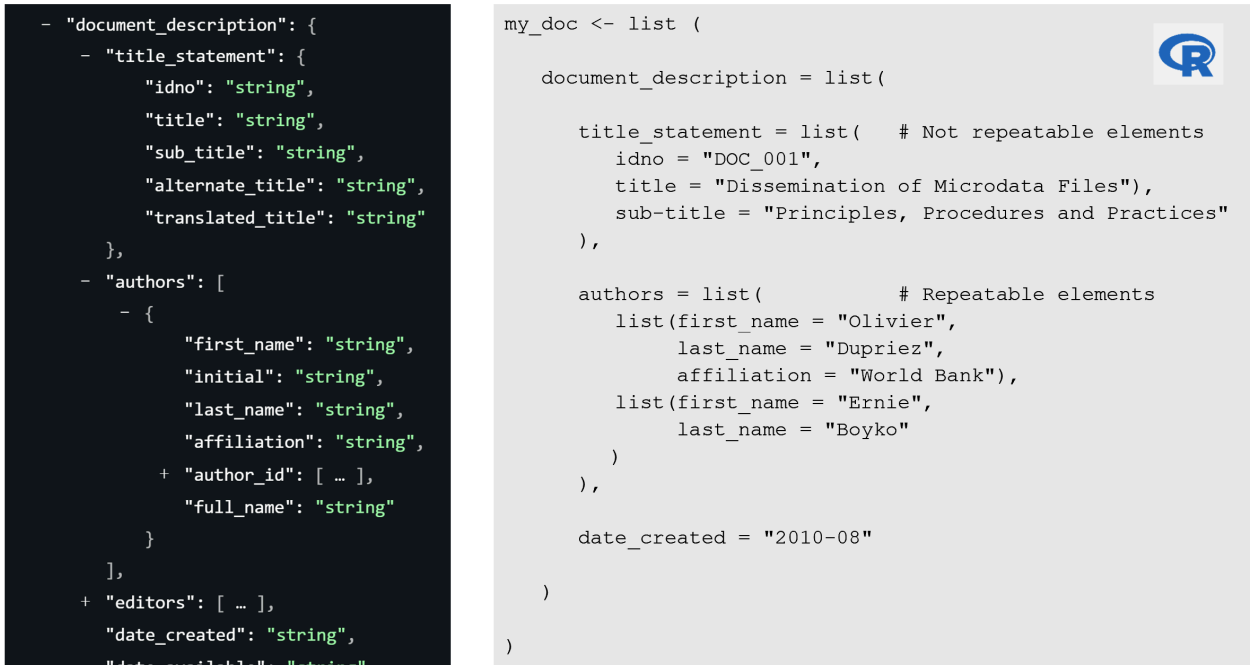
- **Multi-standard support** – Delivers a unified interface for constructing metadata for development data aligned with all metadata standards utilized by the Bank.
- **Automated metadata extraction** – Facilitates dataset importation and automates the creation of essential metadata components, including variable-level data dictionaries for microdata, or feature characteristics for geographic vector datasets.
- **Template management** – Enables the development and maintenance of standardized templates customized to specific data and documentation requirements.
- **API-driven architecture** – Supports automated metadata generation, bulk modifications, and seamless integration with analytical and cataloging workflows via programmatic access (for example, through R or Python libraries).
- **Flexible export options** – Accommodates various export formats (such as JSON, XML, and CSV) to enhance interoperability with external tools and systems.
- **Role-based access control** – Implements user roles (viewer, editor, administrator) to effectively manage contributions, uphold quality assurance protocols, and strengthen metadata governance.
- **AI-readiness** – Designed to interface with external artificial intelligence tools for metadata quality assessment, enrichment, and automation processes.
- **System compatibility** – Operates efficiently with other Bank applications, including those supporting data management, processing, and cataloging, thereby fostering an integrated and efficient data ecosystem.

2.6.2 Generating Metadata Programmatically

Leveraging R and Python for metadata management streamlines workflows, offering both efficiency and scalability while adhering to metadata standards. By employing programmatic methods, users can batch process and automate tasks, which reduces manual work and enhances precision. This is particularly useful for datasets with frequently changing metadata. Additionally, these tools facilitate regular quality checks, diagnostics, and the integration of AI-driven features to boost completeness and consistency. Secure API keys are used to manage access, defining permissions for each user or script. Metadata that aligns with Bank standards can then be uploaded to compatible applications through an API. [2]

Using R

In R, metadata is generated by constructing a *list* object comprising nested lists that correspond to the groups of metadata elements defined in the standard. The organization of elements and their associated data types must adhere strictly to the schema specifications set forth in the Bank's metadata standards. Within the description of standards, curly brackets { } signify that a group of elements must be represented as a list, while square brackets [ ] indicate repeatable elements which are implemented as lists of lists. Figure 1 uses the DOCUMENT metadata standard to illustrate the mapping between schema notation and the corresponding R list structure.



```
- "document_description": {
    - "title_statement": {
          "idno": "string",
          "title": "string",
          "sub_title": "string",
          "alternate_title": "string",
          "translated_title": "string"
      },
    - "authors": [
        - {
              "first_name": "string",
              "initial": "string",
              "last_name": "string",
              "affiliation": "string",
            + "author_id": [ … ],
              "full_name": "string"
          }
      ],
    + "editors": [ … ],
      "date_created": "string",
```

```
my_doc <- list (

   document_description = list(

      title_statement = list(   # Not repeatable elements
         idno = "DOC_001",
         title = "Dissemination of Microdata Files"),
         sub-title = "Principles, Procedures and Practices"
      ),

      authors = list(           # Repeatable elements
         list(first_name = "Olivier",
              last_name = "Dupriez",
              affiliation = "World Bank"),
         list(first_name = "Ernie",
              last_name = "Boyko"
         )
      ),

      date_created = "2010-08"

   )

)
```

Figure 1 – Documentation of the JSON metadata standard (left) and its correspondence to an R list of lists

Using Python

In Python, metadata is organized as dictionaries, with nested dictionaries and lists reflecting the metadata standard's hierarchy. Non-repeatable elements use dictionaries, while repeatable ones are stored as lists of dictionaries. In the documentation of standards, { } denotes a dictionary and [ ] indicates repeatable items as lists of dictionaries. Figure 2 provides the Python equivalent of the example in Figure 1.

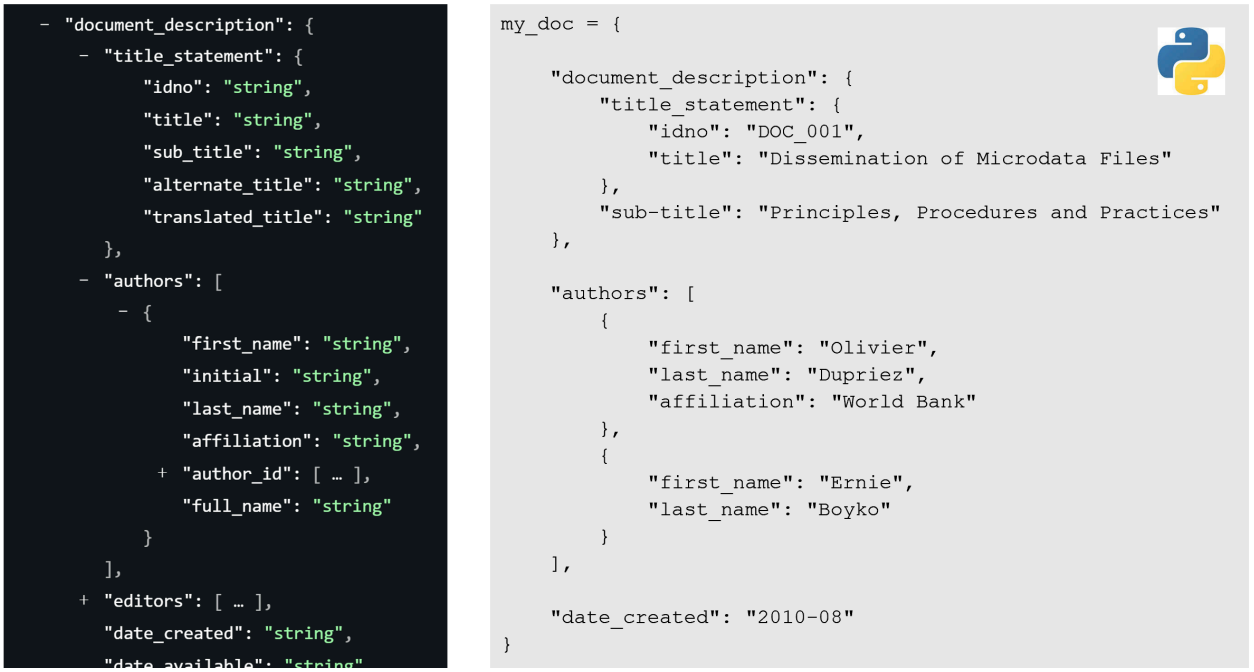


Figure 2 - Documentation of the JSON metadata schema (left) and its correspondence to a Python dictionary

2.6.3 Metadata Generated by Other Software Applications

Metadata standards may be integrated within data management and processing platforms. For instance, Survey Solutions, developed by the Bank, and CSPro, a public-domain software from the U.S. Census Bureau, are capable of exporting data dictionaries using the DDI Codebook format. [3] Similarly, ArcGIS Pro (ESRI) and the open-source QGIS facilitate the creation of metadata adhering to ISO 19115 and ISO 19139 requirements for geographic datasets. [4]

Although these applications and comparable tools can generate metadata that is broadly consistent with the Bank's standards, their outputs are generally limited to structural metadata—including files, variables, and attributes. Critical contextual information, such as methodological details, data collection processes, quality indicators, and usage notes, is typically omitted.

2.6.4 Using Artificial Intelligence to Generate and Enhance Metadata

Artificial Intelligence (AI), particularly through advanced large language models (LLMs), offers significant capabilities for both the creation and enhancement of metadata.

For **new metadata generation**, AI technologies can extract pertinent information from datasets, documents, survey instruments, scripts, and other digital assets. These models convert unstructured data into structured metadata elements that adhere to recognized standards, such as generating abstracts, defining geographic coverage, or summarizing research methodology. The Office of the WBG Chief Statistician is actively engaged in developing and implementing such AI-powered solutions, aiming to facilitate scalable and high-quality metadata creation throughout the Bank's data ecosystem.

Additionally, AI can be leveraged to **enhance human-curated metadata**. AI-driven tools enable the enrichment of descriptions, the completion of minor gaps, the improvement of keyword accuracy, and the automation of quality control

processes. Examples of metadata augmentation include:

- **Generating keywords for datasets**: LLMs can analyze abstracts, definitions, and dataset descriptions to propose relevant keywords, thereby improving searchability and indexing.
- **Extracting and classifying topics**: AI can identify key topics within datasets, either by freely generating new topics for metadata or tagging content according to established classification systems.
- **Creating and refining indicator definitions**: AI can generate or improve indicator definitions, ensuring alignment with standardized terminology.
- **Populating the "relevance" metadata element**: For indicators, AI can suggest succinct relevance statements to help users appreciate the significance and applicability of the data, thereby enhancing semantic search capabilities.
- **Grouping variables by themes**: For microdata, AI can classify variables into thematic groups and composes descriptive summaries of variable groupings.
- **Automating translations**: AI-based translation solutions can facilitate the production of metadata in multiple languages, advancing accessibility for global audiences.
- **Extraction of descriptive content from digital images**: AI methods can examine digital images and automatically derive descriptive details—including objects, individuals, scenes, or textual elements—enabling richer metadata.
- **Generation of transcriptions and quality metrics for audio and video files**: AI systems can efficiently extract technical and descriptive metadata from media files, capturing aspects such as language, speaker or scene identification, quality metrics, and full transcripts.

While AI technologies substantially advance both metadata generation and enhancement, human supervision remains indispensable. These AI-driven tools should serve as assistants, expediting documentation and improving metadata completeness, whereas data curators and subject matter experts are responsible for validating the accuracy and relevance of all AI-generated outputs.

---

1. The application was released as an open-source software application under an MIT license. See https://worldbank.github.io/metadata-editor-docs/ and https://github.com/worldbank/metadata-editor ↩
2. Programmers can refer to the documentation of the standards available at https://worldbank.github.io/metadata-schemas/. ↩
3. See https://mysurvey.solutions/en/ and https://www.census.gov/data/software/cspro.html ↩
4. See https://pro.arcgis.com/en/pro-app/latest/help/metadata/create-iso-19115-and-iso-19139-metadata.htm ↩

2.7 Metadata Quality Control

Evaluating metadata quality is crucial to guarantee that metadata effectively serves its various purposes: enhancing data usability, interpretation, discovery, and accessibility for humans, machines, and AI systems. Achieving these objectives requires systematically reviewing and optimizing metadata for completeness, accuracy, consistency, and suitability for intended use. Structured metadata compliant with standards can be evaluated using a combination of methodical inspections and intelligent analyses, facilitating the identification of opportunities for improvement. A comprehensive assessment of metadata quality should incorporate both **rule-based checks** and **advanced qualitative evaluations**.

2.7.1 Rule-based Checks and Schema Validation

Rule-based checks help ensure data quality by confirming that all necessary fields are present, identifying missing or inconsistent information, and checking that data formats are correct. This involves **schema validation**—making sure the metadata follows the structure and rules established in the metadata standard—and **template validation**, which verifies that any additional and specific rules defined in the organization's template are followed.

2.7.2 Advanced Qualitative Evaluations Using AI

Advanced qualitative evaluations employ artificial intelligence solutions to systematically assess metadata content and style in terms of completeness, coherence, readability, semantic alignment, and grammatical accuracy. In practical application, this entails the design, validation, and deployment of agentic AI tools for comprehensive metadata analysis. The Office of the WBG Chief Statistician oversees the establishment of metadata quality standards and directs the development, upkeep, and implementation of AI-based metadata quality assessment platforms, with the objective of improving consistency and usability within the Bank's development data ecosystem.

2.8 Metadata Governance

2.8.1 Information on Metadata

Metadata must consistently correspond with the associated dataset; any modifications to the data should be promptly reflected in the metadata to ensure ongoing accuracy and consistency. Consequently, it is common practice to maintain multiple versions of metadata.

The metadata standards used by the Bank for documenting development data include a dedicated section for documenting metadata itself (Figure 3). This section is labeled "Document Description" in the DDI Codebook metadata standard and as "Metadata Information" in other metadata standards. It contains key elements to describe metadata versions, such as an identifier, details of the metadata producer(s), production date, and a version statement. The version statement should include both a version description and date, ideally stored in ISO 8601 format (YYYY-MM-DD HH:MM:SS.sss). It is recommended to include a note in the version statement to describe changes compared to previous iterations.


```
- "metadata_information": {
      "title": "string",
      "idno": "string",
    - "producers": [
        - {
              "name": "string",
              "abbr": "string",
              "affiliation": "string",
              "role": "string"
          }
      ],
      "prod_date": "string",
    - "version_statement": {
          "version": "string",
          "version_date": "string",
          "version_resp": "string",
          "version_notes": "string"
      }
  },
```


Figure 3 - Metadata information and versioning

2.8.2 Recording the Provenance of Metadata for Traceability

Metadata may be generated within the Bank or obtained from external sources, such as metadata catalogs managed by other organizations. When using external metadata, it is important to document provenance details for traceability and transparency. The Bank's metadata standards feature a specific *provenance* section intended to record both the source of imported metadata and its original repository.

2.8.3 Locking and Versioning Metadata

The Bank's metadata editor incorporates a versioning mechanism that enables users with reviewer privileges to lock project-specific metadata. Upon review and endorsement, locking renders the metadata read-only, thereby safeguarding it against unauthorized modifications.

When metadata is locked:

- The system assigns a **version number** consistent with semantic versioning (distinguishing major, minor, or patch updates).
- The reviewer supplies a **concise summary** of significant changes introduced in the new release (applicable to versions subsequent to 1.0.0).
- The system logs both the **reviewer's credentials** and the **timestamp** associated with the lock action.

Subsequent to locking, the metadata editor automatically generates a new editable project copy, which allows continued enhancements while preserving the integrity of the locked version. All version information is documented within the *version_statement* section of the metadata, and the editor maintains and displays a comprehensive version history, ensuring transparency and accountability.

To guarantee metadata integrity, the editor generates a checksum—a unique string calculated using a cryptographic hash algorithm. Checksums provide:

- **Integrity assurance** – Any inadvertent alterations or corruption (e.g., from storage or transmission errors) result in a changed checksum, promptly identifying unauthorized changes.
- **Automated monitoring** – Even minimal updates generate a distinct checksum, facilitating automatic comparison without the need for manual intervention.
- **Workflow automation** – Checksums support critical downstream operations in the Bank's data management and dissemination frameworks by:
- Flagging metadata modifications.
- Enabling reversion to earlier versions should inconsistencies arise.
- Synchronizing metadata repositories efficiently.

This approach ensures that metadata versioning remains comprehensive, transparent, and auditable, effectively supporting internal governance requirements and operational processes.

2.9 Metadata Dissemination and Use

Effective metadata dissemination makes development datasets discoverable, accessible, and usable for stakeholders like researchers, analysts, policy makers, and automated systems. To reach both humans and machines, metadata should be available in multiple formats via various channels.

- **Human users** benefit from easily readable formats such as HTML, PDF, and CSV for browsing, reporting, and offline access.
- **Machines** require structured formats like JSON, XML, and RDF for automated processing, integration with catalogs, and interoperability across tools.

Dissemination channels should maximize reach and usability:

- **Application Programming Interfaces (APIs)** – Enable dynamic queries and automated integration with other platforms.
- **Web interfaces and online catalogs** – Offer interactive portals for exploration.
- **Downloadable reports** – Provide metadata in formats like PDF or CSV for offline use.
- **Downloadable structured files** – Supply data in JSON, XML, or RDF for programmatic access.
- **Embedded metadata** – Use standards such as DCAT or schema.org to enhance discoverability and automated harvesting.

To meet varied user needs, the Bank seeks to release metadata openly, balancing transparency with required privacy and access controls. Metadata should be clearly licensed for reuse, and any restrictions must be specified so users understand limitations.

2.10 Archiving and Preserving Metadata

Metadata archiving ensures long-term access, reliability, and usability of the Bank's development data. Systematic preservation maintains historical records, supports provenance tracking, and enables future data reuse for research and policymaking.

Metadata archiving includes:

- **Indefinite preservation** – Descriptive, reference, and structural metadata are archived permanently, with all versions kept unless disposal is required by ethics or contracts.
- **Backup-only preservation** – Administrative, portrayal/presentation, and impact metadata are backed up regularly but not archived long-term.

To maintain long-term accessibility and interoperability, archived metadata is preserved in open, non-proprietary formats such as JSON. This strategy mitigates the risk of obsolescence and facilitates compatibility with emerging technologies.

Integrity and security measures for metadata archiving align with the information security and data protection policies and procedures already in effect at the Bank. These measures include:

- **Regular audits** – Routine checks are performed to verify compliance with archival policies and detect any corruption or loss of metadata.
- **Redundant storage** – Geographically distributed backup systems protect against hardware failures, cyber threats, and other unforeseen disasters.
- **Automation** – Archiving workflows are automated to minimize human error and maintain consistency in metadata preservation.

3. World Bank Group Metadata Standards

This chapter outlines the metadata standards adopted by the Bank for the documentation of development data. These standards are subject to periodic review and may be supplemented with additional standards as necessary.

Chapter outline

- 3.1 Selection, Development, Maintenance, and Versioning
- 3.2 Maintaining Flexibility in the Use of Metadata Standards
- 3.3 Technical Documentation
- 3.4 Elements Common Across Standards
- 3.5 World Bank Group Metadata Standards by Data Type
- 3.6 Administrative Metadata
- 3.7 Relationship Metadata
- 3.8 Compatibility with Other Metadata Standards

3.1 Selection, Development, Maintenance, and Versioning

In line with the Development Data Quality Policy, the WBG Chief Statistician is responsible for choosing, managing, and updating the standards and templates used to document development data. An advisory team provides guidance to the Chief Statistician regarding the selection and maintenance of metadata standards.

Whenever possible, the Bank adopts internationally accepted metadata standards, modifying them as needed to fit its operational and strategic needs. If these standards do not fully meet the Bank's requirements, they are further adjusted or, when necessary, new metadata standards are created in close cooperation with both relevant business units and external partners. Any customized standards go through thorough evaluation to ensure they match global best practices in data documentation and compatibility.

To guarantee consistency and machine-readability, all metadata standards at the Bank are implemented using JSON schemas. The Bank uses the following standards to document development data:

Data type	Standards	Use
Microdata	DDI Codebook version 2.5 (or latest) complemented by a set of Bank-specific metadata elements.	DDI is used for documenting microdata—covering surveys (households, individuals, facilities, enterprises), censuses, administrative records, transactions, and sensor-captured data. It underpins the World Bank Microdata Library (internal and external).
Indicators	WBG bespoke standard. The standard was created by consolidating metadata elements found in multiple WB and external data management and dissemination systems.	The standard is used to document statistical indicators—capturing definition, units, methodology/computation, coverage and disaggregation, source/provenance, quality notes, and revision history. It underpins multiple indicator database and dissemination systems including Data360.
Database (dataset) of indicators	WBG bespoke standard.	The standard is used to document datasets that house indicators (e.g., the World Development Indicators database), providing metadata about scope and purpose, thematic coverage, geographic and temporal extent, and others. It complements the Indicator Metadata standard by contextualizing indicator collections.
Geographic data and services	ISO 19115, 19110, 19119 series, complemented by a set of Bank-specific metadata elements.	The ISO 19100 suite is used to document raster and vector geographic datasets, and geographic data services.
Documents	WBG bespoke standard, based on metadata	This standard is used to document documents of any type (publications, reports, books, manuals,

Data type	Standards	Use
	elements extracted from the Dublin Core, MARC21, and BibTex metadata standards, complemented by a set of Bank-specific metadata elements.	and others) related to development data. This includes documents that form corpora intended to be used as input for training AI models. Note that another standard is used to document documents used as external resources associated with another data type.
Images	Two options are supported: DCMI (light option) and IPTC (for more advanced use), complemented by a set of Bank-specific metadata elements.	The standards are used to document digital images. These can be image files used as input for training AI models related to development data, or photos related to statistical events.
Video	WBG bespoke standard, based on metadata elements extracted from the Dublin Core and schema.org (videoObject) standards, complemented by a set of Bank-specific metadata elements.	The standard is used to document videos, typically videos used as input for training AI models, or whose topic is directly related to development data.
Audio recordings	WBG bespoke standard, based on metadata elements extracted from the Dublin Core, schema.org (audioObject), and PBCore metadata standards, complemented by a set of Bank-specific metadata elements.	The standard is used to document audio recordings, typically audio files used as input for training AI models, or whose topic is directly related to development data.
Data processing and analysis scripts	WBG bespoke standard.	The standard is used to document data processing or analysis scripts, as well as machine learning and statistical models. The purpose is to support reproducibility and replicability of research work, and verifiability of data processing and analysis. This standard enables the WB reproducibility catalog (https://reproducibility.worldbank.org/).
External resources	WBG bespoke standard, based on the Dublin Core standard.	This standard is a complement to each of the above-mentioned standards. It is used to document any digital resource associated with one of the above data types; this could be used for example to document survey questionnaires

Data type	Standards	Use
		or reports, attach videos or images to datasets, etc.

The standards also have spaces for additional elements, which can be added through templates (see section 2.4). The Bank has created at least one recommended template for each metadata standard.

If there are important metadata elements not already included in current standards, they should be added to future versions of the relevant standards. These updates are made collaboratively, with input from the Office of the Chief Statistician and external organizations like the DDI Alliance as relevant, so that consistency and broad applicability are maintained. Creating custom metadata elements outside official standards should only occur in rare cases where existing options do not meet specific needs.

Because metadata standards evolve over time, it is essential to record the specific version used for documenting a dataset. This information on version must be explicitly included in the metadata header. To ensure transparency and interoperability, all metadata standard versions used by the Bank must be maintained in an openly accessible registry with a complete version history. Providing machine-readable access to these versions allows applications to accurately interpret metadata structures. Linking metadata to its corresponding standard ensures that its format and meaning remain clear and consistent over time.

The metadata standards adopted by the Bank are used to generate and maintain cataloging, referential, and structural metadata. No predefined standard is imposed for administrative metadata, usage and impact metadata, or display metadata. Instead, custom schemas are developed on a case-by-case basis, tailored to the specific requirements of the applications that use them (see section 2.6).

Administrative metadata	**Structure and content defined based on needs.**	**Optional; used to provide dissemination systems with instructions related to access, metadata display, etc.**
Usage and impact metadata	Structure and content defined based on needs.	Optional; used to store information (dynamically generated) on web usage and impact of the datasets.
Presentation and portrayal metadata	Structure and content defined based on needs, except for geographic data for which the “portrayal” metadata elements of the ISO19139 are used.	Optional; used to provide dissemination systems with instructions on how to display the data (e.g., the type of chart, color palette, font sizes, axis labels, etc. for an indicator).

3.2 Maintaining Flexibility in the Use of Metadata Standards

Metadata should follow established standards for consistency and usability, while maintaining flexibility for specific use cases. All Bank metadata standards provide an "Additional" section without predefined content, allowing users to add custom elements through templates or programmatically as required. While this supports unique requirements, unnecessary customization is discouraged. If certain custom elements become commonly used, the Office of the Chief Statistician will consider their inclusion in future versions of the standards.

3.3 Technical Documentation

The technical documentation for the metadata standards selected by the World Bank Group Chief Statistician is available at https://worldbank.github.io/metadata-schemas, where comprehensive structures and detailed information on each metadata element can be found (Figure 4).

The metadata standards incorporated within this system are delivered in a JSON format specifically designed by the Bank to advance effective data curation processes. Although foundational standards such as the DDI Codebook and the ISO 19100 set are conventionally defined in XML, the Bank customizes these descriptions to better reflect its internal procedures and operational requirements. Interoperability remains a priority; metadata developed according to the Bank's standards and templates can be exported in formats that are internationally recognized, with only negligible information loss where global specifications do not yet fully support the Bank's advanced elements.

The Bank also documents controlled vocabularies, or standard code lists, to support its metadata standards. These lists are kept up to date and are publicly accessible through the Bank's Fusion Metadata Registry. [1]

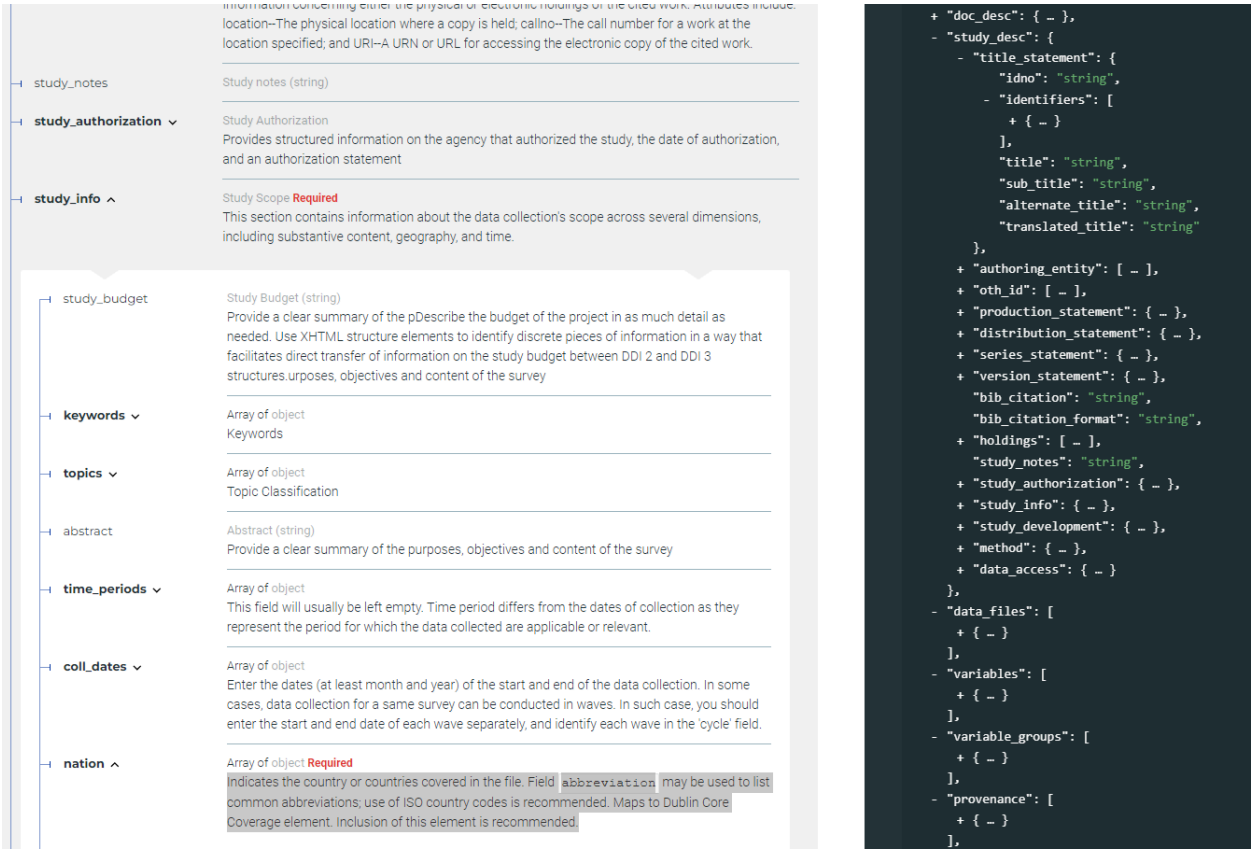


Figure 4 – Online documentation of the metadata standards adopted by the Bank (showing an abstract of DDI)

1. See https://fmr.worldbank.org/FMR/items/codelist.html ↩

3.4 Elements Common Across Standards

A designated metadata standard is implemented for each primary data type—microdata, indicators and databases, geographic datasets, documents, images, videos, audio, and scripts—to ensure the capture of the most pertinent information. In order to promote consistency and interoperability, all standards pertaining to development data incorporate a common set of fundamental metadata elements. In instances where these essential elements are missing from the international metadata standards adopted by the Bank, they have been incorporated by the Bank into its standards. These additions pertain to metadata provenance, tags, and version control.

3.4.1 Information on Standard Used

Metadata files (JSON or XML) produced by the Bank for development data are required to include a header section specifying the metadata standard and its version. This practice enables automated systems to accurately parse, validate, and process the metadata. A clearly defined header enhances usability, consistency, and interoperability of metadata. By explicitly indicating the standard, the header enables systems to:

- ***Interpret structure and semantics***: Machines can reliably understand metadata elements, their relationships, and the expected data types in accordance with the defined standard.
- ***Ensure compatibility and interoperability***: Knowledge of the standard and version facilitates seamless integration with other platforms and tools adhering to the same standards.
- ***Enable automated processing and validation***: Systems can assess compliance with the specified standard, confirming that metadata adheres to prescribed formats and rules.
- ***Facilitate version control and updates***: Explicit versioning in the header allows for smooth transitions when metadata standards evolve, preventing compatibility issues and ensuring metadata integrity over time.

3.4.2 Unique Dataset Identifiers

Every data resource—be it a micro-dataset, an indicator, a geographic dataset, or another type—needs a persistent and unique identifier to set it apart from others. All metadata standards include a mandatory element, *idno*, used to capture the unique internal identifier for each documented resource. This is vital both for accurately cataloguing datasets within the Bank's repositories and for allowing users to cite them properly.

Beyond the *idno* identifier, the Bank's metadata standards allow for capturing various other identifiers, such as Digital Object Identifiers (DOIs) and original identifiers from external sources. Keeping these original identifiers helps maintain data traceability and verify its origins.

3.4.3 Version Statement for the Data

Datasets are subject to periodic updates or revisions, making effective data versioning essential. [1] All metadata standards implemented by the Bank contain components to record the dataset version (Figure 5).

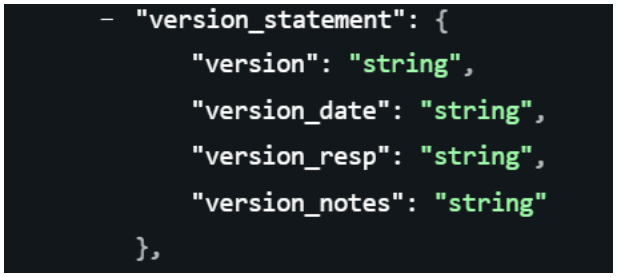


Figure 5 - Versioning data

Semantic versioning is recommended for datasets to clearly communicate changes. It uses a three-part format: MAJOR.MINOR.PATCH (e.g., 1.4.2).

- **MAJOR**: Indicates breaking changes that are incompatible with previous versions. For datasets, this may involve:
  - Significant changes to the data file structure (e.g., new variables added, variable names changed, or format alterations).
  - Removal of key data elements.
  - A shift in methodology or classification that affects continuity.
- **MINOR**: Denotes backward-compatible enhancements, such as:
  - Adding new rows, columns, variables, or fields that do not disrupt existing workflows.
  - Extending the dataset with new metadata.
  - Incorporating data for additional periods or geographies.
- **PATCH**: Reflects backward-compatible fixes, including:
  - Corrections to data errors (e.g., fixing typos or recalculating values).
  - Minor updates to metadata that clarify existing information without altering its meaning.

Workflow example:

- ***Initial release***: A dataset released as 1.0.0 signifies the first stable version.
- ***Patch updates***: A typo is fixed in some records without changing the structure. The version becomes 1.0.1.
- ***Minor updates***: A new variable (or column) is added to the dataset. The version updates to 1.1.0.
- ***Major update***: The dataset's structure is redesigned, and some variables are removed. The version updates to 2.0.0.

The benefits of semantic versioning are:

- **Clarity**: Users can easily understand the scope of changes.
- **Predictability**: Clear guidelines ensure consistent communication with dataset users.
- **Compatibility management**: Data users can evaluate whether their workflows require adaptation based on the change of the version.

When applying semantic versioning, it is recommended to provide thorough documentation for all updates. This process should involve explicitly stating the version number in the dataset's metadata (*version*), recording the version date in ISO 8601 format (*version_date*), and maintaining an exhaustive version history within the documentation (*version_notes*). The

version notes ought to present a detailed summary of each change, specifying both the date and a description of the modifications made.

3.4.4 Metadata Information and Version Statement

It is essential to document both the version of the data and the associated metadata. All metadata standards utilized by the Bank include a designated section for recording details of metadata elements. For instance, the DDI Codebook standard labels this section as "document description," whereas other standards may refer to it as "metadata information" (Figure 6). Additionally, the implementation of semantic versioning for metadata is strongly recommended.


```
"metadata_information": {
    "title": "string",
    "idno": "string",
    "producers": [
        {
            "name": "string",
            "abbr": "string",
            "affiliation": "string",
            "role": "string"
        }
    ],
    "prod_date": "string",
    "version_statement": {
        "version": "string",
        "version_date": "string",
        "version_resp": "string",
        "version_notes": "string"
    }
},
```


Figure 6 - Metadata on metadata in a Bank metadata standard

3.4.5 DataCite Elements

All metadata standards implemented by the Bank include a *datacite* section (Figure 7). This section contains metadata elements necessary for the automated generation of Digital Object Identifiers (DOIs) using DataCite.[2] DataCite, a DOI and metadata management service to which the Bank subscribes, facilitates the registration and management of DOIs for its datasets. The Development Data Group within the Bank oversees this service. Through DataCite, the Bank assigns DOIs to its data products, maintains accurate metadata in accordance with FAIR principles, and ensures persistent links for long-term accessibility and citation.[3] Additionally, DataCite provides an API that enables programmatic DOI creation, thereby increasing the flexibility and scalability of DOI management within the Bank*’s data* workflows. The datacite metadata elements are structured for compatibility with this API.



Figure 7 - Metadata elements for DataCite (for creation of DOIs)

3.4.6 Provenance of the Data Resource

All metadata standards incorporate a *provenance* section (Figure 8), which serves to document the source of both the data and its associated metadata, especially when collected programmatically from external platforms. The elements within this section are designed to provide transparency and enhance the traceability of the data's lineage.



Figure 8 - Metadata on data and metadata provenance

3.4.7 Tags

All metadata standards include a "tags" section (see Figure 9). Tags are user-defined textual elements that may be grouped to support functions such as facilitating filters within data catalogs.


```
- "tags": [
    - {
        "tag": "string",
        "tag_group": "string"
      }
  ],
```


Figure 9 - Tags and tag groups

For instance, a data catalog may encompass both open-access resources and those that require payment. To facilitate user navigation, the catalog interface should provide a filter (facet) enabling users to select datasets according to their pricing model (see Figure 10). Although the Bank's metadata standards presently lack a dedicated element for indicating resource cost status, this information can be efficiently documented through a user-defined tag group, such as "pricing," utilizing tags like "Free" or "For a Fee."

Employing controlled vocabularies for tags and tag groups is often beneficial, as it ensures consistency, reduces ambiguity, and improves searchability across the catalog. These structured tags can then be leveraged to create intuitive facets in the data catalog, enhancing the user experience and supporting more precise data discovery.



Figure 10 - Use of tags and tag groups for faceting in a data catalog

1. The instructions provided here on versioning data do not apply to data that are updated on a continuous basis. For such data, versioning applies only when a derived dataset is generated at a fixed time.
2. https://datacite.org/create-dois/

3. The Bank does not assign DOIs to third party data products. ↩️

3.5 World Bank Group Metadata Standards by Data Type

This section outlines the metadata standards chosen by the WBG Chief Statistician and formally implemented by the Bank for documenting development data and associated resources. For a detailed description of each standard, consult the online documentation of the standards.[1] For guidance on their use and a description of the elements they contain, refer to the Metadata Editor User Guide,[2] which also serves as a reference guide for data curators.

3.5.1 Microdata

Microdata refers to electronic datasets that contain detailed observations at the unit level, including individuals, households, facilities, or enterprises. These datasets are commonly presented in tabular format, with rows corresponding to observational units and columns representing variables. A micro-dataset may comprise multiple related files, which can exhibit hierarchical relationships (such as a household file linked to individual-level files). Microdata is typically sourced from surveys, censuses, administrative records, transactional databases, or sensors.

The Bank has adopted the Data Documentation Initiative (DDI) metadata standard—created and maintained by the DDI Alliance under the GNU General Public License—for microdata documentation.[3] The DDI standard is extensively implemented by data archives, national statistical authorities, and international organizations to facilitate systematic preservation, distribution, and reuse of microdata.

The DDI standard has two main branches: LifeCycle and Codebook. The Bank uses DDI Codebook (version 2.5 or above) to manage, present, exchange, and preserve metadata in social and behavioral sciences. Metadata is organized into five sections:

- **Document Description** – Metadata about the documentation itself, like creation date, version, authorship, and cataloging information.
- **Study Description** – Information about the study, including title, producer or principal investigator, objectives, data version, sampling design, geographic and temporal coverage, and thematic topics.
- **Data Files Description** – A concise description of each data file included in the dataset.
- **Variables Description** – Detailed information on each variable, such as names and labels, value categories, formats and types, interviewer notes, and summary statistics.
- **Variables Groups** – Virtual organization of variables into thematic or other groups. This is optional and not systematically implemented, but adds value for data discovery.

To improve usability and consistency throughout the Bank's metadata systems, several cross-cutting sections have been added to complement the standard, such as "Additional," "Provenance," "Tags," "DataCite," and "Administrative metadata."

File- and variable-level sections create a data dictionary, which is essential structural metadata for microdata assets. Building these manually is time-consuming, but the Bank's metadata editor simplifies the process by importing datasets from formats like Stata or SPSS and automatically extracting file names, variable details, labels, and summary statistics from the data files. Curators can then add more information, such as question text or skip patterns.

3.5.2 Indicators and Databases of Indicators

Two distinct metadata standards are implemented for documenting indicators and their associated databases. The first standard is designed to document each indicator individually, while the second standard pertains to the documentation of databases or datasets containing collections of indicators.

Indicators

Indicators are summary measures related to key issues or phenomena, derived (and possibly modelled) from observed facts. Indicators form *time series* when they are provided for a specific geography and with a temporal ordering, i.e. when their values are provided with an ordered annual, quarterly, monthly, daily, or other time reference.

The Bank developed a bespoke metadata standard for documenting indicators by consolidating and structuring metadata elements utilized by various organizations, including the Bank itself, United Nations agencies, the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), and others. The standard includes elements such as a unique identifier, label, definition, methodology information, details about the data producer and source, geographic and temporal coverage, a list of relevant keywords or topics, and information on known data limitations, among others.

The standard is compatible with the Statistical Data and Metadata Exchange (SDMX) framework. SDMX serves as a data exchange protocol designed to streamline machine-to-machine transfers. It offers a methodology for organizing metadata via *Metadata Structure Definitions* (MSDs), while not governing MSD content. Accordingly, the Bank's metadata standard augments and integrates seamlessly within the SDMX environment. Metadata templates developed by the Bank may be exported as MSDs, and metadata describing indicators may be presented as meta-datasets. This approach allows descriptive and reference metadata produced by the Bank to be incorporated into systems that comply with SDMX-based data management and dissemination requirements.

The indicator metadata must also include **structural metadata** that describes how the data are stored and organized in data files. Structural metadata is closely linked to **data modeling** which establishes both the logical and physical architecture of data by defining entity relationships, associated attributes, and integrity rules.[4] This form of metadata transposes the data model into a machine-readable format, thereby facilitating accurate processing and interpretation by information systems. Clearly articulating the data model and structure is critical for effective data dissemination through APIs. APIs (Application Programming Interfaces) enable standardized requests and exchanges of data between systems.

The Bank metadata standard for documenting indicators has a designated section for outlining an indicator's data structure (Figure 11). This structure follows the SDMX Data Structure Definition (DSD) and is based on the assumption that data are kept in a tabular long format.



Figure 11 - Data structure definition

A crucial part of describing data structures is specifying the *column_type* metadata element. This element defines the function of each column in a long-format data file and includes options such as:

- **Indicator ID**: The indicator unique identifier.
- **Indicator name**: Holds the main title or name of an indicator; there should be only one primary name (alternative names may be documented as *aliases* in the indicator metadata).
- **Geography**: Identifies the geographical area for each value, typically holding a geographic code, with corresponding names available in a code list or an attribute-type column.
- **Time period**: Specifies the time frame, like year, quarter, or month, applicable to each observation.
- **Dimension**: Used for additional breakdowns of an indicator, with one or several columns serving as dimensions—or potentially none at all. When combined with indicator ID, geography, and time period, these create unique identifiers for each observation. Dimensions typically use code lists to list possible values.
- **Periodicity**: The column indicates the (optional) periodicity of the indicator. Periodicity refers to the frequency or time interval at which data is collected, reported, or made available in a dataset. In the context of a Data Structure Definition (DSD), periodicity indicates how often the data for a particular observation is updated or recorded. SDMX provides a standard code list for periodicities.
- **Attribute**: An attribute is information associated with the data that describes a specific characteristic or property of the observation value but is not a geography or dimension. Attributes can have code lists associated with them.
- **Observation value**: The observation value. An indicator must have one and only one column identified as containing the information value (when data are stored in long format).
- **Annotation**: The column contains an annotation for the indicator. An indicator can have more than one annotation. An annotation is a piece of supplementary information or metadata that provides additional context or clarification about specific data or structures within the DSD. Annotations are typically used to provide non-technical explanations, clarifications, or other details that help users understand or interpret the data or metadata more accurately.

Let's consider an indicator "Population by Sex (Thousands)", which provides the population of Belgium for years 2010 and 2020 disaggregated by sex (male/female/total) and is stored as follows in the data file.

ID	Indicator	Country_code	Country_name	Year	Sex	Population	Source	Note
POP_SX	Population by sex	BEL	Belgium	2010	M	5312221	Census 2010	Statbel
POP_SX	Population by sex	BEL	Belgium	2010	F	5527684	Census 2010	Statbel
POP_SX	Population by sex	BEL	Belgium	2010	T	10839905	Census 2010	Statbel
POP_SX	Population by sex	BEL	Belgium	2020	M	5660064	Census 2020	Statbel
POP_SX	Population by sex	BEL	Belgium	2020	F	5832577	Census 2020	Statbel
POP_SX	Population by sex	BEL	Belgium	2020	T	11492641	Census 2020	Statbel

The data structure definition for this indicator would be defined as follows using the *data structure definition* of the metadata standard:

Column name in the data file	Information entered in the data structure definition
ID	Label: Indicator ID Description: Unique identifier of the indicator Data type: String Column type: Indicator ID

Column name in the data file	Information entered in the data structure definition
Indicator	Label: Indicator name Description: Title of the indicator Data type: String Column type: Indicator name
Country_code	Label: Country code Description: ISO country code Data type: String Column type: Geography ; Codelist: BEL = Belgium
Country_name	Label: Country name Description: Country name Data type: String Column type: Attribute
Year	Label: Year Description: Year Data type: Date Column type: Time period Time period format: YYYY
Sex	Label: Sex Description: Sex Data type: String Column type: Dimension ; Codelist: F = Female, M = Male, T = Total
Population	Label: Population Description: Resident population, mid-year Data type: Integer Column type: Observation value
Source	Label: Data source Description: Source of the data Data type: String Column type: Attribute
Note	Label: Note Description: Additional note Data type: String Column type: Annotation

Databases of Indicators

Indicators are typically organized and maintained within databases that aggregate multiple indicators, such as the World Bank's World Development Indicators. To facilitate this, the World Bank has established a customized metadata standard designed to describe each **database** or dataset associated with individual indicators. A component of this indicator metadata standard records the unique identifier for each database, thereby ensuring a clear link between an indicator and its relevant database(s).

3.5.3 Geographic Datasets and Services

Geographic datasets comprise data that specify and depict geographic locations, boundaries, and the physical or administrative attributes of features found on the Earth's surface. These datasets incorporate explicit geographic coordinates, such as latitude and longitude, alongside attribute information associated with particular spatial entities. Geographic datasets can consist of *vector data*, which model features as points, lines, or polygons organized into thematic layers, as well as *raster data*, which illustrate continuous phenomena or imagery through a grid of pixels, including sources like satellite and remote sensing imagery.

Geographic data services refer to the technologies employed for accessing, analyzing, and distributing geographic information. These services are commonly provided via interactive, web-based platforms.

The ISO Technical Committee on Geographic Information/Geomatics (ISO/TC 211) has created several metadata standards to help make geospatial datasets and services easier to find, use together, and understand. These include standards like ISO 19115-1 for describing vector geographic datasets, ISO 19115-2 for raster data, ISO 19110 for detailing feature-level characteristics, and ISO 19119 for geographic data services. Together, these are known as the ISO 19100 series. They serve as a foundation for major global geospatial metadata efforts, including those by the Open Geospatial Consortium (OGC), the U.S. Federal Geographic Data Committee (FGDC), the European INSPIRE Directive, the U.K.'s GEMINI profile, and, more recently, the Research Data Alliance (RDA).

Due to the comprehensive and intricate nature of ISO 19100, the Bank uses a streamlined implementation profile to document its geographic datasets and services. This approach maintains consistency with ISO standards, ensuring that metadata generated within the Bank's systems can be exported as ISO-compliant records whenever external users or platforms need them.

To promote consistency in the Bank’s metadata standards and improve feature descriptions, several new metadata elements have been added beyond those in ISO 19100. These additions were thoughtfully developed to maintain ISO compatibility, providing greater functionality while supporting interoperability with metadata standards used for other data types.

Core metadata elements included in ISO 19115 encompass the dataset title, unique identifier, originator and data source, geographic and temporal extent, data type, as well as the purpose or context of the dataset. For vector datasets, feature-level descriptions—addressed primarily by ISO 19110—are critical for both data discovery and usability; in contrast, for raster datasets, it is essential that metadata records band-level properties to ensure accurate interpretation. Recognizing that generating feature-level metadata can be a time-consuming process, the Bank has developed a Python library that automatically extracts key metadata components from widely used geospatial file formats and aligns these elements with ISO 19100 standards. [5]

3.5.4 Documents

The Bank developed a bespoke metadata standard for the documentation of documents. This metadata standard borrows elements from three international standards:

- **Dublin Core.** A widely recognized set of 15 elements that enables simple, interoperable description of resources across libraries and digital repositories.
- **MARC21** (Machine-Readable Cataloging). A widely recognized set of 15 elements that enables simple, interoperable description of resources across libraries and digital repositories.
- **BibTeX.** A reference tool for LaTeX users that automates citation and bibliography formatting in academic writing.

The standard also includes a set of core elements common to all its standards to ensure interoperability.

3.5.5 Images

In this Guide, **images** are defined as digital files meant for use in online catalogs or visual repositories. The most common file types are JPEG (JPG), PNG, and TIFF.

Digital cameras and devices usually add technical metadata to image files automatically. For instance, many cameras and smartphones generate EXIF (Exchangeable Image File Format) metadata, which includes details like when the photo was taken, what camera model was used, exposure settings, image size, and, if turned on, GPS location. Although EXIF metadata provides helpful technical information, it doesn't explain what the image depicts, why it was created, or how it can be reused. To enhance discoverability, accessibility, and rights management, it is important to include standard descriptive and reference metadata as well.

To meet this need, the Bank adopts two well-established international standards:

- **IPTC Photo Metadata Standard.** Developed by the International Press Telecommunications Council, the IPTC standard is widely used across media, cultural heritage institutions, and digital asset management systems. The Bank applies version 2019.1, which consists of two complementary standards:
- IPTC Core – general descriptive and administrative information (e.g., title, creator, copyright, locations, dates)
- IPTC Extension – enhanced semantic description (e.g., depicted people, objects, events, controlled vocabularies, and rights metadata)

> The richness of IPTC metadata enables sophisticated search, rights tracking, and attribute-based filtering, even though only a focused subset of fields is typically required for development-related imagery.

- **Dublin Core (DCMI).** The Dublin Core Metadata Initiative specification provides a simpler, highly interoperable set of 15 core elements applicable across many resource types. The Bank complements Dublin Core with selected schema.org *ImageObject* elements to improve indexing. The lightweight DCMI option is appropriate when resources are limited, or when broad cross-platform compatibility is prioritized.

AI-assisted metadata generation presents expanding possibilities within this field. Methods such as object and scene identification, optical character recognition for text extraction from images, facial detection, and automated keyword generation can greatly enhance the efficiency of metadata production. Nevertheless, rigorous human expertise remains vital to guarantee that semantic descriptions are accurate, pertinent, and unbiased.

3.5.6 Videos

The Bank documents videos (when considered as development data) using a hybrid metadata standard that combines elements from the **Dublin Core** Metadata Initiative (DCMI), the **IPTC Video Metadata Standard**, and **schema.org *VideoObject***. This approach aligns with the approach used for audio resources, ensuring consistent metadata practices across multimedia content.

Dublin Core's 15 core elements provide a simple and highly interoperable foundation, complemented by the IPTC standard, while selected *VideoObject* fields enable the description of audiovisual-specific technical characteristics—such as duration, encoding format, resolution, and frame size. These are complemented by additional descriptive elements commonly used in the Bank's metadata ecosystem (e.g., keywords, tags, topics, provenance, and administrative metadata).

Videos often contain only minimal embedded metadata, meaning that essential context resides within the spoken narrative or visual components. To expose this information for cataloging and search, the Bank's standard includes a *transcription* metadata element, allowing curators to incorporate human- or machine-generated transcripts of spoken content, and automatically translated versions when appropriate for multilingual access. Affordable speech-to-text

technologies exist for many major languages, enabling scalable and automated metadata generation. Once transcripts are available in the metadata, natural language processing methods—such as keyword extraction, topic modeling, and semantic search—can further enrich descriptive metadata.

3.5.7 Audio Recordings

The Bank documents audio files (when considered as development data) using a hybrid metadata standard that combines elements from the **Dublin Core** Metadata Initiative (DCMI), **schema.org** / *audioObject,* and **PBCore**. This approach aligns with the methodology used for images and video resources, ensuring consistent metadata practices across multimedia content. The Dublin Core and audioObject provide descriptive metadata elements. PBCore and audioObject provide metadata elements needed to document technical aspects of the audio files, such as sound quality, which are required for using audio files as input to AI model training. Metadata elements extracted from these standards are complemented by additional descriptive elements commonly used in the Bank's metadata ecosystem (e.g., keywords, tags, topics, provenance, and administrative metadata).

Artificial intelligence solutions facilitate the extraction of essential metadata from audio recordings, thereby enhancing documentation efficiency and consistency. These technologies can autonomously identify the languages present within an audio file, evaluate technical quality using established metrics such as signal-to-noise ratio, and often produce accurate transcriptions of spoken content. Leveraging advanced models capable of analyzing speech attributes and semantic information, AI enables thorough metadata collection with minimal need for manual input.

3.5.8 Data Processing and Analysis Scripts

The Bank operates an extensive empirical research program, generating a comprehensive and varied knowledge base on development through flagship publications, the Policy Research Working Papers series, and others. Recognizing the value of systematically documenting, cataloging, and disseminating programs and scripts used for data processing and analysis, the Bank has launched a reproducibility initiative and a dedicated reproducibility website.[6] This effort reinforces the Bank's commitment to openness and transparency and aligns with the core principle of verifiability outlined in the Policy on Development Data Quality.

Integrating reproducibility, replicability, and auditability into the Bank's research and data dissemination framework serves multiple purposes:

- **Enhancing research quality**: Public scrutiny encourages contestability and independent verification, promoting rigor and robustness in data analysis.
- **Facilitating reuse and extension**: Researchers can build upon Bank-conducted analyses, increasing the relevance, utility, and impact of both the data and analytical outputs.
- **Strengthening credibility**: Transparent methods bolster confidence in the Bank's research findings.
- **Supporting education**: Students and peers can access real-world examples of analytical workflows.

Technological platforms such as GitHub provide infrastructure for code preservation, sharing, and collaborative development. To produce and publish structured, comprehensive metadata for analytical programs and scripts, and to enhance the discoverability of these resources, the Office of the WBG Chief Statistician has developed a metadata standard for documenting research projects and scripts. It applies both to comprehensive research projects and to individual scripts associated with datasets, regardless of output type—whether publications, visualizations, tabulations, data transformations, or derived datasets and indicators. It underpins the Bank's catalog of reproducible research.

3.5.9 External Resources

External resources refer to various digital materials—such as data files, documents, web pages, images, and scripts—that can be attached as additional reference materials to data of any type.

When compiling metadata for any type of data, information is often gathered from several external sources. For example, when documenting a survey or a census, metadata will often be extracted from questionnaires, interviewer guides, data processing scripts, analytical reports, and other materials. These resources offer important background and should be maintained and published in their original form alongside both the main data and its structured metadata. To keep these supplementary materials properly referenced, accessible, and usable, they must be described using suitable metadata elements.

For documenting external resources, the Bank uses a simple standard based on Dublin Core metadata. Metadata for external resources is used alongside one of the data-type-specific metadata standards described in this Guide.

---

1. Available at https://worldbank.github.io/metadata-schemas/
2. Available at https://worldbank.github.io/metadata-editor-docs/
3. See https://ddialliance.org/
4. The Bank uses a more specialized and advanced tool, the Fusion Metadata Registry, to create and maintain DSDs and the related code lists.
5. Available at https://github.com/worldbank/geometadatatools
6. See https://reproducibility.worldbank.org/

3.6 Administrative Metadata

Administrative metadata is used to manage and operate data systems and is designed mainly for internal use, rather than being shared with external users. It provides important instructions for software applications, like data catalogs, on how to store, display, and access different types of data. This type of metadata helps automate tasks, maintain effective data governance, follow privacy protection requirements, and comply with the Bank's Access to Information Policy.[1] For example, it includes details about how data is classified, ensuring restricted or confidential information is clearly identified.

A dataset may be distributed through multiple catalogs or data systems, each subject to unique administrative protocols. For instance, the Bank operates both an internal and an external version of its Microdata Library. Although descriptive and structural metadata are consistent across these versions, the conditions for data access vary. Administrative metadata is specific to each system and offers tailored guidance regarding the publication and management of data and metadata on each respective platform.

Administrative metadata is created using a systematic process, much like descriptive metadata. IT specialists develop a template with specific metadata elements that match the system's technical needs and functions. This helps ensure the metadata supports how the system operates.

Administrative metadata is developed and maintained through the metadata editor application, allowing each *project* to be assigned one or more administrative metadata templates. These templates promote uniformity and support efficient data management across various systems.

A distinct category of administrative metadata, referred to as *presentation* or *portrayal* metadata, regulates the display of datasets within the system. This form of metadata may be integrated into the broader framework of administrative metadata.

Administrative metadata is maintained securely within the system and is excluded from publicly exported metadata. Authorized users can access this information through an API, enabling data management and dissemination platforms to retrieve and apply it as required.

---

1. See https://www.worldbank.org/en/access-to-information ↩

3.7 Relationship Metadata

Relationship metadata captures the associations between datasets and related digital assets, including additional datasets, analytical scripts, publications, and derivative data products. To facilitate thorough navigation and improved resource discovery, expressing these relationships bidirectionally is advisable whenever feasible. The adoption of persistent identifiers (such as DOIs or handles) is highly recommended to ensure reliable and long-term connections. Additionally, leveraging artificial intelligence for automated extraction and maintenance of these links can further enhance efficiency and accuracy.

There are several main types of relationships to consider:

- **Between datasets** – For instance, a survey might be the source for an indicator, or an anonymized version of a dataset could be derived from raw data.[1] While current metadata standards allow documentation of such links, they are often recorded in only one direction. Keeping such connections updated takes effort and isn't usually prioritized.

- **Between data and publications** – Metadata should show which datasets are cited in publications and which publications mention specific datasets. The Office of the Chief Statistician develops and maintains an AI-based solution to automatically extract data citations, helping build a structured and up-to-date database linking datasets and World Bank publications. The use of persistent identifiers like DOIs greatly improves precision, especially given inconsistent citation practices.

- **Between datasets, scripts, and analytical outputs** – Scripts used for data processing and analysis connect datasets to the results they produce. Metadata about these relationships can link scripts, datasets, and outputs like tables, figures, visualizations, and reports. The metadata standard used to document data processing and analysis scripts provides the elements needed to establish such relationships.[2]

Creating and maintaining metadata that captures relationships between data resources is resource-intensive and time-consuming. Given ongoing needs to strengthen core metadata—descriptive and structural—by improving completeness, accuracy, quality control, and alignment with standards, relationship metadata is not currently prioritized.

---

1. In time series, where each value may have a different source, the identification of the source may be included as a text variable in the time series database instead of being provided in the metadata. ↩
2. For an example of implementation, see the Bank's reproducibility platform at http://reproducibility.worldbank.org ↩

3.8 Compatibility with Other Metadata Standards

The metadata standards outlined in this section are not utilized by the Bank for development data documentation (with the exception of SDMX for indicators' structural metadata). Nonetheless, these standards are related and contribute to the broader metadata ecosystem for development data. Where appropriate and applicable, the Bank aligns its implementation of the metadata standards defined by the Chief Statistician with these other standards to maintain consistency.

3.8.1 Statistical Data and Metadata Exchange (SDMX)

The Statistical Data and Metadata eXchange (SDMX) [1], published as an ISO International Standard (ISO 17369), establishes a framework for exchanging statistical data and metadata using modern information technology. As one of the sponsoring organizations of SDMX, the World Bank is implementing SDMX for selected indicator datasets, with a strategic goal of ensuring SDMX compliance in the dissemination of all its time series data.

SDMX is primarily designed for machine-to-machine exchange of statistical data and metadata. Its core artefacts include:

- **Data Structure Definitions (DSDs)**: templates that specify the dimensions, attributes, and measures for a dataset (i.e., the structural metadata required for data exchange).
- **Metadata Structure Definitions (MSDs)**: templates that define the structure of metadata reports which contain descriptive and reference metadata.
- **Concepts and Code Lists**: Concepts (e.g., "reference area," "time period," "unit," "obs value") are reusable semantic definitions that give meaning to dataset components; they are referenced in DSDs and MSDs. Code lists provide the permissible coded values for those concepts.

SDMX allows users to define MSDs as needed but does not dictate their exact content. To support thorough and useful descriptive and reference metadata for indicators, the Bank's metadata standard offers a well-organized collection of suggested elements. As a result, SDMX and the Bank's metadata standard work together in a complementary way, rather than duplicating efforts.

To facilitate SDMX integration, the Bank's metadata standard and the Metadata Editor application have been designed with interoperability in mind:

- The "data structure definition" section of the Bank's indicator metadata standard allows for the export of SDMX-compliant DSDs.
- The templates and controlled vocabularies developed in the Metadata Editor can be exported as SDMX-compliant MSDs.
- Metadata generated using the Metadata Editor can be exported as SDMX-compliant metadatasets and reports.

3.8.2 DCAT, Schema.org, and Croissant

To improve how data can be discovered, shared, and accessed online, the Bank aligns its development data documentation with three complementary, widely adopted metadata standards: DCAT, Schema.org, and Croissant. While these are not used for managing metadata internally, they enhance interoperability with external applications and increase visibility through internet search engines and data catalogs. Ensuring compatibility with these generic standards keeps the Bank's metadata ecosystem aligned with global data networks, making it easier for external users to find, access, and integrate Bank data across web platforms and search tools."

- **DCAT (Data Catalog Vocabulary)** [2] is a widely used vocabulary for publishing data catalogs on the Web, originally developed for government open data portals.
- **Schema.org** [3] is a community-driven initiative founded by Google, Microsoft, Yahoo, and Yandex to develop a collection of schemas for structured data of any type published on the Internet. These schemas cover descriptive and metadata intended to improve search engine discoverability and facilitate metadata use and exchange.
- **Croissant** [4] is a metadata format for machine learning datasets, developed as part of the MLCommons initiative by a collaboration of industry and academic partners. It extends Schema.org and provides a standardized, machine-readable structural metadata format for machine-learning (ML) datasets that makes datasets easier to discover, load across ML frameworks, and reuse with consistency.

The generation of metadata compliant with these additional, generic standards is automated. A defined subset of fields from the Bank's specialized metadata standards is programmatically extracted and mapped to DCAT, Schema.org, and Croissant profiles as appropriate, producing compliant metadata artifacts without manual intervention by data curators. This approach minimizes burden on teams, ensures consistency, and enables continuous synchronization as source metadata evolves. Since DCAT, Schema.org, and Croissant metadata include only part of the canonical metadata, it is essential to maintain a clear connection to the definitive metadata records.

3.8.3 Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH)

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a simple mechanism for repository interoperability. Data Providers are repositories that expose structured metadata via OAI-PMH. Service Providers then make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP. [5] The Bank seeks to develop data catalogs that comply with the OAI protocol. The metadata standards used by the Bank provide all necessary elements for compliance with OAI-PMH.

1. https://sdmx.org/ ↩
2. https://www.w3.org/TR/vocab-dcat-3/ ↩
3. https://schema.org and https://schema.org/Dataset ↩
4. https://mlcommons.org/2024/03/croissant_metadata_announce/ and https://research.google/blog/croissant-a-metadata-format-for-ml-ready-datasets/ ↩
5. https://www.openarchives.org/pmh/ ↩

Annex 1 – Recommendations to System Developers

Adopting metadata standards from the outset in the design and development of data management and dissemination systems is essential to ensuring that information can be seamlessly shared, understood, and reused across platforms. Standards enable interoperability, improve discoverability, and support accurate interpretation and meaningful use of data by both humans and machines. To realize these benefits, applications for processing, managing, and disseminating data must be architected to generate, store, and expose standardized metadata throughout the data lifecycle.

Storage and Indexing of Metadata

Adopting metadata standards from the outset in the design and development of data management and dissemination systems is essential to ensuring that information can be seamlessly shared, understood, and reused across platforms. Standards enable interoperability, improve discoverability, and support accurate interpretation and meaningful use of data by both humans and machines. Realizing these benefits requires that applications be architected to generate, store, and expose standardized metadata throughout the data lifecycle. This includes the ability to manage metadata in its structured, standardized form (typically provided as JSON) and to integrate with APIs for acquiring, processing, and republishing metadata as part of automated workflows.

JSON Metadata in SQL-based Systems

The adoption of structured and comprehensive metadata compliant with the Bank's standards and stored in JSON format provides significant advantages for developers of SQL-based data management and dissemination systems. Modern relational database management systems (RDBMS) offer extensive support for JSON data, enabling efficient storage, querying, and manipulation of metadata while maintaining the flexibility required to accommodate diverse data structures.

The advantages of JSON metadata in SQL systems are:

- **Efficient storage and retrieval**: SQL databases, such as PostgreSQL, MySQL, and SQL Server, provide native JSON data types and indexing mechanisms that allow for fast storage and retrieval of metadata. JSON documents can be stored in dedicated columns, enabling developers to efficiently query and extract relevant metadata elements without the need for complex joins or schema modifications.

- **Flexible and schema-agnostic structure**: Unlike traditional relational schema constraints, JSON metadata offers a flexible structure that can accommodate evolving metadata standards and custom extensions. This flexibility allows SQL-based systems to ingest and process metadata with varying structures while maintaining consistency and compliance with established standards.

- **Advanced query capabilities**: SQL databases provide powerful functions for processing JSON data, including path-based queries, filtering, and aggregation. Features such as PostgreSQL's jsonb indexing, MySQL's JSON_TABLE, and SQL Server's OPENJSON allow developers to efficiently navigate hierarchical metadata structures and extract insights without compromising performance.

- **Integration with relational data**: JSON metadata can be seamlessly integrated with structured relational data within the same SQL environment. Developers can create hybrid solutions that leverage the strengths of relational modeling while embedding rich, nested metadata structures that do not fit traditional table designs.

- **Ease of updates and versioning**: JSON's self-describing nature allows for easier updates and versioning of metadata, supporting the evolution of data documentation without necessitating database schema modifications. This adaptability ensures long-term maintainability and scalability of metadata management within SQL systems.

- **Interoperability and data exchange**: The widespread adoption of JSON as a data interchange format facilitates seamless integration with external applications, APIs, and web-based platforms. Developers can leverage

standardized JSON metadata files for interoperability across diverse systems while maintaining consistency in metadata representation.

To maximize the benefits of JSON metadata in SQL-based systems, developers should consider the following best practices:

- Use appropriate **indexing strategies**, such as JSONB indexing in PostgreSQL, to optimize query performance.
- **Validate JSON metadata** against predefined schemas to ensure compliance with metadata standards.
- Leverage SQL functions to **perform partial updates and efficient querying** without extracting entire JSON documents.
- Implement appropriate **security measures** to protect sensitive metadata information stored in JSON format.

JSON Metadata in NoSQL-based Systems

NoSQL databases are inherently well-suited for storing and managing JSON metadata, offering high flexibility, scalability, and performance for unstructured and semi-structured data. Developers and administrators of NoSQL-based systems can take full advantage of JSON metadata to enhance their data management and dissemination capabilities.

The advantages of JSON metadata in NoSQL systems are:

- **Native JSON storage and querying**: NoSQL databases such as MongoDB store JSON natively, allowing for seamless ingestion, querying, and retrieval of metadata without the need for transformation. This direct support enhances system performance and reduces data processing overhead.
- **Schema flexibility and adaptability**: Unlike SQL databases, NoSQL systems do not enforce rigid schemas, making them ideal for storing metadata with evolving structures. This flexibility allows organizations to continuously update their metadata standards without the need for schema alterations or migrations.
- **High scalability and performance**: NoSQL databases are designed to scale horizontally, handling large volumes of JSON metadata efficiently. This scalability ensures that metadata management systems can accommodate growing datasets and user demands without degradation in performance.
- **Complex hierarchical data handling**: JSON's hierarchical structure aligns naturally with the document-based model of NoSQL databases, making it easier to store, retrieve, and manipulate nested metadata elements without requiring complex joins or transformations.
- **Integration with cloud and big data technologies**: Many NoSQL databases provide seamless integration with cloud platforms and big data ecosystems, enabling efficient metadata dissemination across distributed environments and enhancing interoperability with data lakes and processing frameworks.
- **Real-time processing and analytics**: NoSQL systems support real-time data processing, allowing to analyze and act on metadata insights dynamically. This capability is critical for applications requiring up-to-the-minute metadata updates and rapid decision-making.

To fully leverage JSON metadata in NoSQL-based systems, developers and administrators should consider the following best practices:

- Design metadata structures that take advantage of NoSQL's flexible schema model while maintaining consistency.
- Utilize indexing and aggregation features to optimize query performance and retrieval times.
- Implement backup and replication strategies to ensure metadata availability and durability.
- Leverage API-driven architectures to facilitate seamless metadata sharing across applications.

Development of Data Dissemination Platforms

Data and metadata stored in back-end (or front end) systems may serve as input for multiple front-end data dissemination systems. The WBG has decentralized systems within, and permissive licensing for outside.

Ensuring that metadata standards are consistently implemented across all systems is critical to maintaining data integrity, interoperability, and usability. One essential aspect of this implementation is making metadata available in back-end systems through well-defined APIs. Providing API access to metadata enables developers of front-end systems and applications, including dashboards and data dissemination user interfaces (UIs), to build solutions that enhance data discoverability and interpretability for end-users.

By offering API access, both internal and external developers can retrieve standardized metadata in real time, ensuring that their applications accurately reflect the most up-to-date descriptions, classifications, and contextual information about the data. This capability empowers developers to create intuitive and user-friendly interfaces that facilitate efficient data navigation and comprehension, ultimately improving the overall user experience.

Furthermore, third-party developers who redistribute the Bank's open data should have the same opportunity to leverage metadata through API access. Providing these stakeholders with standardized and machine-readable metadata allows them to incorporate the data into their platforms and services in the most efficient and cost-effective manner. This not only enhances the visibility and reach of the Bank's data products but also promotes innovation and broader utilization of the data for diverse applications.

API-based access to metadata fosters interoperability by enabling different systems and applications to integrate seamlessly with the Bank's data infrastructure. This ensures consistency in how data is presented and interpreted across various platforms, minimizing discrepancies and enhancing trust among data users.

To maximize the benefits of API access to metadata, the Bank should ensure that:

- **APIs are well-documented and adhere to industry best practices**, providing clear guidance on how to query, retrieve, and utilize metadata effectively.
- **Metadata APIs support commonly used formats**, such as JSON and XML, to accommodate different technical environments and preferences.
- **Authentication and access controls are appropriately implemented**, balancing open access with necessary security measures for sensitive metadata elements.
- **Version control mechanisms are in place**, allowing developers to adapt to updates without disrupting existing integrations.

Development of Other Types of Applications

The integration of the Bank’s metadata standards into software applications fosters the efficiency, interoperability, and long-term usability of data. When software developers embed metadata standards within their applications, they facilitate seamless data documentation, sharing, and reuse across different platforms and organizations.

Software applications that support metadata standards, such as the Data Documentation Initiative (DDI), automate the generation of comprehensive metadata, reducing the need for manual documentation and minimizing the risk of errors. For example, data entry tools like Survey Solutions by the World Bank and CsPro by the U.S. Census Bureau incorporate DDI-compliant metadata generation capabilities.[1] This built-in functionality allows users to capture essential metadata elements such as survey structure, variable definitions, and data collection methodologies in a standardized format. By automating metadata creation, these tools streamline data management processes, improve data accuracy, and enhance operational efficiency.

Embedding metadata standards into software applications promotes interoperability by ensuring that data can be seamlessly exchanged and integrated across different systems and institutions. Standardized metadata formats enable compatibility with other applications and repositories, facilitating collaboration and data sharing. When applications adhere to widely accepted metadata standards, they can easily interface with data dissemination platforms, statistical systems, and analytical tools, enhancing the overall data ecosystem's cohesiveness. This interoperability is particularly crucial in multi-stakeholder environments, where diverse organizations contribute to and consume data.

1. Another example is QGIS which exports metadata compliant with the ISO 19139 metadata standard for geographic datasets. ↩️

Annex 2 – Guidance for the Procurement of Data

The Bank seeks to ensure compliance with metadata standards for acquired and commissioned data. To ensure that the Bank's metadata standards are consistently applied across all datasets acquired or commissioned, specific strategies must be employed depending on the source of the data. These sources include data purchased from vendors, data obtained free of charge from providers, and data collected or generated by consultants (firms or individuals) under contract. The Bank aims to maximize compliance with its metadata standards by providing templates, guidelines, and support to facilitate the adoption of these standards.

The Bank has established formal procedures and Guidance for the acquisition of development data:

- Bank Procedure for Development Dataset Acquisition, Archiving and Dissemination. Procedural instructions to Bank Staff on the archiving and dissemination of development datasets and guidance on acquisition.
- Bank Guidance Development Dataset Acquisition and Archiving. Guidance to Bank Staff on the acquisition and archiving of development datasets from countries, international organizations, commercial vendors, etc.

The recommendations below complement these procedures and guidance.

Data Purchased from Vendors

For data acquired through purchase agreements, the Bank must ensure that metadata accompanying the datasets meets its quality standards to facilitate responsible data use and seamless integration into internal systems.

Vendors should be required to provide datasets with full metadata, preferably structured according to a standardized template provided by the Bank. Clear instructions on metadata content, structure, and formatting must be included in the procurement documentation.

In cases where purchased data is pre-packaged with existing metadata that cannot be structured following a Bank template, the business unit who purchases the data must conduct a thorough verification to ensure that all necessary metadata elements required for responsible use are present. A checklist based on the Bank's metadata standards or templates should be used to assess the completeness and accuracy of vendor-provided metadata.

Data Obtained Free of Charge

The Bank obtains many datasets free of charge, from national statistical offices and many other partners and organizations. Some of these datasets are provided under a formal agreement with the provider, others are downloaded from these organizations' platforms.

For datasets provided at no cost and/or online, the Bank cannot impose specific metadata requirements. However, it can in some cases encourage and support providers in adopting the Bank's standards (which comply with international best practice and align with international metadata standards) through voluntary measures:

- Providers should be offered access to detailed guidelines outlining metadata best practices aligned with the Bank's standards.
- Free tools should be provided to facilitate metadata structuring and formatting according to the Bank's requirements.
- The Bank should highlight the benefits of metadata standardization, such as increased data discoverability and usability.
- Technical assistance should be offered to help providers align their metadata with the Bank's standards.

Even when providers do not fully comply with the recommended standards, the Bank should conduct internal reviews to identify gaps and take corrective actions, such as supplementing missing metadata elements where possible.

Data Collected or Generated for the Bank by Consultants

For datasets collected or generated under consultancy contracts, the Bank has greater control over metadata compliance. Compliance can be ensured through contractual obligations and technical support:

- All contracts with consultants and consulting firms engaged in the collection, production, or acquisition of development data must include a requirement to submit metadata compliant with the Bank metadata standards. Metadata templates and specific requirements should be detailed in the terms of reference (TORs) for each project.
- Consultants must be given access to standardized metadata templates and comprehensive documentation to facilitate adherence to the required standards.
- Clear guidance on metadata content, structure, and formatting should be provided at the project initiation phase.
- Training sessions or onboarding meetings can be arranged to ensure consultants understand metadata expectations.
- A designated point of contact should be available to address queries and provide ongoing support throughout the data collection process.

Final deliverables should be assessed against the agreed metadata standards, and any deficiencies must be addressed before acceptance of the data.

Annex 3 – References and Resources

World Bank resources

Documents

1. **World Bank Group**. *Development Data Quality Policy*. A policy setting forth key principles for quality assurance, control, and improvement of World Bank Group Development Data.
2. **World Bank Group**. *Policy on Access to Information*. A policy governing public accessibility of information in the Bank's possession.
3. **World Bank Group**. *Policy on Personal Data Privacy*. A policy setting forth the core privacy principles governing the processing of personal data by World Bank Group institutions.
4. **World Bank**. *Bank Procedure: Development Dataset Acquisition, Archiving, and Dissemination*. Procedural instructions to Bank staff on the archiving and dissemination of development datasets, including guidance on acquisition.
5. **World Bank.** *Bank Guidance: Development Dataset Acquisition and Archiving*. Guidance to Bank staff on the acquisition and archiving of development datasets from countries, international organizations, and commercial vendors.

Standards and Tools

6. **Metadata standards for development data.** API description of metadata standards adopted by the World Bank for development data. https://worldbank.github.io/metadata-schemas/
7. **Metadata Editor software application.** An open-source multi-standard metadata editor for development data. https://github.com/worldbank/metadata-editor
8. **Metadata Editor User and Practice Guide.**https://worldbank.github.io/metadata-editor-docs/
9. **GeoMetadataTools.** An open-source Python library for the extraction of metadata from geographic data files. https://github.com/worldbank/geometadatatools
10. **NADA cataloguing application.** An open-source multi-standard data and metadata cataloguing application, developed for the International Household Survey Network (IHSN). https://github.com/ihsn/nada

External resources

11. **BibTeX**. A file format used to describe lists of references. https://www.bibtex.org/

12. **Croissant.** A high-level format for machine learning datasets that integrates metadata, resource file descriptions, data structure, and machine learning semantics. https://github.com/mlcommons/croissant

13. **Data Documentation Initiative (DDI) Alliance**. Official website of the DDI Alliance. https://www.ddialliance.org

14. **Dublin Core (DCMI).** Official website of the Dublin Core Metadata Initiative. https://www.dublincore.org.

15. **GEMINI (GEo-spatial Metadata INteroperability Initiative)**. The UK metadata template for the documentation of geographic metadata standard, built on ISO 19115-2003. <https://agiorguk.github.io/gemini/1037-uk-gemini-standard-and-inspire- implementing-rules.html>

16. **INSPIRE (INfrastructure for SPatial InfoRmation in Europe)**. The European Union template for the documentation of geographic datasets and services based on ISO standards. https://inspire.ec.europa.eu.

17. **IPTC (International Press Telecommunications Council)**. Open standards for the news media (metadata standards for images and videos). https://iptc.org/

18. **ISO 19110.** Geographic information — Methodology for feature cataloguing. https://www.iso.org/standard/57303.html

19. **ISO 19115.** Geographic information — Metadata for vector (ISO 19115-1) and raster (ISO 19115-2) geographic datasets. https://www.iso.org/standard/53798.html;

20. **ISO 19119.** Geographic information — Metadata for geographic data services. https://www.iso.org/standard/57303.html;

21. **ISO 19139.** XML Schema Implementation for geographic datasets and services. <https://committee.iso.org/sites/tc211/home/projects/projects---complete-list/ iso-19139-1.html>

22. **MARC21 (Machine-Readable Cataloging).** Format for Bibliographic Data by the US Library of Congress. https://www.loc.gov/marc/bibliographic/.

23. **National Information Standards Organization (NISO)**. 2004. Understanding Metadata. https://www.niso.org/publications/understanding-metadata.

24. **PBCore**. A cataloging standard for the description of audiovisual content. https://pbcore.org/

25. **Gordon Priest.** 2010. The Struggle for Integration and Harmonization of Social Statistics in a Statistical Agency. IHSN Working Paper No 4

26. **Schema.org**. A collaborative initiative for developing schemas to structure data on the internet. https://schema.org.

27. **SDMX (Statistical Data and Metadata eXchange)**. Official website of the SDMX community. https://sdmx.org.