Open Data Capacity Development
Module 2: Open Metadata
Training Syllabus

This training module was developed under the guidance of the United Nations Statistical Division as part of the Data4Now Initiative.

Introduction

This syllabus is intended to guide trainers who are presenting a course on open data for staff in NSOs and other agencies of the national statistical system responsible for documenting and disseminating open data and databases of indicators. It may also be of interest to anyone who wishes to know about the implementation of open data principles. The training module is part of a larger program on the development of open data capacity in official statistics agencies. It is accompanied by a PowerPoint presentation that can be used for group presentations or for individual learning.

What do I need to know before using this module?

This module provides a broad survey of the principles of open data as applied to the work of national statistical agencies. It does not require prior familiarity with open data or the operation of a statistical agency.

Learning objectives

- What are metadata and their importance for open data?
- Become familiar with international standard schemas for metadata
- Learn about types of metadata and their functions
- How to use standards and code lists to facilitate sharing
- How to draft and compile metadata

Table of Contents

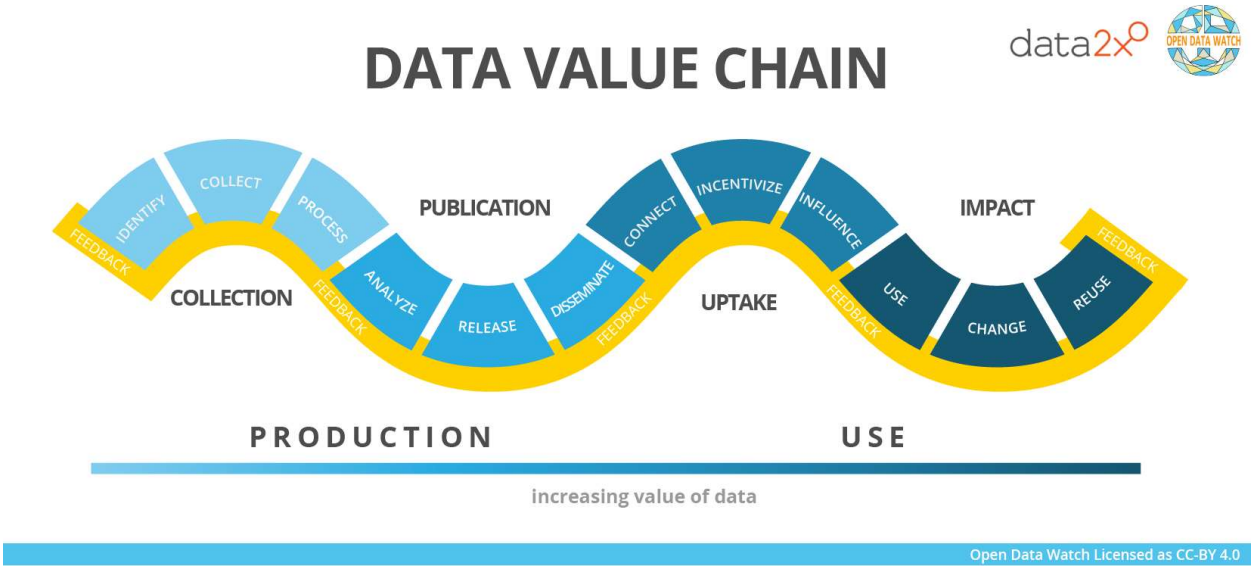
1. Open Data and Metadata

Metadata are data about data.

Metadata are an essential component of open data because they provide public information about who or what was measured, how they were measured, where they were collected, their time coverage and frequency, who is responsible for managing the data, and who can access and use them.

Metadata increase the value of data at every stage of the data value chain. Without metadata, the data would be of little value to anyone; they would simply be numbers on a page or in a file. During data collection, metadata record technical details about the production and management of data, helping the NSS to improve its processes. They improve the coordination within the NSS by making it easier to exchange data between agencies or external partners. And they promote transparency and accountability within the statistical system.

Figure 1 The Data Value Chain



Source: Open Data Watch

Metadata can describe many types of data: microdata from surveys or administrative sources, graphic images, videos, reports, or books. Here we are concerned with indicators—sometimes called macrodata— and time series of indicators. For example, the indicators of the Sustainable Development Goals or the official statistics produced by national statistical systems.

When data are published, metadata guide the analysis of data and the design of data visualizations. When data are disseminated outside the NSS, metadata make it possible to find and use data by naming and defining the data, describing how they can be accessed, and identifying the responsible agency. And to qualify as open data open data, the metadata must include a data license that says the data may be freely used and reused.

Metadata may occur at different levels of the statistical system. In this module, we are primarily concerned with metadata that apply to an indicator or a time-series of indicators. Metadata may also describe a database and its content. At a higher level, information about the statistical law and management of the statistical system are metadata that apply to entities in the statistical system. There can also be high level metadata that describe the mandate and activities of a statistical agency.

2. Types of metadata

In this section we look at the three principal types of metadata used to describe time series data and indicators. Each type has a different but complementary function.

Structural metadata

Structural metadata are information about how the data are organized and stored. They specify the identify of an indicator and its dimensions along with other attributes of the statistical observations that constitute the data.

- **Dimensions** describe the conceptual organization of the data. An indicator typically has a time dimension, a location dimension, and sometimes other characteristics differentiate the observations on an indicator.

  For example, to retrieve data on an SDG indicator you need to know the indicator name (its identity), the geographic reference area, and the time period. The metadata may also include a dimension that specifies the sex of the subject, the unit of account, or a subnational region. When talking about an indicator, these dimensions are often spoken of as its disaggregations.

  Attributes provide additional information about the dimensions. For example, an attribute may specify the scale factor used to represent the value of the data.

Reference metadata

Reference metadata (sometimes called descriptive metadata) include the unique identifier of the data and a basic description of the data, such as how and where they were collected and any adjustments or revisions that have been made; they may also provide information about the uses or applications of the data. Reference metadata can be quite extensive. For example, the SDMX Global Metadata Concept lists 80 elements that can be included as reference data. The major headings are:

- Scope
- Methodological Information
- Statistical processing
- Quality
- Dissemination
- Miscellaneous

Administrative metadata

Administrative metadata provide information about data management and the legal responsibilities for managing and using data. They are sometimes treated as part of the reference metadata, but they have a distinct function for the statistical agency. They are important for maintaining a record of institutional arrangements for producing and disseminating data and defining policies for data dissemination.

Although much of the administrative metadata may be intended for internal use, the terms of use or data license that define the rights and permissions for use and reuse of the data are a crucial element of open data and must be available to all users. Information on the agency responsible for producing the data and contact information should also be shared with end users.

In the next section we will look more closely at the reference metadata needed by data users.

3. Metadata to support data users

Metadata provide information that is very important for the effective use of open data. Different users will need different elements of metadata. For example, someone wanting to access data remotely through an API will use the structural metadata, including associated code lists, to populate the API request. At the same time, they will want to download the descriptive and administrative metadata needed for their use of the data. Researchers may want more

detailed information on the data collection and computation methodologies, while more casual users may require only a brief definition of the indicator.

Data managers who are responsible for documenting and disseminating data should choose a data structure and metadata schema that best describes their data. In doing so, they can select from the long list of metadata elements included in the standard schemas or they may include custom elements to reflect special characteristics of their data and the needs of their users.

Example: The World Development Indicators

The World Bank's World Development Indicators (WDI) database must serve a wide range of data users and provide metadata that robustly describe the contents of the database. Although the WDI metadata do not conform to a standard schema, they constitute a schema in their own right, providing a mix of reference and administrative metadata. Here is an example of the metadata provided for GDP data retrieved from the WDI. Note that the License type and a link to the full license are provided with the metadata.

Metadata element	**Description**
Indicator Name	GDP (current US$)
Long definition	GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars. Dollar figures for GDP are converted from domestic currencies using single year official exchange rates.
Source	World Bank national accounts data, and OECD National Accounts data files.
Topic	Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators
Periodicity	Annual
Aggregation method	Gap-filled total
Statistical concept and methodology	Gross domestic product (GDP) represents the sum of value added by all its producers. Value added is the value of the gross output of producers less the value of intermediate goods and services consumed in production before accounting for consumption of fixed capital in production. … Total GDP is measured at purchaser prices. Value added

	by industry is normally measured at basic prices.
Limitations and exceptions	Gross domestic product (GDP), though widely tracked, may not always be the most relevant summary of aggregated economic performance for all economies, especially when production occurs at the expense of consuming capital stock. … Among the difficulties faced by compilers of national accounts is the extent of unreported economic activity in the informal or secondary economy. In developing countries, a large share of agricultural output is either not exchanged (because it is consumed within the household) or not exchanged for money.
License URL	https://datacatalog.worldbank.org/public-licenses#cc-by
License Type	CC BY-4.0

(Some of the descriptions have been abbreviated, denoted by an ellipsis.)

With the downloaded data, the WDI provides three important pieces of information: the database name, the date of the last update, and the indicator code.

Data from database	World Development Indicators
Date Last Updated	03/28/2024
Series code	NY.GDP.MKTP.CD

In addition, metadata for this indicator provides information on each countries' compilation of its national accounts, including the national accounts base year, national accounts reference year, the system of national accounts (SNA) used, and SNA price valuation.

4. International Metadata Schemas

A metadata schema is a standardized method for recording and disseminating information about a set of data or other objects. The objects described by the metadata may be digital such as databases, videos, or web pages or physical such as books or works of art. The metadata themselves are usually stored in machine readable formats according to the specifications of the schema so that they can be retrieved and used to locate the objects they describe.

In this section we look at three well-known metadata schemas used by international organizations and others. The use of a standard schema and its descriptors facilitates

interoperability and data sharing by allowing the exchange of information between systems that will be consistently interpreted by all users. But data producers can also construct a customized schema based on the nature of their data and the needs of data users.

Data Documentation Initiative

Data Documentation Initiative (DDI) is an international standard maintained by the Data Documentation Initiative Alliance. It has been widely used for documenting the metadata for microdata produced by surveys in the social sciences. Earlier versions were focused on the information found in survey codebooks. Version 2.0 covers the whole data lifecycle from survey instrument design to dissemination.

DDI is the standard used by the International Household Survey Network's Central Data Catalog and National Data Archive (NADA) application for documenting and cataloging surveys. Because the microdata documented by DDI metadata are often aggregated to higher level indicators, there is an effort underway to align SDMX and DDI with each other and with other metadata standards.

Statistical Data and Metadata Exchange

Statistical Data and Metadata Exchange (SDMX) is an international standard (defined by ISO 17369:2013) for documenting and exchanging data and metadata. SDMX is supported by the UN, OECD, ILO, IMF, World Bank, Eurostat, and the European Central Bank, Bank for International Settlements, and national statistical offices.

Data and metadata can be stored using any application used by the statistical agency in whatever form is most effective for its requirements. When the data (and metadata) are exchanged with other applications or shared between agencies, they are rendered as SDMX-ML messages (a form of the XML markup language) or in JSON. The content of the message is described by the Data Structure Definition (DSD). Metadata alone can be described using the SDMX Metadata Structure Definition (MSD).

In SDMX the Data Structure Definition (DSD) is used to define the concepts or characteristics of a statistical observation and its values. Three types of concepts can be declared in a DSD: dimensions, attributes, and the measures that are the observational values. If code lists are used to represent dimensions, they must also be declared in the DSD.

The Working Group on SDMX Sustainable Development Goals has developed a data structure definition for the SDG indicators. Guidelines for the Global SDG DSD are available here. There are also guidelines for the customization of the Global SDG DSD.

More information about SDMX is available on the SDMX website.

Dublin Core

Dublin Core is an international metadata standard (ISO 15836) for describing digital or physical resources. The core schema specifies 15 metadata elements that have now been extended to 55 metadata terms. The specifications for the Dublin Core are maintained by the Dublin Core Metadata Initiative.

The Dublin Core was the first metadata standard for describing web content. The core consists of metadata tags that are based on RDF (resource description framework) and are designed to facilitate interoperability across the semantic web. However, the Dublin Core vocabulary can also be used within databases to specify the contents of metadata records.

Relationship to national metadata schemas

The international schemas provide a framework for constructing national metadata schemas. They can be adopted in their entirety, but often they are adapted to national schemas based on local requirements.

5. Using standard vocabularies to facilitate data sharing

Sharing of data between individuals or organizations is made easier when common terminology or "vocabularies" are used to describe the data. These commonalities can occur at three levels: the definition of the indicator; the definition and tag of the metadata element; and the code lists used to describe the dimensions and attributes of the data.

Data definition

The definition of the data or data series depends on the methodology used to construct the data. For many types of data in the social and economic domains there are internationally recognized definitions, classifications, and standards for constructing and reporting indicators. Examples of such standards are the System of National Accounts, the Balance of Payments Manual, International Classification of Disease, and the International Standard Classification of Education. National and international bodies provide definitions and classifications for many other measures. To facilitate interoperability, the descriptive metadata should specify the standards used to construct a data series and note any deviations from the standards. Data sharing will be further facilitated by using standard descriptions for the indicators and their dimensions.

Metadata definition

Standard metadata schemas provide names (tags) and definitions for metadata items. For example, among the 80 reference metadata items included in the SDMX Core Metadata Concepts are the following.

Name	Definition	Concept ID
Source data type	Characteristics and components of the raw statistical data used for compiling statistical aggregates.	SOURCE_TYPE
Data collection method	Method applied for gathering data for official statistics.	COLL_METHOD
Reference period	Timespan or point in time to which the measured observation is intended to refer.	REF_PERIOD
Base period	Period of time used as the base of an index number, or to which a constant series refers.	BASE_PER

Two of these items – source data type and data collection method – will require free form answers to describe them, but the definitions should be kept as short as possible using well-recognized terminology. The latter two items that refer to time periods are examples of metadata that should use code lists included as part of the reference metadata. In this case, if the metadata specify that time period is annual, the code list is simply the numerical year, but monthly data require a code or agreed way to represent the year and month.

Code lists
Metadata schemas should specify the code lists used to describe the dimensions or attributes of the data. The code lists may be stored separately and identified by a URI (uniform resource identifier). In SDMX code lists are specified as part of the DSD.

The use of codes provide a degree of language independence that further facilitates interoperability. For example, consumption data classified using the Classification of Individual Consumption According to Purpose (COICOP) can use the code CP011 for Food in English or the equivalent description in any other language.

SDMX provides cross-domain code lists for many topics. Other domain-specific code lists are also available. When custom metadata concepts are used, it may be necessary for the data provider to provide custom code lists.

6. Drafting and compiling metadata

All indicators produced and disseminated by a statistical organization should be accompanied by sufficient metadata to ensure that an indicator can be understood and used properly. The metadata schemas previously discussed provide generic lists from which metadata elements may be selected. Some applications, such as SDG reporting, have templates that specify required metadata elements and tools for editing and publishing metadata. But even without such tools, statistical offices can compile metadata and make them available to data users.

Creating a metadata template

The template is a guide for recording the structural and reference metadata in a standard format for all the indicators in a database. The elements can be selected from a standard metadata schema or customized by the data manager. In either case, the template should include definitions of each element –metadata for the metadata -- to guide the reporter completing the template.

In constructing the template, think about the information needed to answer these questions.

- What are the dimensions needed to identify an observation?
- What attributes are needed to provide additional information about the observations?
- What codes are used to represent values of the dimensions or attributes?
- What reference metadata are needed to fully describe data?
- What administrative data are needed to inform data users of the data's provenance and licensing terms.

Here are some tips for creating clear and useful metadata

- Metadata should be comprehensive but not overwhelming
- Use clear and simple language
- Keep sentence and paragraphs short
- Avoid technical terms, jargon, and unexplained acronyms
- Use standard terminology or refer to a standard glossary
- Produce metadata in the languages of your primary users

Selecting the reference metadata

The standard schemas discussed here all have model lists of metadata elements to choose from. Here are examples of user-oriented elements that could be included. Other elements may be needed for internal data management or to document particular types of data. The SDMX Global Metadata Concept Scheme, which provides definitions and code names for reference

metadata, can be used as a check list. Not all elements need to be included. In constructing the template, some elements may be marked as required and others as optional.

Metadata topic	Metadata elements
Indicator identification	Long name Short name Indicator code
Coverage	Geographic coverage Population coverage Sector coverage Time coverage
Methodology	Data description Data collection method Periodicity Classification systems Statistical concepts and definitions Limitations and exceptions to standards
Statistical processing	Adjustments to data Aggregation method Imputation method
Dissemination	Date of last update or release Date of next release Alternate sources and formats
Contact information	Contact organization Contact email address
License terms	Short form of license (e.g. CC-BY 4.0) Link (URL) to full license terms

Maintaining metadata

Updates and revisions to metadata need to be included in the regular cycle of database maintenance. Care should also be taken to maintain cross-domain consistency. If, for example, a code list is revised for one data set, it should be revised for all other data sets using the same codes.

Publishing the metadata

Ideally metadata should be available in machine readable formats through an online data portal. Systems using SDMX can publish their data and metadata to an SDMX registry. But when

this is not possible, metadata may be published in the form of spreadsheets or PDF documents. The availability of metadata should be announced on the same websites or data portals used to disseminate data.

6. Summary and key takeaways

- Metadata are an important component of open data because they provide public information that describes the data and permits the appropriate use and sharing of data.
- Metadata are often described by their purpose as structural, reference (sometimes called descriptive), or administrative.
    - Structural metadata describe the dimensions of the data set and attributes such as scale factors.
    - Reference metadata provide information needed to use the data. They describe how and where the data were collected; adjustments or revisions that have been made; they may also provide information about uses and applications of the data.
    - Administrative metadata provide information about data management and the legal responsibilities for managing and using data. They are important for maintaining a record of institutional arrangements for producing and disseminating data and defining policies for data dissemination.
- An important element of the administrative metadata is the data license or terms of use. For open data, the data license must satisfy the terms of the Open Definition.
- A metadata schema is a standardized method for recording and disseminating information about a set of data. Some important international metadata schemas are the Dublin Core, Digital Data Initiative (DDI), and Statistical Data and Metadata Exchange (SDMX).
- Standard metadata schemas provide names (tags) and definitions for metadata items. The use of standard vocabularies and code lists facilitates the exchange of data and language independence.
- Metadata schemas can be customized to accommodate the characteristics of local data sets. More important than the use of a standard schema is providing uses with well-organized metadata that fully describe the data.