सत्यमेव जयते

Government of India

Ministry of Statistics and Programme Implementation

Data Dissemination: National Metadata Structure (NMDS) For Statistical Products

August 2021

NSO INDIA

Data for Development

Government of India

Ministry of Statistics and Programme Implementation

Policy Implementation & Monitoring Division (PIMD)

Sardar Patel Bhawan, 4th Floor

Sansad Marg, New Delhi – 110001

Ph: (011) 23341867

Data Dissemination: National Metadata Structure (NMDS) For Statistical Products

August 2021

Introduction

National Statistical Office (NSO), Ministry of Statistics & Programme Implementation, presents and disseminates data and metadata through different products like Census data (Economic Census); Survey data such as NSS Surveys viz. Household Surveys, etc, Annual Survey of Industries (ASI), Consumer Price Indices (CPI), and macro-economic aggregates like National Income, Index of Industrial Production (IIP). In addition, statistical data is presented in analytical publications such as NSS Reports, Annual Survey of Industries Reports, National Indicator Framework (NIF) for monitoring SDGs, Energy Statistics, EnviStats India, Women & Men in India etc., which provide analysis of data, supported by the visual presentation of that data in the form of graphs and maps.

The production of data and presentation of metadata structure requires an overview of the arrangements, technical infrastructure and skills required for a holistic and integrated approach to the presentation and dissemination of statistical data and metadata to different user groups. National Metadata Structure (NMDS) is to provide guidelines for the data producer to adhere to a basic minimum quality standard in order to establish and maintain the quality of data and enhance ease in sharing data. The specific objectives of this document are:

- to promote reporting for each type of statistical process and its outputs across different Ministries/Divisions/Departments of NSO, hence facilitating comparisons across processes and outputs;
- to ensure that producer reports contain all the information required to facilitate identification of quality issues and potential improvements in statistical processes and their outputs; and
- to ensure that user reports contain all the information required by users to assess whether statistical outputs are fit for the purposes they have in mind.

A. What is Metadata?

A.1. Metadata should contain all the information users need to analyse a dataset and draw conclusions. It increases data accessibility by summarizing the most important information (i.e. methodology, sampling design, interview mode, etc.) required for analyzing a dataset which alleviates the need for users to search for supporting documents and reports. Furthermore, good metadata clearly articulates the potential uses for a dataset, preventing potential misuses. Metadata is also a tool for rendering complex microdata structures into something meaningful, navigable, and user-friendly. Finally, the adoption of well-known metadata schemas and vocabularies allows for semantic interoperability.

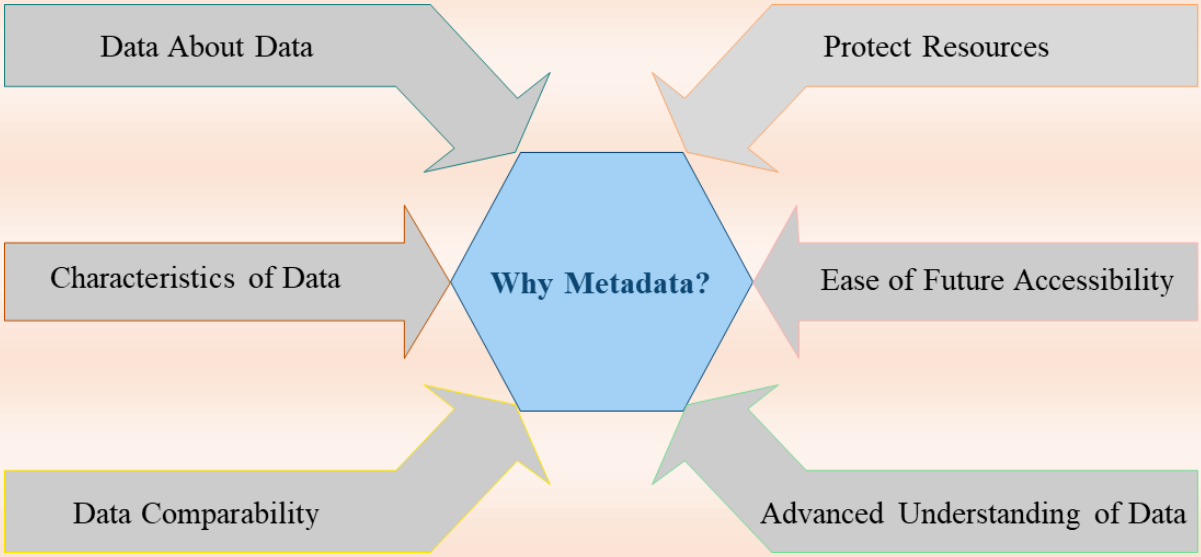
The Metadata process is fully integrated in the Generic Statistical Business Process Model[1] (GSBPM) which has metadata as one of the key elements in the version 5.1.

[1] UNECE: United Nations Economic Commission for Europe, https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

B. Why Metadata?

B.1. In most information technology usages, the prefix of meta conveys "an underlying definition or description." So it is that, at its most basic, metadata is data about data. More precisely, however, metadata describes data containing specific information like type, length, textual description and other characteristics. Metadata makes it much easier to find relevant data and to use a dataset, users need to understand how the data is structured, definitions of terms used, how it was collected, and how it should be read.

B.2. Metadata is an important way to protect resources and their future accessibility. For archiving and preservation purposes, it takes metadata elements that track the object's lineage, and describe its physical characteristics and behaviour so it can be replicated on technologies in the future.



B.3. In today's modern data driven world and in the era of digital transactions, huge amount of data is generated on real time basis, and lately, a large number of organisations/agencies have started producing data, the quantum of which is huge, and thus arises a need of standard regulatory framework to be laid down to assure the quality of data produced by different producers. It will also serve the purpose of ensuring data comparability across time horizons so as to enable better understanding of different social and economic movements.

B.4. Although metadata may not seem exciting or impressive, the true importance of metadata can never be underrated and hence, is important to take a concerted effort to build sound metadata structure to draw maximum gains from varied data sets.

C. Role of MoSPI in Building Metadata Structure

C.1. MoSPI being a nodal agency for planned development of the statistical system in the country is also responsible for maintaining the highest standards of data quality which adhere to basic guidelines of International Agencies so as to ensure India's statistical system is one of the frontrunners in quality data producer. MoSPI aims at raising the National Statistical System (NSS) to the epitome of being one of the best professionally equipped government data producing agencies by building the best of IT infrastructure among others in the system, and Metadata is one of the building blocks to achieve the objective.

C.2. The document presents the NMDS in two formats- the first one is the indexed version as NMDS concepts (Section F), and the second one presents details of concepts through definition and guidelines (Section G).

D. Metadata Management

D.1 It is advisable to put in place a metadata policy by the official statistical producing agencies, ab initio. The policy is a set of broad, high level principles that form the guiding framework within which metadata management can operate.

D.2 Once the metadata policy is put in place, for an organisation, metadata should be compiled and maintained actively. Otherwise, the currency, and thus use of Metadata will degrade with time. To realise the full capabilities of Metadata, it is necessary that the Metadata are maintained over a long period of time. Even with investment in technically sophisticated search tools, such systems may find little stakeholders acceptance, if the data are incomplete or is not updated regularly.

While preparing the NMDS, the following core principles should be borne in mind:

i. Metadata Handling:
   a. Statistical Business Process Model
   b. Active, not Passive
   c. Reuse for Efficiency
   d. Version Preservations

ii. Metadata Authority
   e. Registration
   f. Single Source
   g. One Entry/Update
   h. Standards Variations

iii. Relationship to Statistical Business Processes
   i. Integrity
   j. Matching Metadata
   k. Describe Flow
   l. Capture at Source
   m. Exchange and Use

iv. Users
   n. Identify Users
   o. Variant Formats
   p. Availability

E. Retention, Preservation, and Destruction

National Statistics constitute valuable and irreplaceable assets whose value can increase through widespread and long-term use. National Statistics should thus be backed by the Data Management Policy eliciting the arrangements it has in place for the retention, long term preservation, and destruction of its resources including Metadata.

F. National Metadata Structure (NMDS) Concepts - Index

G. Details of NMDS Concepts

Item No	Concept name	Definition	Guidelines
1	**Contact**	**Individual or organisational contact points for the data or metadata, including information on how to reach the contact points.**	
1.1	Contact Organisation	The name of the organisation of the contact points for the data or Metadata.	Provide the full name (not just acronym/code name) of the organisation responsible for the processes and outputs (data and metadata) that are the subject of the report
1.2	Compiling agency	Organisation collecting and/or elaborating the data being reported	Provide the full name of the Department/Division under the organisation responsible for the processes and outputs (data and metadata) that are the subject of the report
1.3	Contact Details	The details of the contact points for the data or metadata.	Provide contact details of contact point(s) in following format: a. Name of Organisation owning the processes and outputs b. Author (if different from (a)) c. Disseminating Agency (if different from (a) and (b)) d. Name (first and last names) e. Designation f. Postal address g. email address (preferably designation based) h. Contact number i. Fax number If more than one name is provided, the details of main contact should be indicated. If the author of the report is different from the person(s) responsible for process and its outputs, provide this name also with his/her details

Item No	Concept name	Definition	Guidelines
2	**Statistical Presentation and Description**	**Description of the disseminated data which can be displayed to users as tables, graphs or maps**	
2.1	Data description	Main characteristics of the data set, referring to the data and indicators disseminated.	Describe briefly the main characteristics of the data in an easily and quickly understandable manner, referring to the main variables disseminated.
2.2	Classification system	Arrangement or division of objects into groups based on characteristics which the objects have in common	List all classifications and breakdowns that are used in the data (with their detailed names) and provide links (if publicly available). Type of dis-aggregation available in the data sets - for example rural-urban, male-female, etc. and whether data is available at the sub-national level or not, should be clearly specified.
2.3	Sector coverage	Main economic or other sectors	List the main economic or other sectors covered by the data and the size classes used, for example, Health/ Education/ Manufacturing etc for sectors and classes based on number of employees for size classes
2.4	Statistical concepts and definitions	Statistical characteristics of statistical observations, variables	Define and describe briefly the main statistical variables that have been observed or derived. Indicate their types.
2.5	Statistical unit	Entity for which information is sought and for which statistics are ultimately compiled.	Define the type of statistical unit about which data are collected, e.g. enterprise, household, etc.
2.6	Statistical population	The total population of a defined class of people, objects or events	Define the target population of statistical units for which information is sought. For example, agricultural household, general household, industrial unit, etc.

Item No	Concept name	Definition	Guidelines
2.7	Reference Period	The length of time for which data are available	State the time period(s) for which data is collected
2.8	Data Confidentiality[2]	Rules applied for treating the datasets to ensure statistical confidentiality and prevent unauthorised disclosure.	Describe the procedures that are used in protecting confidentiality, viz., anonymisation, legal provision, if any.
3	**Institutional Mandate**	**Law, set of rules or other formal set of instructions assigning responsibility as well as the authority to an organisation for the collection, processing, and dissemination of statistics**	
3.1	Legal acts and other agreements	Legal acts or other formal or informal agreements that assign responsibility as well as the authority to an agency for the collection, processing, and dissemination of statistics	State the national legal acts and/or other reporting agreements
3.2	Data sharing	Arrangements or procedures for data sharing and coordination between data producing agencies.	Describe the arrangements, procedures or agreements to facilitate data sharing and exchange between data producing agencies within the national statistical system

[2] All statistical information published by any agency shall be arranged in such a manner so as to prevent any particulars becoming identifiable by any person (other than the informant by whom those particulars were supplied) as the particulars relating to the informant who supplied it, even through the process of elimination (Source: Collection of Statistics Act, 2008).

Item No	Concept name	Definition	Guidelines
3.3	Release policy	Rules for disseminating statistical data to all interested parties	State if the release of the products is governed by some policy etc.
3.4	Release calendar	The schedule of statistical release dates.	State whether there is a release calendar for the statistical outputs from the process being reported, and if so, whether this calendar is publicly accessible and if yes, give a link or reference.
3.5	Frequency of dissemination	The time interval at which the statistics are disseminated over a given time period.	State the frequency with which the data are disseminated, e.g. monthly, quarterly, yearly.
3.6	Data access	The conditions and modalities by which users can access, use and interpret data	State the conditions and link on website from where the user can access the data For easy access of users, following details should also be mentioned about the dataset: **Title:** Name by which the data is known **Dataset Edition:** Edition of data (ex: first, second, final etc) **Dataset Reference data type:** Type of data entered in the field (ex: .txt, .dbf, .xls) **Presentation Format:** Presentation format of the data (ex: document, map, table, etc.) **Dataset Language:** language of any text in the data **Status/Version:** How updated is the data?

Item No	Concept name	Definition	Guidelines
4	**Quality Management**	**Systems and frameworks in place within an organisation to manage the quality of statistical products and processes.**	
4.1	Documentation on methodology	Descriptive text and references to methodological documents available.	List reference metadata files, methodological papers, summary documents and handbooks relevant to the statistical process.
4.2	Quality documentation	Documentation on procedures applied for quality management and quality assessment.	List relevant quality related documents, for example, other quality reports, studies.
4.3	Quality assurance	All systematic activities implemented that can be demonstrated to provide confidence that the processes will fulfil the requirements for the statistical output.	Describe the procedures (such as use of a general quality management system based on ISO 9000 series) to promote general quality management principles in the organisation. Describe the quality assurance framework used to implement statistical quality principles. Describe the quality assurance procedures specifically applied to the statistical process for which the report is being prepared, for example training courses, process monitoring, benchmarking, assessments, and use of best practices. Describe any ongoing or planned improvements in quality assurance procedures.

Item No	Concept name	Definition	Guidelines
4.4	Quality assessment	Overall assessment of data quality, based on standard quality criteria.	Summarise the results of the most recent quality assessments and cross reference to the chapters in the report where the results are presented in more detail.
5	**Accuracy and Reliability**	**Accuracy of data is the closeness of computations or estimates to the exact or true values that the statistics were intended to measure. Reliability of the data, defined as the closeness of the initial estimated value to the subsequent estimated value.**	
5.1	Sampling error	That part of the difference between a population value and an estimate thereof, derived from a random sample, which is due to the fact that only a subset of the population is enumerated.	If probability sampling is used: • for user reports-provide the range of variation of the A1[3] indicator among key variables at user report level of detail; • for producer reports-provide the range of variation of the A1 indicator among key variables at producer report level of detail; • indicate the impact of sampling error on the overall accuracy of the results; • state how the calculation of sampling error is affected by adjustments for nonresponse, misclassifications and other sources of uncertainty, such as outlier treatment. If non-probability sampling is used: provide an assessment of representativeness, a motivation for the invoked model for estimation and risk of sampling bias

[3] In international terminology, A1 is simply used for Sampling Errors. In other words, indicator for Sampling Errors is denoted by A1. For further details, refer to: https://ec.europa.eu/eurostat/documents/64157/4373903/02-ESS-Quality-and-performance-Indicators-2014.pdf/5c996003-b770-4a7c-9c2f-bf733e6b1f31

Item No	Concept name	Definition	Guidelines
6	**Timeliness**	**The timeliness of the data collection release to be compiled.**	
6.1	Timeliness	Length of time between data availability, the event or phenomenon the data describe, and final release to its users.	Outline the reasons for the time lag, if any. Outline efforts to reduce time lag in future.
7	**Coherence and Comparability**	**Adequacy of statistics to be reliably combined in different ways and for various uses and the extent to which differences between statistics can be attributed to differences between the true values of the statistical characteristics**	
7.1	Comparability – over time	The extent to which statistics are comparable or reconcilable over time	Provide information on possible limitations in the use of data for comparisons over time. Distinguish three broad possibilities: 1. There have been no changes, in which case this should be reported. 2. There have been some changes but not enough to warrant the designation of a break in series. 3. There have been sufficient changes to warrant the designation of a break in series.
7.2	Coherence	The extent to which statistics are reconcilable with System of National Accounts	For producer reports only. Where relevant, the results of comparisons with the System of National Account framework / Other Statistical Standards and feedback from System of National Accounts / Other Statistical Standards with respect to coherence and accuracy problems should be reported and should be a trigger for further investigation.

Item No	Concept name	Definition	Guidelines
8	**Statistical Processing**	**Any statistical processing undertaken to finalise the data**	
8.1	Source data type	Characteristics and components of the raw statistical data used for compiling statistical aggregates.	Indicate if the data are based on a survey, administrative data, multiple data sources, or macro-aggregates. In the event of multisource or macro-aggregate processes describe each source. For each dataset from an administrative source, summarise the source, its primary purpose, and the most important data items acquired
8.2	Frequency of data collection	Time interval at which the source data are collected	Indicate the frequency of data collection (e.g. monthly, quarterly, annually, or continuous).
8.3	Data collection method	Method applied for gathering data for official statistics	For each source of survey data: • describe the method(s) used to gather data from respondents; • annex or hyperlink the questionnaires(s). For each source of administrative data: • describe the acquisition process and how it was tested. For all sources: • describe the types of checks applied at the time of data entry.

Item No	Concept name	Definition	Guidelines
8.4	Data validation	Process of monitoring the results of data compilation and ensuring the quality of statistical results	Describe the procedures for checking and validating the source data and how the results are monitored and used. Describe the procedures for validating the aggregate output data (statistics) after compilation, including checking coverage and response rates, and comparing with data for previous cycles and with expectations. List other output datasets to which the data relate and outline the procedures for identifying inconsistencies between the output data and these other datasets
8.5	Data compilation	Operations performed on data to derive new information according to a given set of rules.	Describe the procedures for imputation, the most common reasons for imputation and imputation rates within each of the main strata. Describe the likely impact of imputation. Describe the procedures to derive new variables and to calculate aggregates and complex statistics. Describe the procedures for adjustment for non-response and the corrections to the design weights to account for differences in response rates. Describe the calculation of design weights, including calibration (if used). Describe the procedures for combining input data from different sources.

Item No	Concept name	Definition	Guidelines
			Provide the ratio of the number of replaced values to the total number of values for a given variable. Specific reference to formula shall be made. The formula or mathematical equation used while computing different variables in the report may be described here in a structured format showing the Numerator; Denominator and Multiplier used for computing the same
9	**Metadata Update**	**The date on which the metadata element was inserted or modified in the database.**	
9.1	Metadata last posted	Date of the latest dissemination of the metadata	The date when the complete set of metadata was last disseminated as a block should be provided (manually, or automatically by the metadata system).
9.2	Metadata last update	Date of last update of the content of the metadata.	The date when any metadata were last updated should be provided (manually, or automatically by the metadata system).