NITI Aayog

DMEO
DEVELOPMENT MONITORING AND EVALUATION OFFICE

ADMINISTRATIVE DATA TOOLKIT

CLEAR
South Asia Center

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
SOUTH ASIA AT IFMR

NITI Aayog

DMEO
DEVELOPMENT MONITORING AND EVALUATION OFFICE

CLEAR
South Asia Center

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
SOUTH ASIA AT IFMR

NITI Aayog

DMEO DEVELOPMENT MONITORING AND EVALUATION OFFICE

ADMINISTRATIVE DATA TOOLKIT

CLEAR South Asia Center

J-PAL ABDUL LATIF JAMEEL POVERTY ACTION LAB SOUTH ASIA AT IFMR

The authors are Dr. Radha R. Ashrit (Dy. DG), Ms. Gujan Saini (Research Officer) and Ms. Ankita Gupta (Young Professional) at DMEO, NITI Aayog and Ms. Bhavya Pandey, Mr. Mayank Bhushan, Mr. Evan Williams, Ms. Harini Kannan, Ms. Mrignaina Tikku, Mr. Nishant Lodha, Mr. Ramakrishnan N, Ms. Sohini Mookherjee, and Ms. Aparna Krishnan at CLEAR/J-PAL South Asia.

Suggested Citation: Administrative Data Toolkit (2024), Development Monitoring and Evaluation Office, NITI Aayog

NITI Aayog

DMEO
DEVELOPMENT MONITORING AND EVALUATION OFFICE

CLEAR
South Asia Center

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
SOUTH ASIA AT IFMR

Table of Contents

NITI Aayog

DMEO
DEVELOPMENT MONITORING AND EVALUATION OFFICE

CLEAR
South Asia Center

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
SOUTH ASIA AT IFMR

Acknowledgements

The Toolkit was a joint effort by DMEO, NITI Aayog and CLEAR/J- PAL South Asia based on secondary research as well as inputs provided by key stakeholders.

This work would not have been possible without the incessant efforts put in by the team at DMEO, NITI Aayog especially Dr. Radha R. Ashrit (Dy. DG), Ms Gujan Saini and Ms Ankita Gupta for providing valuable inputs and suggestions to structure the report and facilitating State department interactions. We would like to thank CLEAR/J-PAL South Asia for collating all the information, providing technical expertise on the subject, and writing the report, especially Ms. Bhavya Pandey, Mr. Mayank Bhushan, Mr. Evan Williams, Ms. Harini Kannan, Ms. Mrignaina Tikku, Mr. Nishant Lodha, Mr. Ramakrishnan N, Ms. Sohini Mookherjee, and Ms. Aparna Krishnan.

Nidhi Chhibber

Director General
Development Monitoring and Evaluation Office
NITI Aayog, New Delhi 110001

Preface

Administrative data are those that institutions routinely collect data for operational purposes such as administering public services, rather than for a specific research objective. Administrative data could be used effectively for evidence-based policymaking, as it provides access to comprehensive information over a long period, greater accuracy of research insights, and a faster turnaround of analyses and results. To this end, this toolkit is intended to act as a resource for enabling the effective and safe use of administrative data for evidence-based policy-making by strengthening administrative data systems and processes. The toolkit and the associated checklists are intended for the audience of the wider central and state government stakeholders like departments and line ministries that collect and digitally store administrative data and are interested in enabling its access and use for research and analysis purposes for more effective policy making.

In this toolkit, we use the lifecycle of the administrative data in a digital information system as a framing to describe processes and checklists. A data life cycle is the order of stages data go through in an information system — from its initial generation or collection to its archival or destruction.

This practical toolkit is an accompaniment to the "Compendium of Case Studies on Using Administrative Data for Evidence-Based Policy Making" which summarises the advantages of using administrative data and discusses some of the best practices that can ensure better quality and usability of administrative data. It also highlights via case studies and examples how government institutions have adopted some of these best practices and developed effective and replicable solutions to use administrative data.

The following key sections of the toolkit may enable the readers to set up actionable processes and checklists while working with any administrative data:

1. **Data Quality -** The quality of data reflects the fitness of data for usage for purposes of analysis and interpretation of insights. It entails the following dimensions – relevance, accuracy, timeliness, accessibility, interpretability, and coherence[1].
2. **Data Handling -** Data handling ensures that data are stored, archived or disposed of safely and securely before, during, and after its usage.
3. **Data Security** - Data security is critical for the protection of confidential data[2] and prevention of breach of information or its misuse towards unintended objectives.
4. **Data Privacy** - Maintaining data privacy protects personal or sensitive information while enabling data access by others (beyond those involved in its collection) by using techniques such as de-identification of data, masking, anonymisation or pseudonymisation[3].

Stages of a data life cycle

The above themes of data standards (i.e. data quality, handling, security and privacy) may not be executed sequentially. They are interconnected and need to be executed parallelly during the lifecycle of an administrative dataset. The data lifecycle comprises five phases in general:

[1] Source: Brackstone, Gordon. 1999."Managing Data Quality in a Statistical Agency." Survey Methodology, Vol. 25, No. 2, pp. 139-149 Statistics Canada,United Nations Statistics Division. December 1999. Accessed at https://unstats.un.org/unsd/dnss/docViewer.aspx?docID=194#start .

[2] Reference: O'Toole, Elisabeth, Kenya Heard, Laura Feeney and Rohit Naimpally. 2015. "Data security procedures for researchers." J-PAL North America. Last Modified March 2023. Accessed at https://www.povertyactionlab.org/resource/data-security-procedures-researchers .

[3] Reference:Guidance Note. "Data Privacy, Ethics And Protection : Guidance Note On Big Data For Achievement Of The 2030 Agenda." United Nations Development Group. n.d. Accessed at https://unsdg.un.org/sites/default/files/UNDG_BigData_final_web.pdf .

Collection, Storage, Usage, Sharing and Archiving. It is important to note that the overall efficacy of the administrative data depends on how the four cross-cutting data standards are executed across each of the five phases of the data lifecycle.

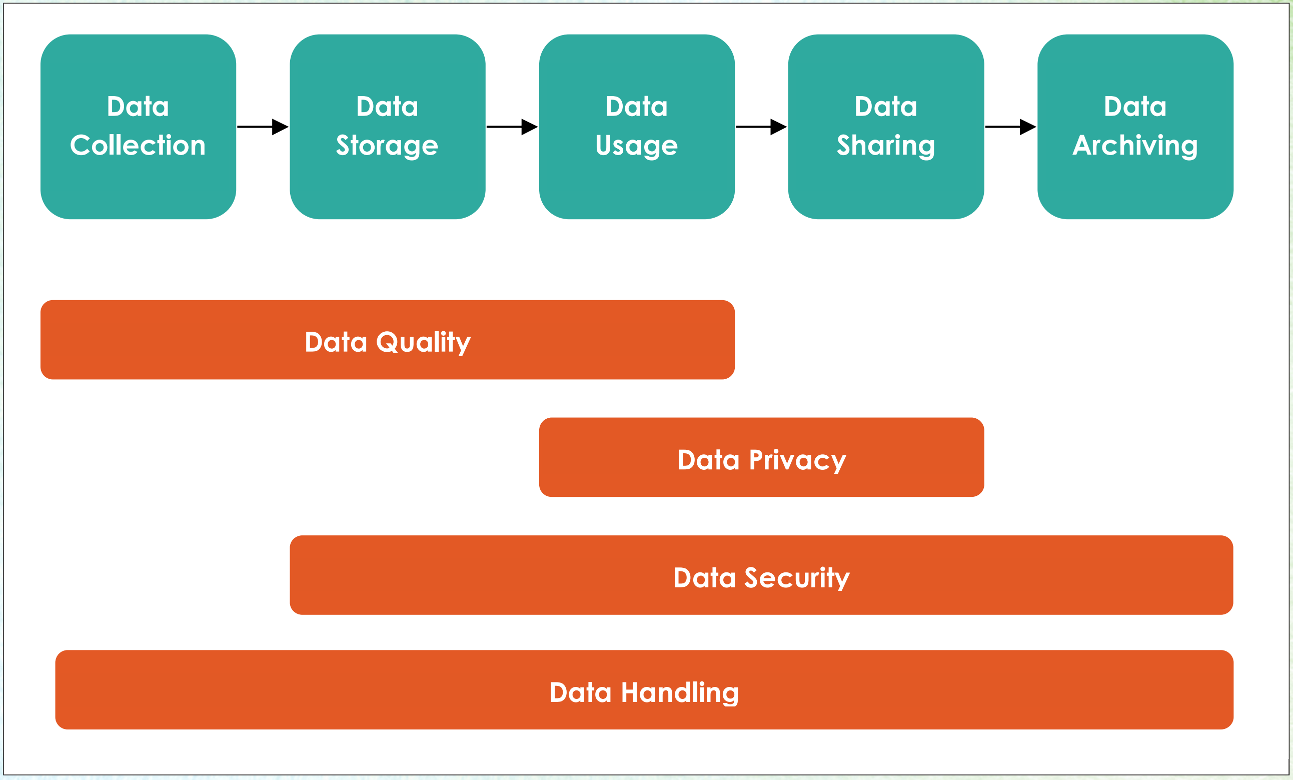


Figure 1: Data standards executed across each of the five phases of the data lifecycle

The first phase of the data life cycle is the collection or capture of data. At this stage, it is important to document how the data are collected. Metadata standards should be uniformly applied and information on transformations of data should be well-documented. The application of data standards is crucial at this stage as any challenges to the robustness of a dataset can be eliminated at this stage through following thorough protocols. Since administrative data contains identifiers such as personal information, socio-economic markers, or other sensitive information, it is essential that security and privacy measures are followed at this stage.

In order to be used effectively for statistical analysis or informing policy decision-making, administrative data must undergo processes that retain its quality and make it suitable for analysis such as outlier detection and linking of datasets. At this stage, ensuring the safe and appropriate usage of data remains an important process consideration.

Inter-agency data sharing can enable new and innovative uses of data, both within and beyond administrative systems. It is important to share data in a way that protects privacy and confidentiality while making the data useful to inform decision-makers.

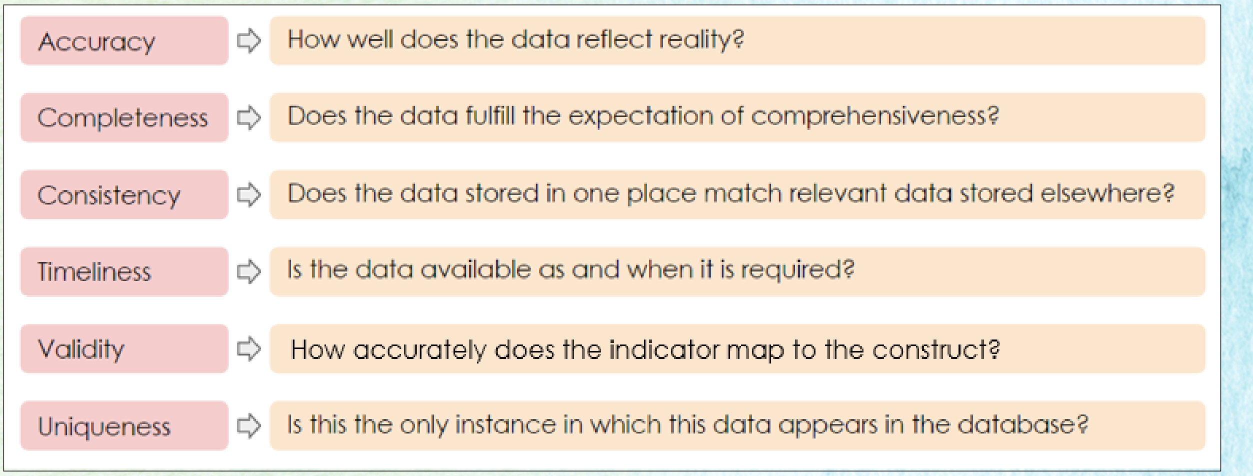
Data archiving is the stage at which data that is no longer actively used, is catalogued for long-term retention. At this stage, it's important to have strong security and privacy measures in place. When data agencies create high-quality policies and practices that govern the various phases of the administrative data life cycle, they can be confident they are on the right path to effectively and safely utilise administrative data to answer critical stakeholder questions and to inform decisions to support continuous improvement.

Section 1
Data Quality

1.1 Data Quality Dimensions

The effective use of administrative data rests on maintaining data quality. This entails maintaining accurate records and consistent processes on how the data are collected, collated, cleaned and cross-checked, independent data audits to ensure high accuracy, and ensuring safeguards in places where subjects have incentives to misreport information by implementing additional validation checks. Further, data quality standards can enhance operational and statistical uses of administrative data.

A high-quality dataset should adhere to the following dimensions of data quality:



Accuracy	⇨	How well does the data reflect reality?
Completeness	⇨	Does the data fulfill the expectation of comprehensiveness?
Consistency	⇨	Does the data stored in one place match relevant data stored elsewhere?
Timeliness	⇨	Is the data available as and when it is required?
Validity	⇨	How accurately does the indicator map to the construct?
Uniqueness	⇨	Is this the only instance in which this data appears in the database?

Figure 2: Dimensions of Data Quality

A detailed description of the data quality dimensions is provided below:

1. Accuracy

Accuracy refers to the closeness of the administrative record data values to their (unknown) true values. It is important to check for information on any known sources of errors in the administrative data such as missing records, missing values of individual data items, misinterpretation of questions, and keying, coding, and duplication errors.

2. Completeness

Data completeness refers to the comprehensiveness of the data. For example, in the case of the delivery of a welfare programme, it is important to ensure that all the required data for the identification and delivery of schemes and programmes is available in that dataset.

3. Consistency

Consistency refers to the overall reliability of the data. It is imperative to ensure that the dataset produces similar results under consistent conditions (such as when a type of analysis is repeated).

Some errors to look out for while keeping a check on the consistency of the data are:

» **Data collection errors**: The procedure involved in primary data collection is prone to errors of misreporting since the staff manually enters the respondents' information or response into the system. There could be errors in spelling, incorrect addresses, incomplete information etc.

» **Incentives to misreport:** In the case that the performance of frontline workers is tied to outputs or outcomes such as school attendance, immunisation completion, thresholds of malnourishment indicators, agriculture output etc., the staff in charge of service delivery may have an incentive to over-report or under-report outcomes based on the scheme benefits.

4. Timeliness

Timeliness refers to how well the administrative data meets the needs of the user at the time of need, as well as how up-to-date the data are.

5. Validity

Validity is how accurately an indicator maps to the construct. A construct is a theoretical concept, theme, or idea which is not easy to measure. To measure this empirically, one needs to use "proxies" i.e. indicators. An indicator is "valid" if it can be measured accurately and without bias. For instance, cortisol levels may be an example of an unbiased measure of stress, where stress levels would be the construct but the stress test is still very noisy. While we think about the mapping of the indicator onto a construct, we would refer to this as the validity of the measure. Another example of validity (measures what it claims to measure) is a test of intelligence i.e. the construct, which should measure only intelligence and not something else (such as memory).

6. Uniqueness

Uniqueness of data means that there's only one instance of it appearing in a database. A common problem in databases is data duplication. To meet this data quality dimension, it is important to review the information to ensure that none of it is duplicated.

1.2 Types of data quality checks

In order to ensure a high quality of administrative data being collected, it is imperative to incorporate quality checks at the time of data collection, as well as post-data collection. It is important to note that a significant amount of administrative data is collected using technology, where the data collection is either conducted by frontline workers actively or passively through machines. Researchers have used a variety of checks to ensure the data they collect is of good quality. The checks we describe below borrow from the checks conducted on survey data. The type of data quality checks conducted for administrative data are mentioned in detail below:

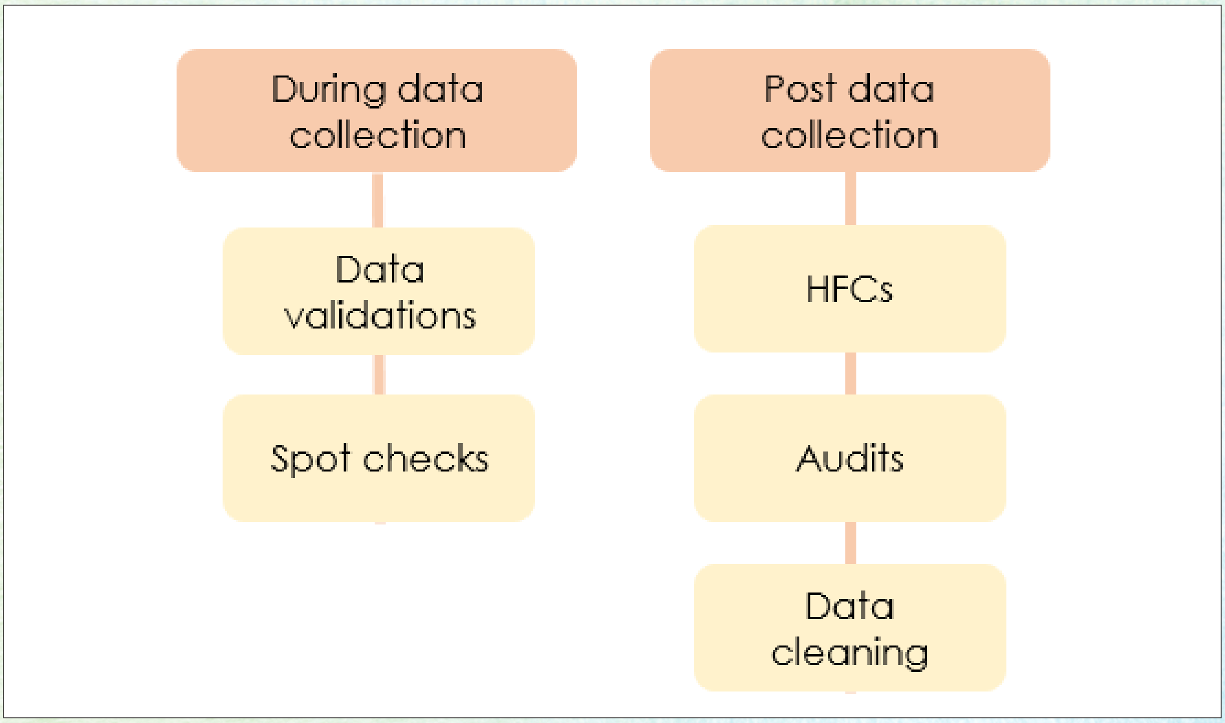


Figure 3: Types of Data Quality Checks

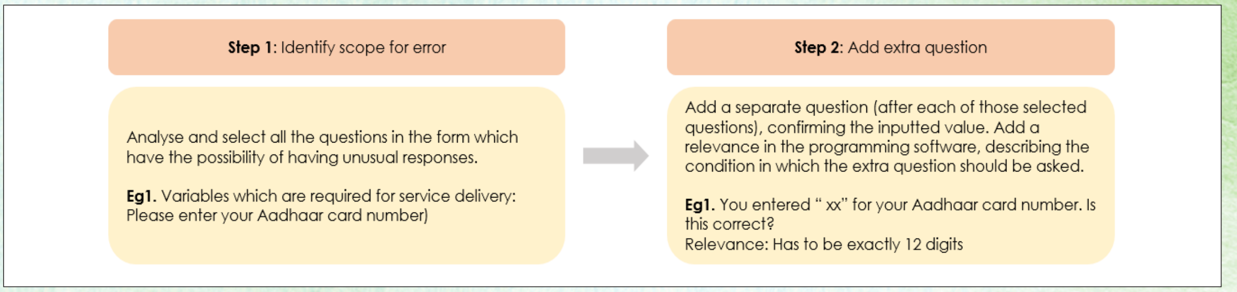
1.3 Pre-data collection checks

1. Data validations

The first iteration of data quality checks at the point of data collection can be in-built within the data collection application to reduce the possibility of human error. The different types of validations one can incorporate are as follows:

a) Soft constraints:

These constraints can be used to double-check a response in cases where the responses for a particular variable can be unusual but not impossible. For example, if the input for a variable regarding the date of birth or Aadhaar card number is higher than a certain number of digits. The following steps could be followed to incorporate a soft constraint in the form:



Step 1: Identify scope for error

Analyse and select all the questions in the form which have the possibility of having unusual responses.

Eg1. Variables which are required for service delivery: Please enter your Aadhaar card number)

Step 2: Add extra question

Add a separate question (after each of those selected questions), confirming the inputted value. Add a relevance in the programming software, describing the condition in which the extra question should be asked.

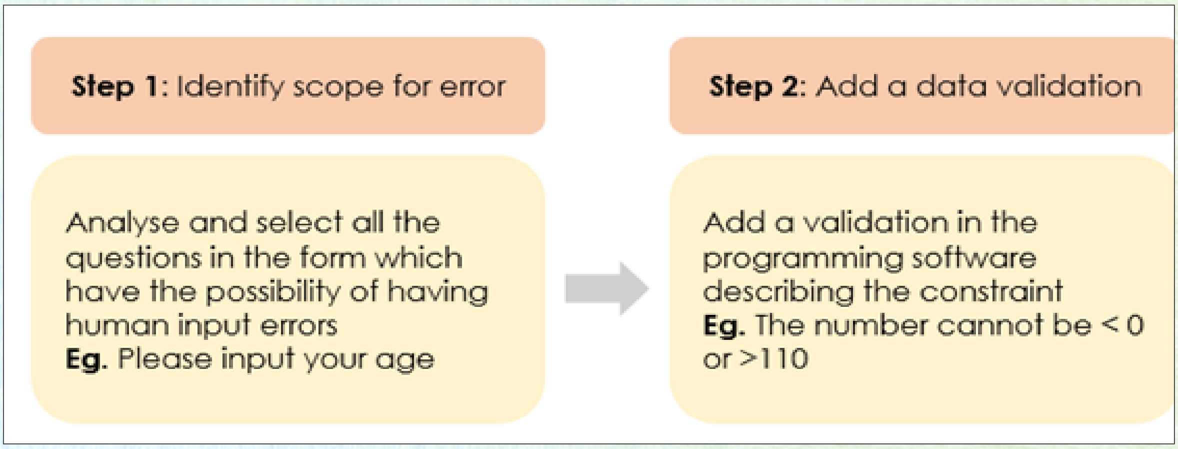
Eg1. You entered " xx" for your Aadhaar card number. Is this correct?
Relevance: Has to be exactly 12 digits

Figure 4: Steps to incorporate soft constraint

b) Hard constraints

These constraints can be incorporated into a form to prevent responses that are impossible or extremely unlikely. For example, if age has been reported as less than zero.

The following steps could be followed to incorporate a hard constraint in the form:



Step 1: Identify scope for error

Analyse and select all the questions in the form which have the possibility of having human input errors
Eg. Please input your age

Step 2: Add a data validation

Add a validation in the programming software describing the constraint
Eg. The number cannot be < 0 or >110

Figure 5: Steps to incorporate hard constraint

2. Spot Checks

Generally, there are two ways in which data might be collected on the field:

a. The field team or implementation team might reach out to targeted individuals or households and directly collect data and input them using a digital platform or using paper-based records e.g. ANMs collecting data at Anganwadis, personnel collecting data for a national level census survey, this is known as active data collection.

b. Another type of data collection happens where citizens or individuals may input or enter data using a digital platform (app or portal) e.g. for applying for citizen services eSewa, the citizens register or apply for a particular service. This is also known as passive data collection.

Spot-checks are unanticipated visits by field staff (e.g. cluster resource coordinators or health visitors) to verify whether the data collection/data recording is happening when and where it should be, at the time of service delivery. Such checks are particularly useful in the context of administrative data where there is an instance of active data collection.

The following things should be kept in mind when conducting spot checks:

a. Ensuring the appropriate usage of data collection tools is in place and following the required compliance of the data collection processes.

b. Conducting spot checks on a higher percentage of processes at the initial stages of deployment of new hardware/software/data collection formats to catch errors early, and then to decrease the percentage checked over time.

For example, in the context of administrative data collection, spot-checks can be particularly useful to check the data collection at points of service delivery such as provision of immunisation services, Public Distribution Systems (PDS), etc.

1.4 Post-data collection checks

1. High-Frequency Checks (HFCs)

In order to ensure high quality of data, HFCs are routine checks made on incoming data, ideally daily, to check for data irregularities.

Objectives of conducting HFCs:

a. To keep a check on the accuracy of the data collected, for example by looking for information on any known sources of errors in the data.

b. To monitor the progress of data processes, and to measure the performance of data recording.

c. To check on the possibility of data fraud in administrative systems.

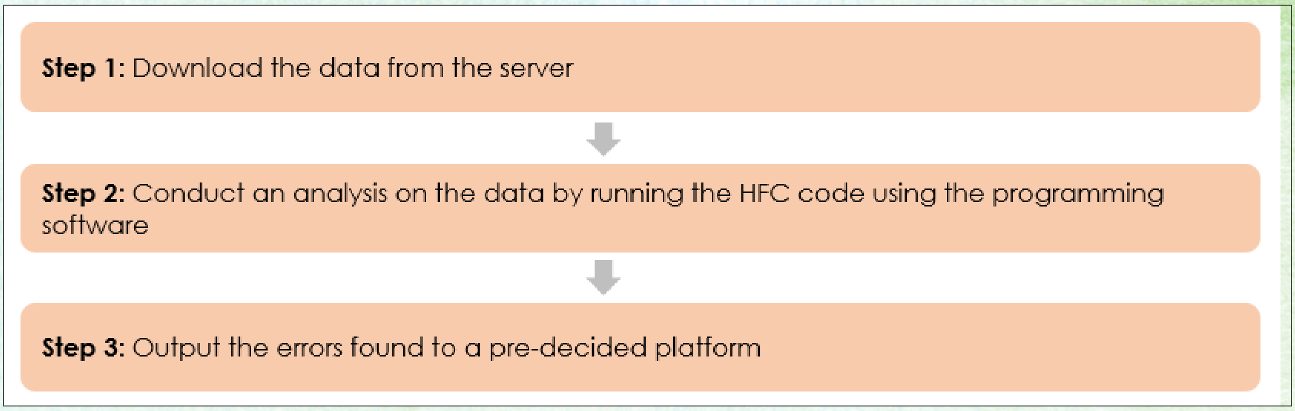
Before the data collection/data recording begins, a few steps should be taken to set up the HFC process:

a. Identify and create a list of the types of checks that would need to be conducted. A detailed list has been provided below, for reference.

b. Analysing the data and creating dashboards to monitor these checks, this code would be run, to analyse the data every time the HFC is conducted.

c. Decide on how the errors that would be detected, after running the HFC, should be outputted. HFCs are run throughout data collection. It is important to decide on the frequency of conducting the HFCs. HFCs should be run on a monthly basis, weekly basis, or every two days based on the size of the sample.

It is advisable to complete the setup of the HFCs before the data collection process begins. Once the data collection begins, the HFC should be run on a daily or bi-weekly basis. The following steps can be followed at the time of conducting the HFC each time:



Step 1: Download the data from the server

Step 2: Conduct an analysis on the data by running the HFC code using the programming software

Step 3: Output the errors found to a pre-decided platform

Figure 6: Steps for running HFC

HFC Checklist

1. Check for missing data
   a. Number and type of variables which have all missing values
   b. % of missing observations in remaining variables
2. Check for duplicates
   a. List of duplicate IDs which need to be rectified
   b. Reasons for duplicates arising e.g. Incorrect spelling, number written
3. Check if there is any inaccuracy or outliers in the data
   a. If the values in some entries are drastically higher or lower than the average entry value
4. Check for the distribution of missing values and responses to skip-order questions
   a. If all valid skips have been compiled and there are no invalid skips
   b. If the entries are logically consistent (e.g. age must increase with time)
5. Completeness/validity of variables
   a. Validity - blank and other invalid entries e.g. Phone numbers such as, '9999999999'
   b. If the field implementers are not forging data e.g. If they are entering their phone numbers at the time of service delivery instead of the service receiver's phone number.
6. If the data are comparable with similar data collection contexts (e.g. the average household income in this data should not be drastically different from the average household income reported in another data source)

Box 1: HFC Checklist

2. Backchecks or Data audits

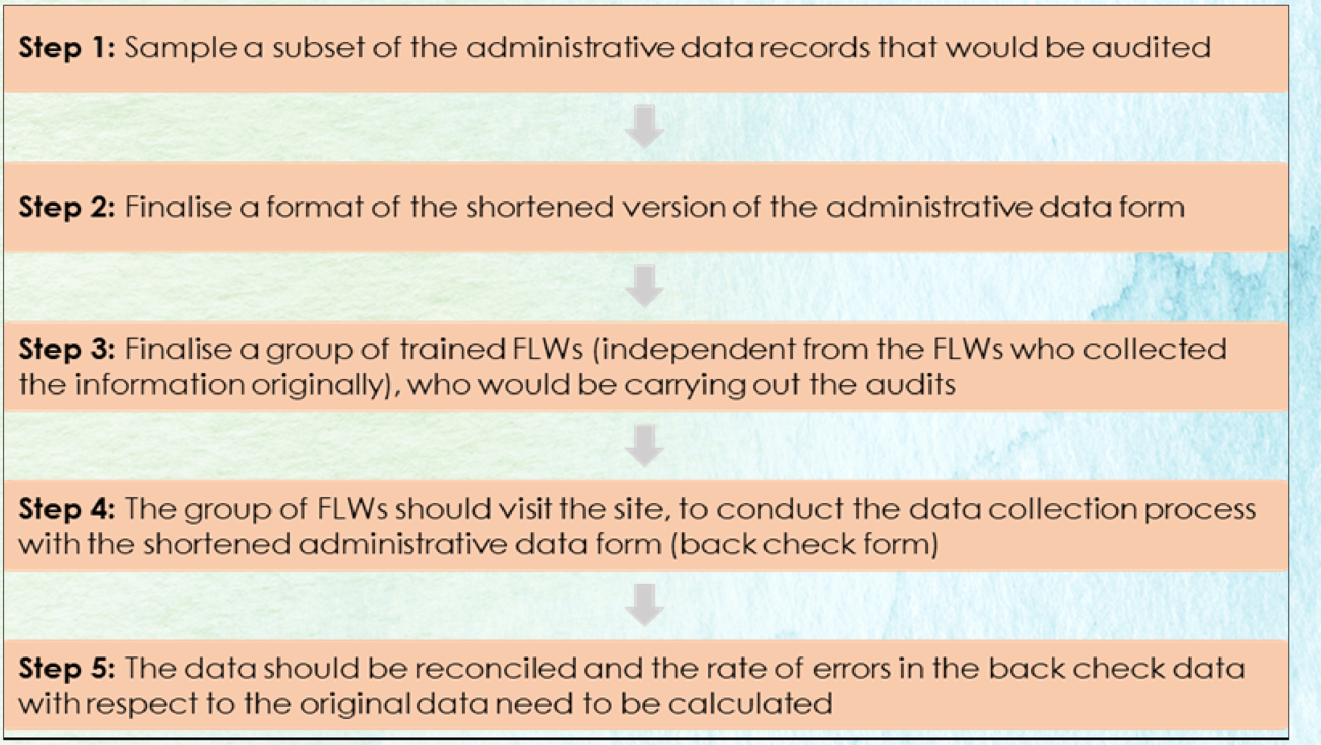
In order to ensure that accurate data is being collected, it is imperative to set up a system of field audits to verify the authenticity of collected information for a subset of records. These audits should be carried out by a group of trained front-line workers (independent of front-line workers who collected the data originally), to revisit the place of data collection and verify the information which was collected.

The data collection format for the backcheck should be a shortened version of the original administrative data format which was used. Approximately 10-20% of the form entries should be randomly assigned to be backchecked. While finalising the shortened version of the administrative data form for the backcheck, the following variables should be included:

1. **Type 1 variables:** These are generally demographic questions such as about marital status, education level etc. which usually do not allow room for any error. Discrepancies between the original form entry and the back-check form entry indicate poor-quality data.
2. **Type 2 variables:** These are variables which are key outcome variables for the programme in question. Examples of such variables would be recall questions such as "Did the child get immunised" or "Did you go for an ANC visit?". Discrepancies between the original form entry and the back-check form entry indicate poor-quality data.

The option of back-checks and audits may not be relevant for all types of administrative data. For example, in the case of income tax forms, the salaried income entered in the form can be corroborated with the income statements from the employers. However, in many cases backchecks and audits are relevant for validating the same information from different sources to maintain data accuracy and for use as a tool for regulation of administrative processes using data.

The following steps can be followed to carry out an audit of an administrative data set:



Step 1: Sample a subset of the administrative data records that would be audited

Step 2: Finalise a format of the shortened version of the administrative data form

Step 3: Finalise a group of trained FLWs (independent from the FLWs who collected the information originally), who would be carrying out the audits

Step 4: The group of FLWs should visit the site, to conduct the data collection process with the shortened administrative data form (back check form)

Step 5: The data should be reconciled and the rate of errors in the back check data with respect to the original data need to be calculated

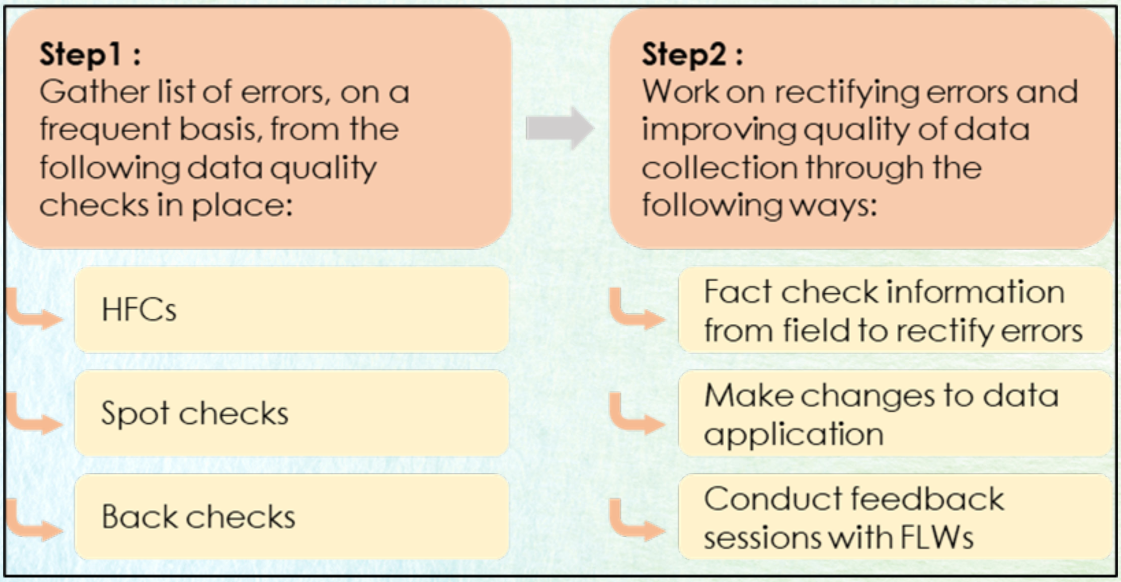
Figure 7: Steps for conducting audit of Administrative data set

1.5 Progress monitoring, feedback into processes, and data flow

Through this process of conducting timely spot-checks, HFCs and back-checks on the data being collected, an accurate list of errors in the data is identified.

As and when these errors are identified, there are various methods to rectify them while improving the overall quality of the data collection process:

1. Procure information from the field to be able to accurately correct the errors arising in the data.
2. Conduct feedback sessions with the front-line workers, to understand their challenges during data collection and explain to them the errors arising to reduce mistakes going forward.
3. Make relevant changes to the data collection application if that may prevent some common errors from arising going forward.



Step1 : Gather list of errors, on a frequent basis, from the following data quality checks in place:

- HFCs
- Spot checks
- Back checks

Step2 : Work on rectifying errors and improving quality of data collection through the following ways:

- Fact check information from field to rectify errors
- Make changes to data application
- Conduct feedback sessions with FLWs

Figure 8: Steps for running HFC

It is important to note that the above checks are primarily applicable for active administrative data collection. For passive data collection, only HFCs and subsequently making changes to the data application would be imperative.

Section 2

Metadata documentation and accessibility

Administrative data can be an excellent source of information for use in research and evaluating the impact of programmes. The purpose of this section is to provide a step-by-step guide for documenting administrative datasets in a standardised format called a metadata catalogue. This format can be self-administered by staff within the department or the agency or compiled by non-departmental staff tasked with the cataloguing activity with inputs from departmental staff.

2.1 Why document department datasets?

There are numerous challenges in identifying the necessary digital infrastructure to enable interlinkages or interoperability within existing data systems. Combining and analysing data to detect patterns and generate insights enhances the potential of the use of data for policy decisions. Interoperability is the extent to which the data are capable of undergoing integration or of being integrated with other datasets. The effectiveness of administrative data is amplified when datasets can be linked with one another. The ability to link datasets, such as birth registry to immunisation data, and Anganwadi enrolment data to school data, can enable policymakers to quickly identify beneficiaries who may fall through the cracks while transitioning from one stage of life to another. Interoperability also helps with analysing the intended benefits of any policy stimulus more holistically.

Challenge 1: Datasets are collected and stored in different formats and locations, at varying digitization levels (semi-structured to structured data).

For instance, social sectors such as health, education, social welfare etc. have many beneficiary-focused schemes that capture information on services and benefits delivered to individuals, typically entered by a frontline worker into mobile applications. Other datasets within these same departments could be at the institutional level (health centres, hospitals, schools, Anganwadi centres etc.). There are departments such as Food, Revenue, and Taxation which primarily have registration and transactional databases, automatically generated through digitised processes (online applications, point of sale devices, SMS, etc.), whereas Land revenue, Pollution Boards, Power, and other infrastructure departments have diverse area-based and asset information such as road networks, satellite images of land and water bodies, electricity stations, air and water pollution levels etc. Finally, there may be Government-wide information systems such as call records from grievance redressal mechanisms, bills and financial transactions from the Treasury, human resource data, census data and others.

A combination of various formats of these datasets is required to be interlinked and joined together to draw any meaning or for any further analysis. Governments and agencies are making strides towards bridging siloed datasets and connecting them to better inform decisions and deliver public services effectively.

Challenge 2: Lack of information on what data are available is another challenge in facilitating the use of the data. Even when information is available it is often scattered, not maintained in standardised formats or limited to a few persons.

In order to facilitate greater use of data, it is important not only to know what data sets are available but also the details about the data collection process, management, storage and the existing use.

Solution: A central and **comprehensive metadata catalogue** would help first identify all the available administrative and secondary data sources available with the government. Documenting datasets in a standardised format (or a metadata catalogue) helps address the above challenges. Such a central and routinely updated metadata catalogue would enable ideas for linking across datasets and innovative use of data contained therein for planning and decision-making purposes. With increasing focus on open data and data sharing policies, the data catalogue can then be made accessible for external use with a process for accessing data. The feedback from users can further be used to update the documentation and improve data quality and accessibility.

Cataloguing datasets will allow the department, partner and agency to understand the magnitude of the data that is generated and owned, and a standardised format of data across various departments within the organisation or government will facilitate interoperability between departments

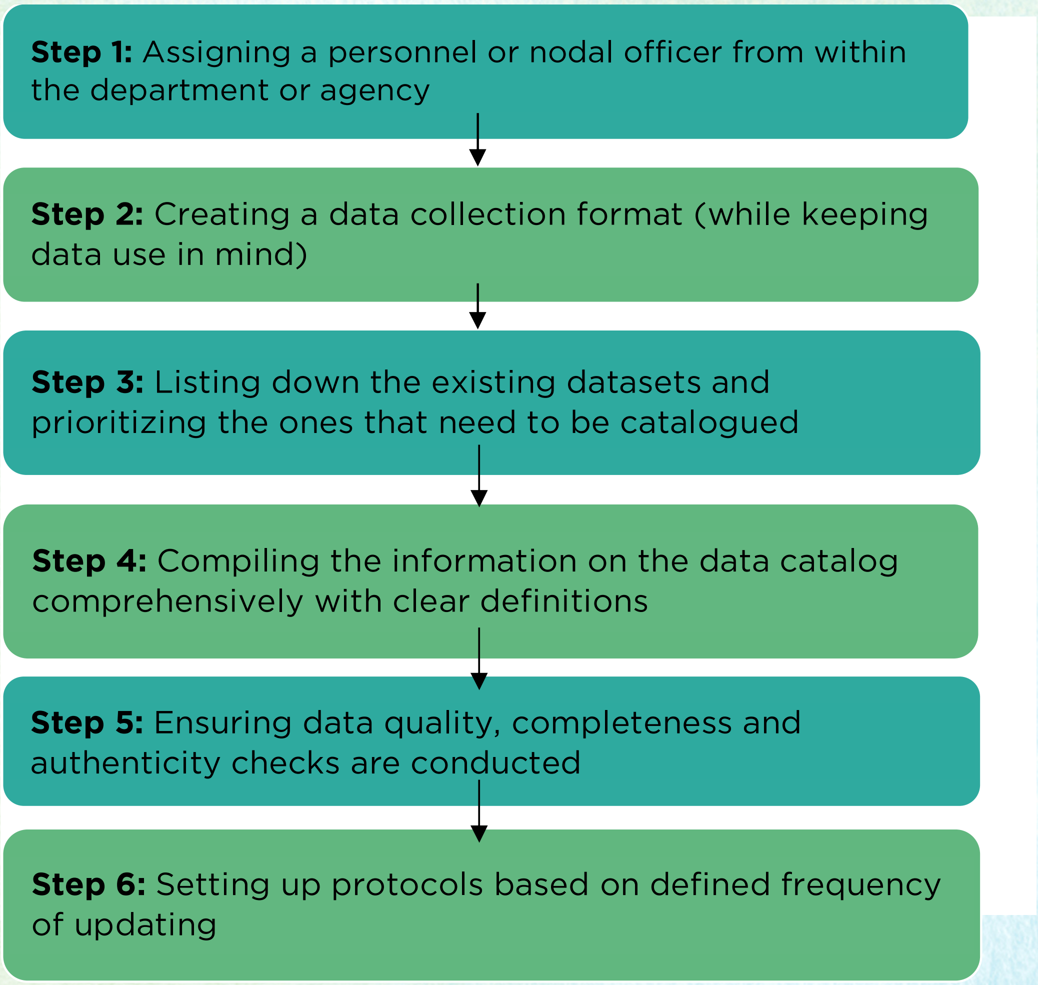
2.2 What is a metadata catalogue (MDC)?

Metadata is often simply defined as "data about data." Metadata provides the information necessary to use a particular source of data effectively and may include information about its source, structure, underlying methodology, topical, geographic and/or temporal coverage, licence, when it was last updated and how it is maintained. An MDC is a standardised format for collation of information on all existing data sets available with the department. Each data set captures a description of basic information such as key data contents or fields, the process of data generation, rules to access, storage, types of data available and potential usability. These data sets can be closed i.e. accessible only by the department or publicly available; they can be administrative data or survey data. A few examples of public data catalogues include data.gov.in, World Bank datasets etc.

2.3 Populating a metadata data catalogue

This section provides a step-by-step guide on how to create a metadata catalogue (MDC) for your purpose.

Flowchart 1: The Metadata Cataloging Process



Step 1: Assigning a personnel or nodal officer from within the department or agency

Step 2: Creating a data collection format (while keeping data use in mind)

Step 3: Listing down the existing datasets and prioritizing the ones that need to be catalogued

Step 4: Compiling the information on the data catalog comprehensively with clear definitions

Step 5: Ensuring data quality, completeness and authenticity checks are conducted

Step 6: Setting up protocols based on defined frequency of updating

- **Step 1: Assigning personnel or a nodal officer from within the department or agency**

  Each department or agency should identify nodal personnel who will be responsible for identifying the data sets maintained/owned by the department and ensuring all relevant information is populated in the MDC. Usually, the person most familiar with the data set should be assigned to populate the MDC and ensure the entire compilation process. This could be the IT personnel; programme officer or any other staff member.

- **Step 2: Creating a data collection format (while keeping data use in mind)**

  The critical step while cataloguing the datasets is to develop a standardised format for collating all information. This step helps organise all information into meaningful categories. These minimal sets of indicators or tags would help create a master index for all the key datasets, contain information about the datasets, and help prioritise the datasets required to be further catalogued. This tabular format can be designed collaboratively by the nodal officer who can work with multiple people or teams who have the knowledge about the individual datasets being entered into the format (A data catalogue template is shared here for reference)

For instance, information could be classified, namely into,

Basic Information (scheme name, title of the dataset)

Data contents related information (granular unit of data, scale of data, personal identifying fields and so on)

Data collection-related information (point of entry of data, eligibility for inclusion, frequency of data collection and so on)

Data quality-related information (data validation/ measures taken, format in which data are available)

Data storage-related information (database management system used, list of stakeholders involved in data management and their roles)

Data use related information (who uses the data and for what purpose, frequency of generation of insights, actions taken using the variable in the dataset)

- **Step 3: Listing down the existing datasets and prioritising the ones that need to be catalogued**

  The next step is to gather all information regarding the schemes and programmes implemented by the department along with the contact information of nodal officers for each scheme in the tabular or spreadsheet format. Identifying which datasets need to be catalogued can be done based on the usability, ownership and depth of the dataset. For example, the Tamil Nadu e-Governance Agency (TNeGA) is developing a State Family DataBase (SFDB) of beneficiaries across Tamil Nadu in an effort to help departments better formulate and target social welfare schemes. The SFDB is being created by consolidating beneficiary data from multiple departments for the social welfare schemes. As each of the departments will have its own data collection and storage processes, efforts are made to understand how the information can be linked using common unique identifiers. The catalogue will then be shared with departments to help them correct data quality issues to promote better use of data for monitoring, planning and decision-making.

  The following criteria can be used to prioritise the datasets to be catalogued:

  a) Type and the usability of the dataset (Is the data for external use or internal use? Is the dataset a welfare scheme/programme dataset or is it a reference dataset?[4])

  b) Ownership of data (Is the state/department owner of the data collected? Is the dataset a part of a central scheme?)

  c) Level of data available (Aggregated or granular dataset)

- **Step 4: Compiling the information on the data catalogue comprehensively with clear definitions**

  Before starting with the compilation of information, understanding who is the owner of each dataset, whether there are multiple departments who own a single dataset, or specific datasets owned by many departments helps coordinate for each dataset

[4] Reference data are the data used to classify or categorise other data. For example, a dataset contains the inclusive criteria for a beneficiary by defining the cut-off age or weight, for enrolment in a particular scheme. Another operational dataset (such as ICDS), accesses the former dataset to track ongoing activities across all Anganwadis. Here the first dataset acts as a standard metric and is referred to as the reference dataset.

to be catalogued along with the partner line departments. Defining the roles of each of the owner, reviewer and approver helps standardise the accountability and define the intended use and scope of the data by the owners (who can help with identifying datasets, who can own the dataset other than the listed owners, and who can help with documentation). Then the process of compilation of all information can begin. The next step is to collate all the synthesised information in a standardised format. This documentation is a crucial step to enable and improve the use of the datasets.

Indicative list of resources that can be used to fill information into different sections:

- Department websites for comprehensive information on all schemes/programmes
- Operational documents of the scheme/programme
- Registration forms/performance under the scheme
- Manuals on the MIS linked to the programme/scheme
- J-SON files for schemes and back-end forms for information about the data contents, frequency of updating information, access modalities
- Detailed interviews with the IT head/officer that manages the data for the programme for information on data contents, data architecture, data documentation & quality
- Interview/discussion with the programme officer for information on the data collection process

- **Step 5: Ensuring data quality, completeness and authenticity checks are conducted**

  Once the MDC format is filled by an assigned staff member, the document must be reviewed to make sure that all basic information is filled in correctly, check for fields that are being recorded in the fields of the table, whether observations for a field are completely missing, are there any fields where more than a threshold percentage of the observations are missing, or whether the primary key is duplicated.

  Several other quality checks can be undertaken to ensure that the data contents are ready for any further analysis.

- **Step 6: Setting up protocols based on defined frequency of updating**

  The usual documentation process relies heavily on creating a one-time reference for datasets, but the efficacy of a metadata cataloguing process relies on regularly updating the information, such that it is relevant and ready for use at any given period of time. Once the above processes are undertaken, there is a need to establish clarity on how the catalogue will be maintained and updated on a periodic or regular basis. Setting up protocols while looking at each dataset and the frequency at which it is updated and used is necessary to be clearly articulated when the catalogue is being prepared. It is important to also update the catalogue at least once a year as there may be changes to information systems, and new data sets could be added on or upgraded.

The purpose of the MDC is to facilitate greater data use with a focus on interlinkages and interoperability across departments. To this end, the MDC should be hosted on the department's/ Ministry's website or maintained by a central agency such as Planning or Economics & Statistics, or Information Technology departments. Selected fields of the catalogue can be made viewable (publicly) and more detailed information through registered access (requiring login credentials) for those outside of the government. After operationalising the above steps, it is important to include protocols for sharing aggregated and non-aggregated but anonymised datasets with various partners, setting up standardised processes to make sharing of datasets secure, streamlined and sustainable. Please refer to the section on Data Sharing Standards for more details on the procedures for data exchange and external data use.

Section 3
Data Privacy

Administrative data are very likely to contain particularly sensitive information and therefore data privacy measures are critical to ensure that the privacy of individuals is protected. Minimising the contact with individually identifiable and sensitive data may substantially reduce the vulnerability, inadvertent disclosure of, and targeted attacks on, individuals in administrative datasets. Strong privacy safeguards can be enablers in the use of administrative data for research and evidence-based policymaking.

3.1 Data privacy and data security

Before delving further into the specific measures for data privacy, it is important to distinguish data privacy from data security.

While both data privacy and data security fall under the ambit of data protection, data privacy deals with protecting the identities of individuals captured in a dataset while data security deals with protecting the dataset itself. When addressing data privacy issues, the focus is on the collection, processing, storage and transmission of data with the attention and consent of the parties involved. When an organisation/agency collects data, individuals need to know what data is being collected, why it is needed, and with whom it is shared for transparency. Furthermore, the data subject must agree to these terms. When developing a data security policy, the focus of safeguarding measures is to prevent unauthorised access to data.

Data security includes all measures, policies, and technologies put in place to protect data from external and internal threats. However, the application of data security measures alone does not always meet data protection requirements. Data protection still requires compliance with regulations regarding the collection, sharing, and use of the data an organisation/ agency protects. Data security protects data from malicious threats while privacy is about the responsible management or use of that data.

3.2 Data privacy laws and regulations

Protecting privacy is critical while conducting any analysis or research with administrative data in an ethical way. Data systems must use rigorous system design to safeguard people's privacy and give them control over their data, in addition to responsive and adaptable architecture. A solid legal framework ought to complement these measures. The OECD's Fair Information Practices (FIPs) and emerging international good practices, such as the European Union's General Data Protection Regulation (GDPR), are two examples of global standards for data protection that should be followed when managing data used for identification and authentication.

The European Union's (EU) 2016 General Data Protection Regulation (GDPR) is the most recent example of comprehensive regulation of data protection and privacy, setting a new threshold for international good practices. It has become an important reference point for global work in this area. Article 5 of the GDPR, enshrines the core principles described above, requiring that personal data collection, storage, and use by:

- Processed lawfully, fairly and in a transparent manner concerning the data subject
- Collected for specified, explicit and legitimate purposes
- Adequate, relevant and limited to what is necessary concerning the purposes for which they are processed
- Accurate and, where necessary, kept up to date
- Kept in a form that permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; and
- Processed in a manner that ensures appropriate security of the personal data.

In addition to the above, implementing states are required to provide a supervisory authority to monitor the application of the regulation (Article 51(1) of the GDPR).

To guarantee that privacy and data protection laws are adhered to, as well as the rights of individuals, personal identification systems and data protection are frequently subject to the supervision of an independent supervisory or regulatory authority. A single government official, an ombudsman, or a group with several internal and external members could serve as the supervisory authority. A key factor is the authority's true independence, which is measured by structural factors like the composition of the authority, the method of appointment of members, the authority's authority to exercise oversight functions, the allocation of sufficient resources, and the authority to make decisions that are meaningful without interference from outside parties.

3.3 Ensuring data privacy while accessing data

Access to datasets can be classified in a simple form as shown below based on personal identification of data & the level of aggregation. Different access levels that are desired for these combinations are given below.

Type of data	Personally Identifiable	Personally Non-Identifiable
Non-Aggregated/ Transactional	Restricted Access	Registered Access
Aggregated	Restricted Access	Open Access

Box 2: Data Access Matrix

- **Non-Aggregated & Personally Identifiable**

  This represents any data like health records, transport data, and education where some uniquely identifiable personal information is present. This needs to be clearly set in a restrictive mode and access should be given only to authorised users within the government or agency.

- **Non-Aggregated & Personally Non-Identifiable**

  A personally identifiable record could be anonymised by removing personally identifiable fields. For example, health records could be anonymised and be given for research purposes. Educational data can help in identifying eligible students for disbursement of benefits like scholarships and create a database of employable resources.

- **Aggregated, but personally identifiable**

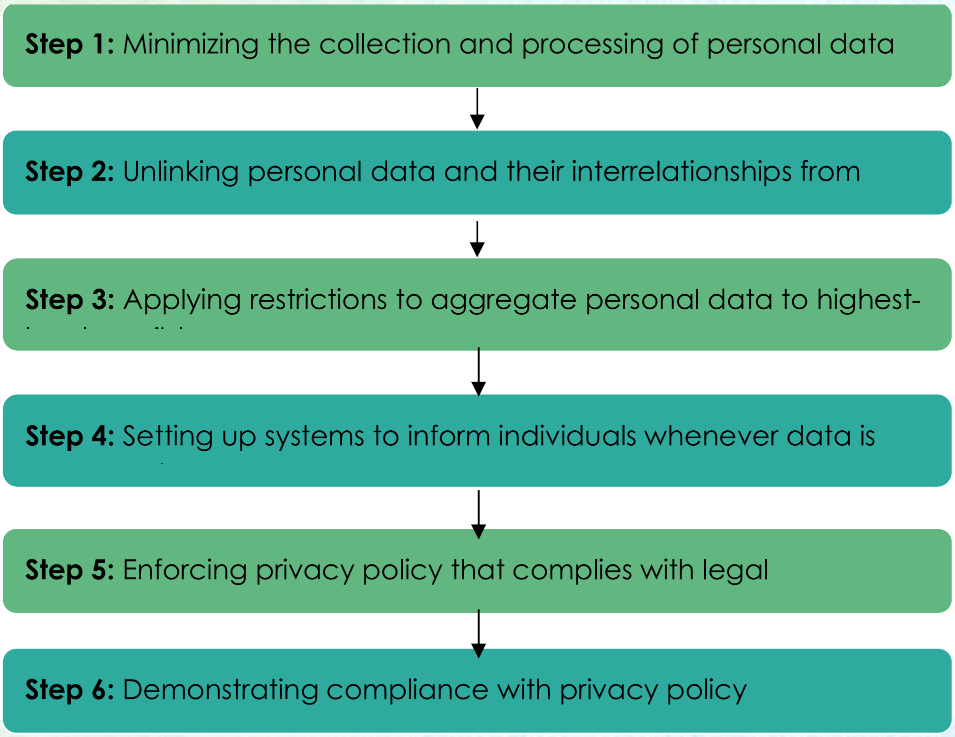
  The sum-total of benefit disbursements or transactions such as aggregated electricity usage of a person. Even though it is aggregated, these are personally identifiable. So, access should be given only to authorised users in the Government.

- **Aggregated & Personally Non-Identifiable**

  These are typical datasets that can be published under OGD platforms like data.gov.in. In other words, this information can be petitioned and obtained under RTI. Such datasets are to be proactively released by every department under the OGD initiative.

3.4 Setting up data privacy measures to protect administrative data

Flowchart 2: Setting up privacy measures for administrative data



Step 1: Minimizing the collection and processing of personal data

Step 2: Unlinking personal data and their interrelationships from

Step 3: Applying restrictions to aggregate personal data to highest-

Step 4: Setting up systems to inform individuals whenever data is

Step 5: Enforcing privacy policy that complies with legal

Step 6: Demonstrating compliance with privacy policy

It is important to note that some of these steps may be common to both data privacy and data security.

Step 1: Minimising the collection and processing of personal data

To minimise the system's impact on privacy, reduce collecting and processing personal data to a minimum. Personal data should be adequate, relevant, and limited to the purpose of its collection. Any specific points that a person brings to your attention, such as an objection, request for rectification of incomplete data, or request for the erasure of unnecessary data, should be taken into consideration. If personal data is not sufficient for the purpose for which it was collected, it should not be processed. In some cases, there may be a need to collect more personal information than originally planned to use in order to have enough for the purpose at hand. Hence, it is important to regularly check the relevance of personal data and periodically get rid of sensitive data that is not usable. The agency should only keep the information necessary to create a basic record of a person they have removed from their search, removing all other personal data from their repository.

Step 2: Unlinking personal data and their interrelationships from potential abuse

Encryption, anonymisation or use of pseudonyms can help with hiding personal data and their interrelationships from plain view to achieve unlinkability and unobservability and minimise potential abuse. Anonymisation and pseudonymisation are methods that can be used to hide identities and personal data but in distinct ways. "The processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information" is how the GDPR defines pseudonymisation. As a result, personal data is replaced with non-identifying data ensuring that additional information will be required to recreate the original data. The term "anonymised data" refers to information that has been anonymised in such a way that no longer identifies the registered user. Separate, compartmentalise, or distribute the processing of personal data whenever possible to avoid the ability to make complete profiles of individuals.

Step 3: Applying restrictions to aggregate personal data to the highest level possible

Aggregate personal data to the highest level possible when processing to restrict the amount of personal data that remains. Anonymise data using k-anonymity, differential privacy and other techniques (e.g., aggregate data over time, reduce the granularity of location data, etc.). "Differentially private methods provide strong promises to prevent outside parties from learning whether any individual is in the data, regardless of the background information available to others. In this, it differs from traditional methods, which typically protect against specific, rather than general, methods of breaching privacy. Differentially private methods are being used more and more for releases of tabular data, for instance by the US Census Bureau (Machanavajjhala et al. 2008), Google (Erlingsson, Pihur, and Korolova 2014), Apple (Differential Privacy Team 2017), SafeGraph (SafeGraph 2020), but can also be challenging to implement."[5]

Differential privacy is a strict mathematical definition of privacy. At its simplest, imagine an algorithm that analyses a dataset and computes the statistics (mean, variance, median, mode, etc.) of the data. Such an algorithm is said to be differentially private if it cannot be determined from the output whether personal data was included in the original record.

[5] Cole, Shawn, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. 2020. "Section 1.3.1 : Different Levers for Protecting Sensitive Data: The Thematic Chapters. Chapter: Using Administrative Data for Research and Evidence-Based Policy: An Introduction." In: Cole, Dhaliwal, Sautmann, and Vilhuber (eds), Handbook on Using Administrative Data for Research and Evidence-based Policy.First Published January 1, 2021.Accessed at https://admindatahandbook.mit.edu/book/v1.0/intro.html .

In other words, the Differential Private algorithm guarantees that its behaviour hardly changes when one person joins or leaves the dataset. Anything an algorithm might output to a database that contains information about an individual is very likely to come from a database that has no information about that person. Best of all, this guarantee applies to everyone and all datasets. Differential privacy guarantees apply no matter how fancy your personal details are, or if your database contains other people's details. This formally ensures that no individual-level information about participants in the database has been compromised.

Step 4: Setting up systems to inform individuals whenever data is processed

Inform individuals when, for what purpose, and by what means their data is being processed by sharing transaction or data breach notifications. A key principle of the by-design approach and international principles on privacy and data protection is to hold governments, third parties and their actors accountable for potential abuses. In addition, these standards should require general openness and transparency regarding policies and practices related to the management of personal information and should be readily available to individuals. One way of implementing personal oversight of data usage is by creating a platform or portal where individuals can log in to see their personal information and a record of who accessed the data, when and why. India also has a portal where a resident can view her record of authentication using her Aadhaar number. Such portals enable users to have control over their data. At the same time, platforms that require internet access may exclude people in less-connected areas or those with low digital literacy. Practitioners therefore ensure that people have access to other procedures (such as physical offices) and grievance mechanisms to report and correct errors in data and to monitor who uses data and for what purposes.

Step 5: Enforcing a privacy policy that complies with legal requirements

Enforce privacy policies that meet legal requirements through role-based access control and authorisation. Authentication is the process of confirming that you are who you say you are. This involves matching an individual's claimed identity, verified by a credential (such as an ID card or a unique ID number), against one or more of the authentication factors associated with that credential. Secure authentication (that is, for higher security levels) requires a multi-factor approach. In general, any combination of authentication factors should include some or all of the three categories above. Additionally, sub-factors such as location (where you are) and time (when you are trying to authenticate) can be used in combination with other core factors to further condition authentication. Both online and offline authentication mechanisms share a common set of requirements to protect individuals claiming identity and to provide adequate security to consumers of identity (services, individuals, or dependents). In general, an authentication mechanism following legal requirements should do the following:

- A known and easily accessible exception handling and complaint resolution protocol in case of authentication mechanism failure (e.g. false negative biometric results). No right, benefit, or entitlement may be denied (or made difficult to access) due to a failure of the identity system.
- Eliminate opportunities for identity authorities or other actors to use transaction

metadata to track or profile identity owners (e.g., by encryption, hashing, anonymizing data, decentralising such data, etc.). Mandated by relevant laws and regulations, certain relationships between identity systems and relying parties are governed by legal agreements (e.g., memorandum of understanding) that set out their respective responsibilities.

- Determines the attributes, if any, that will be passed from the identity provider to the relying party/service provider upon successful authentication of the user.
- Establish a secure communication channel between the relying party and her identity provider to enable authentication workflows between the service provider and her identity provider applications. Manage digital identities including expiration, revocation, and renewal.

Step 6: Demonstrating compliance with privacy policy

Demonstrate compliance with privacy policies and applicable legal requirements using tamper-resistant logs and audits. To ensure that only authorised users have access to personal data for their authorised purposes, you need a way to track transactions and identify who accessed the data and when. Automatic tamper-proof logging of transactions involving identity data is a best practice method for organisational and personal oversight of how that data is used. Any log or audit data collected must comply with the data protection requirements of the identity system in question. The log should contain at least the following[6]:

- Protected from unauthorised access (and have that use monitored);
- Protected from unauthorised copying or exfiltration; and
- Devoid of personal data

For example, in India and Estonia, logs are digitally signed to detect tampering. Additionally, Estonia has digitally signed logs that are concatenated, making it difficult to change history. New technologies such as blockchain could potentially make these protocols more secure, even against the authorities that control them, by making them immutable. Ensure logs are removed from application servers as soon as possible and sent to a central log management system. Whenever possible, send log data directly to a centralised log management system. Ensure log files and log transactions are encrypted in transit and at rest. Analyse log file activity to identify gaps in logging or corruption patterns that may highlight suspicious activity.

[6] Clark, Julia. 2019. Section III Topics: Privacy and Security .ID4D Practitioner's Guide: Version 1.0 .October 2019. Washington, DC: World Bank. License: Creative Commons Attribution 3.0 IGO (CC BY 3.0 IGO). Accessed at https://id4d.worldbank.org/guide/privacy-security .

Section 4
Data Security Procedures

Data security procedures are the formalised, internal processes and standards that the data agencies can employ to protect the administrative data that they collect. Data security requires adequate planning, development of procedures, and training and supervision to ensure that data are stored, archived, or disposed of safely and securely preserving the integrity of data.

4.1 Rationale behind setting up data security procedures

Data security is essential for safeguarding private information, maintaining the privacy of the subjects, and complying with requirements and regulations. Ensuring data security involves the "interaction of legal, technical, statistical and, above all, human components".

The Five Safes Framework outlines a set of considerations to be kept in mind to check for secure data procedures:

Safe projects: Is the use of the data ap propriate?

Safe people: Can the users be trusted to use it appropriately?

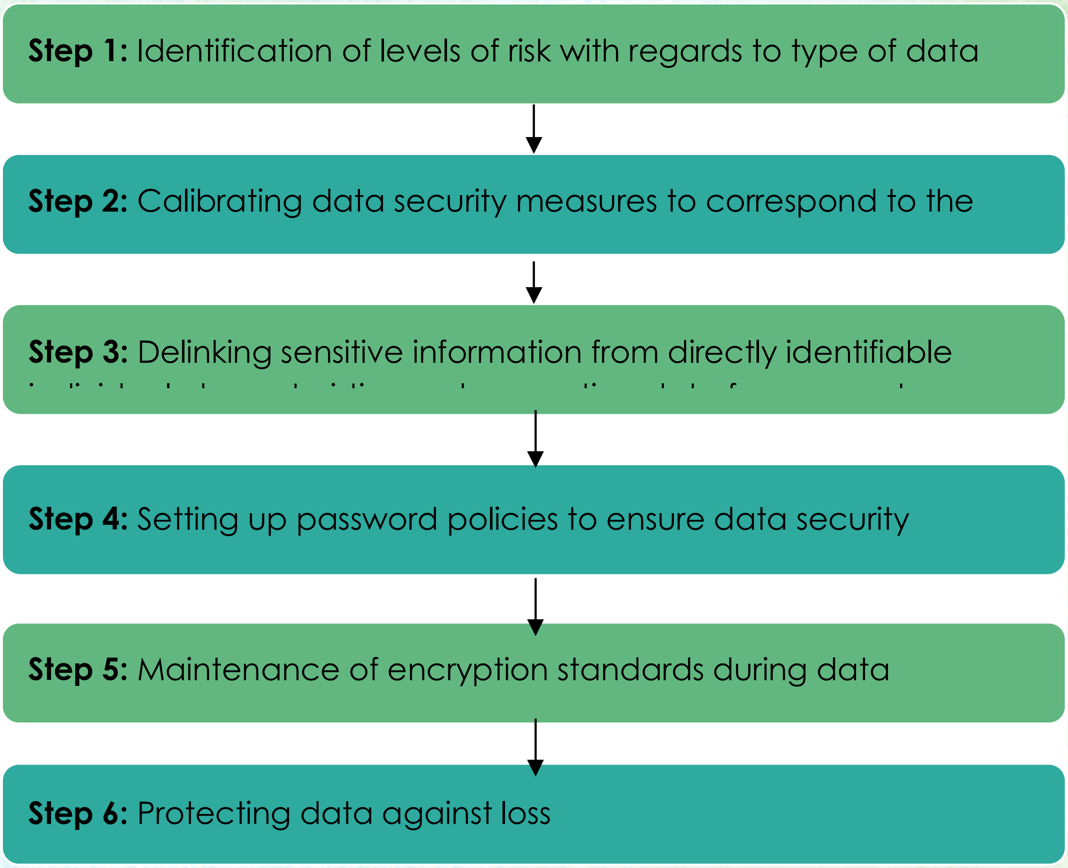
Safe data: Is there a disclosure risk of personal information in the data itself?

Safe settings: Does the access limit unauthorised use?

Safe outputs: Are the statistical results or outputs meeting confidentiality requirements or are released without the required data owner's approval?

4.2 Setting up data security procedures to protect administrative data

Flowchart 3: Setting up data security procedures



Step 1: Identification of levels of risk with regards to type of data

Step 2: Calibrating data security measures to correspond to the

Step 3: Delinking sensitive information from directly identifiable

Step 4: Setting up password policies to ensure data security

Step 5: Maintenance of encryption standards during data

Step 6: Protecting data against loss

Step 1: Identification of levels of risk regarding the different types of data being handled

To enable the access and usage of administrative data securely, it is essential to identify safe data flows[7] while considering applicable regulations, confidentiality, data security capabilities, and the potential need to maintain the ability to add follow-up data or new data sources. The first step here is to determine what type of information is stored and processed in your environment, such as medical or financial data, as defined by your organisation or by law. The next step is to classify the data[8] by the level of risk associated with handling it; using a standard classification table (that details the risks involved for each level) For example; Level 1 - Information intended and released for public use (de-identified data which can be published) or Level 2 - Low-Risk confidential information that may be shared only within the smaller group (unpublished intellectual property) or Level 3 - Medium risk confidential information intended only for those with a "need to know." or Level 4 - High-risk confidential information that requires strict controls (contains personally identifiable sensitive information). This is also relevant to data privacy.

Step 2: Calibrating data security measures to correspond to the different levels of risks

[7] Bauman, Jason, Geeti Mehra, Julia Chabrier, Laura Feeny and Michelle Woodford.2015. "Section:Data Flow. Using Administrative Data for Randomized Evaluations." Abdul Latif Jameel Poverty Action Lab North America .December 2015. Accessed at https://www.povertyactionlab.org/resource/using-administrative-data-randomized-evaluations .

[8] Harvard University. "Data Classification - Administrative Examples." Information Security and Data Privacy. Last Modified n.d. https://privsec.harvard.edu/data-classification-table .

Data security measures must be calibrated to respond to these levels of risk, and corresponding requirements for security can be imposed. We must use information security policies that classify data into levels based on confidentiality and set standards for use and sharing such that the higher the data level accorded to a type of data, the greater the required protection. For example, according to the policy, a level 1 classification of data is "information that is considered public", such as research data that has been de-identified in accordance with applicable rules.

Step 3: Delinking sensitive information from directly identifiable individual characteristics and encrypting data for secure storage and access

Data poses the most risk when sensitive information is linked directly to identifiable individuals. Once delinked, the data must be handled separately, and the identifiers should remain encrypted at all times.

In order to ensure safety, the selected sensitive data must be first transformed into an encrypted code that needs a password or pair of "keys" to decipher it. This encryption of data may take place at multiple levels (such as at the device, folder, or file levels), at various phases of the data lifecycle, and using a range of software and hardware packages as well as methods to balance privacy and usability[9]. Once separated, the "identifiers" data set and the "analysis" data set should be stored separately, analysed separately, and transmitted separately. Once separated, the identifiers should remain encrypted at all times, and the two data sets should only link again if necessary to adjust the data matching technique. Access to the "identifiers" data should be limited only to a few personnel. Please refer to the section on data privacy for the importance of securing sensitive data.

Step 4: Setting up password policies to ensure data security

To guarantee data security, strong passwords are necessary. Each high-value account should have a unique password. For instance, the passwords for institutional servers, email, and encrypted files should all be unique. Furthermore, passwords must not be shared using file-sharing mechanisms or over the phone. It is important to rely on password storage systems and other safer alternatives for password sharing.

The following mechanisms must be incorporated to ensure secure data transactions:

- Methods to handle inactivity or timeouts during remote access,
- Processes to handle non-retrievable passwords (i.e. if a user forgets his or her password, the password is reset by the system, rather than the original password being returned)
- Restriction on the number of password guesses permitted before account lockout
- Storing access logs that describe who signed in, from where, and when

Step 5: Maintenance of encryption standards during data transmission and sharing

Data must be safeguarded both at rest and while being transferred between the data provider, data recipients and their collaborators. "Even though using encryption may decrease convenience (a password or a hardware key needs to be used each time decryption occurs), utilising encryption for data and devices should be mandated as a minimum-security feature as part of any data access mechanism[10]."

A whole-disk encrypted laptop or a secure server that stores encrypted data may

[9] Schmutte, Ian M., and Lars Vilhuber. 2022. "Section 5.2: Methods .Chapter: Balancing Privacy and Data Usability: An Overview of Disclosure Avoidance Methods." In: Cole, Dhaliwal, Sautmann, and Vilhuber (eds), Handbook on Using Administrative Data for Research and Evidence-based Policy, Version v1.1. First Published January 1, 2021. Accessed at https://admindatahandbook.mit.edu/book/v1.1/discavoid.html .

or may not provide protection for data in transit. In order to maintain the security of data in transit, it is essential to keep in mind the encryption standards while sharing files in any form or medium. Advanced Encryption Standard (AES) is most commonly used by governments and security organisations as well as everyday businesses for classified communications.

Step 6: Protecting data against loss

Besides securing against threats of misuse, it is equally important for data security to prevent the loss of data. Data and processing information must be backed up securely and regularly, with timely maintenance of passwords. These may be device-level, institutional-level, or cloud-based backups, depending on resource availability and sensitivity of data. An encrypted hard drive may also be used to maintain backups, especially in areas with low connectivity. The access to this storage must be regulated and closely monitored using a digital log system for the personnel intending to use data from the storage. The backup storage must be updated regularly based on the frequency of updates of the data. General practice is to back up every 24 hours, with incremental backup every 3 hours for real-time data, and differential backups for sporadic data flows. The data backup must be protected both physically in case of instances of fire or flood, and also protected from intentional unauthorised attacks by using a strong firewall system.

[10] SShen, Jim, and Lars Vilhuber. 2022. "Section 2.3.3: Encryption. Chapter: Physically Protecting Sensitive Data." In: Cole, Dhaliwal, Sautmann, and Vilhuber (eds), Handbook on Using Administrative Data for Research and Evidence-based Policy, Version v1.1. Accessed at https://admindatahandbook.mit.edu/book/v1.1/security.html .

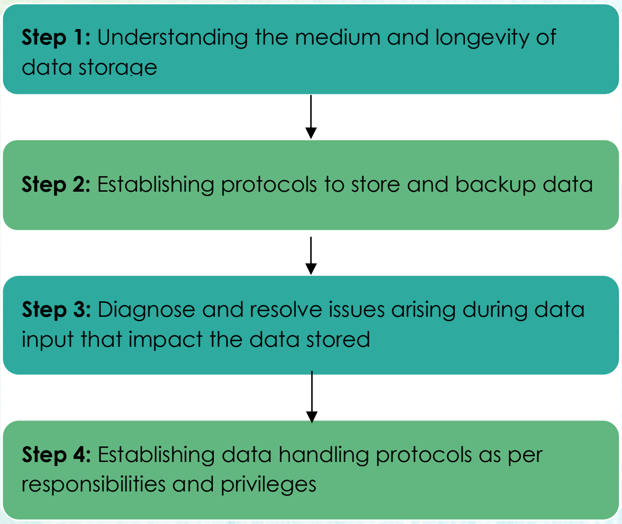
Section 5

Data Handling

Data handling is the process that ensures the secure storage, archiving or destruction of data during and after a project or duration of the scheme is completed. This includes the development of policies and procedures to manage data processed electronically and non-electronically.

Data handling[11] is important in ensuring the integrity of administrative data since it addresses concerns related to confidentiality, security, and preservation/retention of administrative data. Data must be protected both when at rest and in transit between the data provider and stakeholders. Data that is encrypted while at rest on a whole-disk encrypted laptop, or a secure server, will not necessarily be protected while being transmitted. In the case of data handled electronically, data integrity is a primary concern to ensure that recorded data is not altered, erased, lost, or accessed by unauthorised users.

Flowchart 4: Data Handling Process



Step 1: Understanding the medium and longevity of data storage

Step 2: Establishing protocols to store and backup data

Step 3: Diagnose and resolve issues arising during data input that impact the data stored

Step 4: Establishing data handling protocols as per responsibilities and privileges

[11] O'Toole, Elisabeth, Kenya Heard, Laura Feeney and Rohit Naimpally. 2015. "Data security procedures for researchers." J-PAL North America. Last Modified May 2023. Accessed at https://www.povertyactionlab.org/resource/data-security-procedures-researchers .

Step 1: Understanding the medium and longevity of data storage

Deciding how long the data or dataset for a particular scheme or programme should be kept may depend on the nature of the scheme, project, department or agency's guidelines, ongoing interest in or need for the data, cost of maintaining the data in the long run, and other relevant legislative considerations. Under current requirements of data retention, records are usually maintained for three years to ten years depending on the category of dataset. Physical records are digitised or microfilmed. Institutional guidelines may require that data be retained for longer periods. Understanding the longevity along with the medium of storage is the first step before handling any data. The storage of data or datasets could be physically stored or stored on the cloud or a combination of both. While storing data it is important to create a restriction on the access of the stored data i.e., who can read and overwrite the data. Further, encrypting the stored data with the use of passwords as well as creating a backup of the data to protect the data from loss is imperative. Understanding the medium of storage would help deploy compatible data security and privacy measures.

Step 2: Establishing protocols to store and backup data

Before setting up the protocols to store and backup data, it is important to classify the indicators or individual data into sensitive and non-sensitive data. The personally identifiable data must be separated and stored while being encrypted and should be accessible by a limited number of data users. In the case of data stored electronically, the potential for altering, erasing, losing, or unauthorised access is high. Several years of valuable data can be compromised or lost, if any intruder or unauthorised personnel breaks into a server. Although some aspects of protection from these threats are the responsibility of IT personnel, researchers and other users who gain temporary access are also responsible for ensuring the security of their data. "If the data are recorded electronically, the data should be regularly backed up on a hard copy; and should be made of particularly important data; relevant software must be retained to ensure future access, and special attention should be given to guaranteeing the security of electronic data" (ORI website, 2003).

Step 3: Diagnose and resolve issues arising during data input that impact the data stored

There may be multiple issues while data are being collected or inputted that need to be diagnosed and mitigated at this stage. Issues in data input can directly impact its storage, retrieval, and preparation. The following questions need to be addressed:

- **Clean data values:** The data entry process often introduces typos, error codes and errors in the data. Do different data storage strategies support different levels of data manipulation? Do the changes in data values need to be tracked?
- **Static or dynamic data:** Is the data static or will updates be available throughout the lifetime of the analysis? [Is new data constantly being added? Will there be new data in the same format? Are new fields added?]
- **Sources of collection:** Was the data obtained from different sources or by a single source? Since the data are taken from a greater number of sources, the need to transform data into a common format is more critical.
- **Common standardised definitions:** What if different origins use the same definition for a common variable? For example, does one source define "medium delay" in the same

way as another? If not, can we perform a simple transformation between definitions?

- **Classifying data values by use:** Will all data be used in the analysis or will a subset of the data be analysed? To speed analysis, work with a subset of the fields in the record, rather than with all the fields in the database at the same time.
- **Automation requirements for repeat analysis:** Is this a one-off analysis or is it planned to repeat this analysis, how often will the analysis be repeated? The more iterative the analysis, the more important it is to automate the analysis process.

Step 4: Establishing data handling protocols as per responsibilities and privileges

To establish and define the roles and responsibilities of all kinds of data users in the system, it is suggested to articulate using a simple RACI matrix (See below).

Function	Data Executive	Data Owner	Data Steward	Data Trustee
Data entry	Accountable	Consulted	Responsible	Informed
Data quality	Accountable	Responsible	Consulted	Informed
Data verification	Accountable	Consulted	Informed	Responsible

Box 3: RACI Matrix

The functions and responsibilities of each data user must be defined further to ensure clarity in terms of who might mitigate what kind of issues. Sample descriptions of the responsibilities upheld by the most common data user categories are shared below:

Sample roles and responsibilities of different stakeholders for data handling:

- **Data Executive:** Data Executives have the overall accountability for ensuring that the data are fit for a purpose and create the internal and external standards and ensure these are met. Procurement and selection of hardware and software, allocating personnel to various roles, conducting reviews of the data quality with the team members, and ensuring that the issues in data are resolved.
- **Data Owner:** Primarily responsible for ensuring that only the relevant subset of the data from the owned and managed database is being used across the system. Provides day-to-day leadership and direction, accountable for ensuring the integrity of data, allocating resources and resolving escalated issues. Authorises access to and use of data. Able to modify the dataset and set protocols and rules for editing values within a dataset for other users.
- **Data Stewards:** Responsible for the day-to-day management of data. Ensure that the data standards are followed while inputting data, monitor and track data quality, identify data entry errors, correct the data to match with the pre-defined standards and handle enquiries about data. Usually input and track the data and not edit the data retrospectively. Can work with a cross-section of the data and not access or download the complete dataset.
- **Data Trustee:** Responsibilities include following the policies and standards, ensuring the appropriateness, accuracy and timeliness of data, reporting any unauthorised access, misuse or data quality issues to the data steward for remediation, complete

all necessary training required. Perform quality checks, maintain backups and monitor the data. May analyse the data periodically to ascertain that the information captured is valid and reliable.

5.1 Data Sharing Standards

There are multiple forms of data exchange possible between the government and other partners. For data exchange between government departments, standardising field-level data elements for all datasets would help create uniformity between exchanges with all schemes/ departments.

1. *Government to Citizen*

   The government may want to set up a platform where citizens can access their personal data on scheme eligibility, benefits received, entitlements, certificates etc. Or, the government may want to, for accountability, set up a website, or dashboard to allow citizens in general and civil society to monitor, and understand how government programmes are being implemented, expenditure and utilisation patterns.

2. *Government to Government*

   Such exchanges may be inter-department, required to create linked datasets across departments such as linking ICDS data with school education data, agricultural production with land records etc. It could also involve intra-departmental transfers, for example, enrollment information of primary schools is shared with higher education departments to assess drop-outs, or for creating single and overarching platforms like a comprehensive HEALTH MIS or an urban stack which requires pooling across different data systems.

3. *Government to third parties*

   This could involve an exchange of data with tech vendors to populate MIS and other tech solutions for programme delivery. Alternatively, data could be shared with academics, researchers, think tanks and civil society organisations for use in policy-oriented research to feedback into government decision-making and/or for independent research.

For each scenario, there is a need to create broader guidelines concerning the level of data access, privacy and processes issued. Some cases may need more privacy and data protection measures than others. Some may require formal data access approval, some could be automatic. Data-sharing standards involve legal and regulatory contexts that must be incorporated into the data owner's effort to share data. In this section, we outline practices that facilitate the responsible use of administrative data for evidence-based policymaking to the full extent of existing laws. It identifies common issues to consider when negotiating an agreement to securely share data. This will provide a set of guidelines for determining how to share data in a way that protects privacy and confidentiality while making the data useful to inform decision-makers.

5.1.1 Defining procedures for data exchange and external data use

Data exchange for internal or external use requires a formal structure to be put in place. Formal agreements ensure the validity and purpose of the data use between any individual or institutional parties, such as governments, agencies, technology partners, researchers or consultants. These parties must ensure that the following details are shared with the data owners, in their request for data.

- **Dataset name:** The name of the dataset, scheme or subset of data required
- **Data structure:** The data schema and model (structure of the data, variables, data types, or any interdependencies)
- **Data dictionary:** To aid interpretation and understanding of the data provide supporting documentation, if necessary. E.g. in the case of education data, what is the definition of a 'student'?
- **Data security and encryption:** The request should specify how the data can be transferred safely following the relevant data encryption and security protocols. If data encryption is required, what encryption method will be used? How will the security certificate or encryption key be transferred?
- **Data exchange process flow:** What will be the data exchange flow between the provider and requester i.e., what is the process for exchanging data from when the data are collated, transmitted, used and disposed of?
- **Format of exchange:** The format that will be used to transfer the data, such as:

  a. CSV, comma-separated file

  b. TXT, plain text file

  c. SQL, query for the relational database

  d. Data Interchange Format

  e. Open Document Format

  f. Others
- **Frequency of data:** Will the data be exchanged once, or will it be recurring? If recurring, how frequently will the exchange occur (real-time, weekly, monthly, or yearly)? Also, the start and end date of the data access request must be mentioned.
- **Responsibilities:** Who is the technical person responsible for the environment in which the data will reside?
- **Data retention:** How will the data be disposed of if the Requestor is to only retain it for a limited time?

5.1.2 Data Use Agreements

The following section provides an overview of potential clauses that can be included in formal data-sharing agreements. Any data being shared between two or multiple agencies must be governed by the clauses outlined below. Actions must be taken to amend the data use agreement[12] from time to time, for all parties to remain in compliance with applicable regulations. The following sections can be included for creating cross-department or cross-agency data-sharing agreements.

Section 1: Key Definitions

- **Purpose -** *Rationale:* Including this section would help clearly state the scope of this specific data exchange, the usage of the data intended to be exchanged and the timelines involved. *Clauses:* The overall purpose of the specific data exchange, scope of usage of the data once received, broad stages and timelines defined must be included.
- **Data - Rationale:** This section would help articulate the entire list of datasets intending to be shared, along with its ownership and storage details. This would be helpful to understand the complexity of data storage/ retrieval and the additional permissions required to access the data or specific datasets. *Clauses:* The list of datasets or subsets of data with their description, current ownership of these datasets, storage platform/ type of servers where the dataset is stored, and a brief plan on the use of the datasets must be mentioned within this section.
- **Confidential information -** *Rationale:* This section suggests how the personally identifiable information would be treated and handled. Confidential information (also termed as Personally Identifiable Information) includes personally or unit-identifying information obtained in the process of conducting data collection. Aggregate statistics as well as information that would be available to a third party under the ambit of Right to Information are not Confidential Information. Clauses: The parties and personnel involved in working and accessing the data must be listed here along with the specific indicators or groups of datasets which reveal personally identifiable information.

Section 2: Policies

1. Sharing and Transfer

a. Consent and format of the data - *Rationale:* Once the data to be used has been identified by all the parties, access to the required data will be provided by the data-providing party to the data-requesting party after the due approval and consent. Clauses: The criteria for data usage and the kinds of permissions (accessing the dataset, other parties who own a part of the dataset, use in the agreed manner, and publishing aggregate statistics) required by authorities must be stated here.

b. Transfer of data: *Rationale:* Data access may be enabled against the proposed scope of work and data request either through a transfer of relevant data through physical hard drives or online (web access). Clauses: The data-providing party would state how relevant data as

[12] O'Hara, Amy. 2020. "Model Data Use Agreements: A Practical Guide." In: Cole, Dhaliwal, Sautmann, and Vilhuber (eds), Handbook on Using Administrative Data for Research and Evidence-based Policy. First Published January 1, 2021. Accessed at https://admindatahandbook.mit.edu/book/v1.0-rc4/dua.html .

requested and agreed upon in the scope of work will be shared with the data-providing party. The data sharing party can also add clauses relating to acknowledgement of receipt of the data. In case the data requesting party will use the data to survey or sample individuals then the goals and objectives of the survey, and transfer protocols to cite and acknowledge the dataset owners must be provided.

2. Storage and Security

a. Release of data and handling sensitive information: *Rationale:* This section would share the need and handling protocol for both non-PII and PII data. There might be a case when only non-PII data would be sufficient for the exchange, while in certain cases PII data might be required for exchange or it might not be possible for the data sharing authorities to separate the PII from the data while sharing. Clauses: If the data consists of any PII data while sharing, the data requesting party must include clauses regarding their encryption protocols and the applicable national, state and local laws and regulations.

b. Retention and destruction of data post analysis: Rationale: It is important to restrict the use of data only to the purpose outlined in the agreement and not for other purposes. The data must be only retained for a stipulated time and destroyed afterwards. In case the data are required to be retained beyond the stipulated time a written approval by the data owning department must be requested. Clauses: Specific guidelines must be laid for the return or destruction of the identifiable information. The stipulated time for which the data will be stored must be mentioned and the steps that would be followed to destroy this data and revoke data access must be mentioned, even if it is in a phased manner.

3. Access and use

a. Use of data: *Rationale:* It is important to constrain the data requesting party's use of data to the purpose outlined in the agreement and not for other purposes. Clauses: Therefore, in this section, the use to which the data will be put should be articulated by the data requesting party. It should also include clauses requiring Non-Disclosure Agreements from any external parties/persons wishing to access the data for the purpose of analysis, such as subject experts and researchers.

b. Amendments and reviews to the access: *Rationale:* The data shared should be used to achieve the scope of work mutually agreed upon. Any further use of the data beyond the activities defined in the scope should require an amended / additional scope of work to be shared with the data-providing party. Clauses: The agreements should include language indicating that all appropriate administrative, technical, and physical safeguards must be ensured to prevent unauthorised use of or access to the data until reviewed by the data-providing party.

4. Ownership and Publication

a. **Explicitly stating the ownership of the data and the analysis:** *Rationale:*The data provided for the purpose of projects are owned by the data-providing party and the parties reserve the right whether to make aggregate (de-identified) statistics available publicly or not. Clauses: Clauses indicating ownership of datasets, what aggregate statistics or analysis can be published and what cannot be published, what would be the review process and timeline for the data providing authority to review and enable the publishing of any information related to the data exchanged should be included.

b. **Publishing information:** *Rationale:* The rights of the data requesting party and the bona fide collaborators to publish or publicly disclose material or information related to the results of research undertaken must be provided to avoid any limitations, such that they do not violate any of the confidentiality and privacy of data. Clauses: In this section, clauses for transparent public disclosure must be added. Before any materials are published in the public domain or on any platform outside the data owner's purview, the data-providing party can request for - a public disclosure of information about the data ownership, results or analysis undertaken, acknowledgements and citations provided explicitly in all publications and project materials. The data-providing party must have a pre-decided number of days based on the agreement for their review process and ensure that the data are appropriately protected, aggregated and represented.

Concluding Thoughts

Governments responsible for delivering welfare-enhancing public services face numerous decisions: will the new programme address the development need? Are programmes reaching intended beneficiaries and being delivered as designed? Do programmes positively impact developmental outcomes? Traditionally, such decisions were informed by administrators' experience, political considerations, expert opinions, or public sentiment. However, increasingly the notion that data and evidence should guide programme design, implementation, and scale-up decisions is growing.

The Government of India (as well as various state governments) have repeatedly expressed a desire to make effective use of data for decision-making. This is underscored by their investments in initiatives, policies, guidelines, staff and infrastructure to strengthen this effort. This toolkit is envisioned as a resource for staff in government ministries and departments to manage and leverage data for use in policy research and analysis by strengthening administrative data systems and processes.

This toolkit can serve as a framework (or basis) for a conversation between data providers and data users on important data management aspects over the data life cycle that need to be addressed to enable secure and effective use of data for decisions. The preceding sections of the toolkit, therefore, describe the function and best practices relating to data quality, privacy, security and handling over the data lifecycle moving from data generation to use and finally its archiving and destruction.

While these guidelines and best practices are in no way exhaustive and need to be contextualised considering institutional capacity, resources, technical skills and ultimate use by the data owner/stakeholder; they are robust and based on widely accepted and used practices in rigorous academic research. We hope this toolkit contributes to efficient d ata use for policymaking to help achieve the goal of Viksit Bharat by 2047.

References

1. J-PAL NA.2019."The Lessons of Administrative Data: High-Profile Policy-Relevant Research Powered by Administrative Data." Last modified March 2019. https://www.povertyactionlab.org/sites/default/files/research-resources/2019.03.28-The-Lessons-of-Administrative-Data.pdf .

2. Bauman, Jason, Geeti Mehra, Julia Chabrier, Laura Feeny and Michelle Woodford.2015. "Using Administrative Data for Randomised Evaluations." Abdul Latif Jameel Poverty Action Lab North America.December 2015. Accessed at https://www.povertyactionlab.org/resource/using-administrative-data-randomized-evaluations .

3. Cole, Shawn, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. 2020. "Using Administrative Data for Research and Evidence-Based Policy: An Introduction." In: Cole, Dhaliwal, Sautmann, and Vilhuber (eds), Handbook on Using Administrative Data for Research and Evidence-based Policy. First Published January 1, 2021. Accessed at https://admindatahandbook.mit.edu/book/v1.0/intro.html .

4. Groves,Robert M. and George J. Schoeffel. 2018. "Use of Administrative Records in Evidence-Based Policymaking." ANNALS, The American Academy of Political and Social Science, 678, Pgs 71-81. July 2018. Accessed at https://drive.google.com/file/d/1IfNzf_Qm85DiSfeJEAnkhSdN9RW7APVW/view .

5. Gibson, Michael. 2021. "Data Quality Checks". Abdul Latif Jameel Poverty Action Lab. Last modified March, 2021.Accessed at https://www.povertyactionlab.org/resource/data-quality-checks .

6. O'Toole, Elisabeth, Kenya Heard, Laura Feeney and Rohit Naimpally. 2015. "Data security procedures for researchers." J-PAL North America. Last Modified May 2023. Accessed at https://www.povertyactionlab.org/resource/data-security-procedures-researchers .

7. Clark, Julia. 2019. Section III Topics: Privacy and Security .ID4D Practitioner's Guide: Version 1.0 (October 2019). Washington, DC: World Bank. License: Creative Commons Attribution 3.0 IGO (CC BY 3.0 IGO). Accessed at https://id4d.worldbank.org/guide/privacy-security .

8. Schmutte, Ian M., and Lars Vilhuber. 2020. "Balancing Privacy and Data Usability: An Overview of Disclosure Avoidance Methods." In: Cole, Dhaliwal, Sautmann, and Vilhuber (eds), Handbook on Using Administrative Data for Research and Evidence-based Policy. First Published January 1, 2021. Accessed at https://admindatahandbook.mit.

edu/book/v1.0/discavoid.html .

9. Wood, Alexandra, Micah Altman, Kobbi Nissim, and Salil Vadhan. 2020. "Designing Access with Differential Privacy." In: Cole, Dhaliwal, Sautmann, and Vilhuber (eds), Handbook on Using Administrative Data for Research and Evidence-based Policy. First Published January 1, 2021. Accessed at https://admindatahandbook.mit.edu/book/v1.0/diffpriv.html .

10. Machanavajjhala, Ashwin, Daniel Kifer, John Abowd; Johannes Gehrke and Lars Vilhuber.2008."Privacy: Theory meets Practice on the Map."IEEE 24th International Conference on Data Engineering, Cancun, Mexico, 2008, pp. 277-286. April 2008. Accessed at https://ieeexplore.ieee.org/document/4497436 .

11. Korolova, Aleksandra, Úlfar Erlingsson and Vasyl Pihur. 2014. "RAPPOR: Randomised Aggregatable Privacy-Preserving Ordinal Response."CCS '14: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. Pages 1054-1067. November 2014. Accessed at https://dl.acm.org/doi/abs/10.1145/2660267.2660348 .

12. Apple. 2017."Learning with Privacy at Scale." Highlight, Machine Learning Research, Differential Privacy Team. Last Modified December 2017. https://machinelearning.apple.com/research/learning-with-privacy-at-scale

Reference Resources: Data Quality

1. **Data Quality Checks - J-PAL Global**
2. **Data Quality Assessment Tool for Administrative Data**
3. **Data Quality Assurance Toolkit**
4. **Checklist for Quality Evaluation of Administrative Data**
5. **Data Quality Dimensions**
6. **Data Quality Dimensions - ScienceDirect**

Reference Resources: For setting up data security procedures

7. Physically protecting sensitive data - IDEA Handbook
8. Data Classification - Administrative Examples by Harvard University and corresponding requirements for security
9. J-PAL Guide on data security procedures for researchers
10. J-PAL Guide on using administrative data for randomised evaluations
11. Resources from MIT's Information Systems & Technology Department:
    a. Secure computing
    b. Encryption (including software recommendations) and whole-disk encryption
    c. Removing sensitive data

Reference Resources: For creating a Metadata Catalogue (MDC)

1. Generic Template of the MetaData Catalogue
2. Checklist: Microdata Catalog
3. Dataverse (Harvard University)
4. J-PAL Guide to Publishing Research Data

Reference Resources: For Data Handling

1. Data quality — ISO 8000:150 Data quality management: Roles and responsibilities
2. IDEA Handbook Chapter 2: Physically Protecting Sensitive Data
3. UN Data Strategy of the Secretary-General for Action by Everyone, Everywhere

Sample Data Catalogue Template

Category of information	Indicator	Description
BASIC INFORMATION	Name of the dataset	A dataset is an organised collection of data on the same unit in the form of rows and columns. It can be associated with a scheme or not. For the MDC, each row will represent a unique data set. Naming the dataset: 1. If it is part of an MIS then the name of the data set = "Name of the MIS" + "Name of the module within the MIS". 2. If it is part of a scheme then the name of the data set = "Abbreviation of Scheme" + "name of module/universe of the data set (e.g applicant/ beneficiary/project)" 3. If it is a standalone data set then the name can just signify the purpose of the data set
	Name of the department	Name of the department filling the MDC
	Name of Programme/ Scheme	Full name of the programme/ scheme
	Data owner	name the department/ ministry/ agency that has complete access and decision authority on the sharing and use of data
	Sector	
	Brief description of the purpose of the dataset	A short text description of the primary purpose(s) of the dataset

Category of information	Indicator	Description
	Geographical extent of the dataset	Mention whether this is for the entire state or a few geographical regions. If only for a few geographical regions then mention the names (if the list is not too extensive)
	Is similar data available in other states? E.g. nationwide MIS	
	If linked to an MIS, write the name of the MIS	Not all data sets may be linked to an MIS. Write NA if not linked to an MIS
	If linked to a scheme/ programme, the name of the scheme/ programme	Not all datasets may be linked to a specific scheme. Write NA if the data set is not linked to a scheme.

Category of information	Indicator	Description
DATA CONTENTS	Data description	Describe any other descriptive information about the dataset - what is it used for, summary of what information it contains and types of data collected
	Geographical coverage	
	Most granular units of observation	Specify the lowest unit (or level) for which data are collected, such as individual, household, school, district, farm holding, transaction etc.
	Unique ID of the lowest unit of observation	What ID is collected for the lowest data unit - e.g. if an individual, then an Aadhaar, phone number, ration card etc. If an institution like a firm/ hospital - registration number / / PAN/ TIN/ Department ID etc.
	List of geographical identifiers in the data set	List out all the geographical identifiers in the data set (e.g. address/village/facility/area, block, district)
	Code directory used for geographical identifiers	Drop-down: LGD, Census, State directory, Others, NA
	Personal Identifying Information (PII fields) in the dataset	List out the fields that can be used to identify the individual units in the data set (e.g. name, address, geo-location, phone numbers etc.). If the data set is not at an individual/entity level or no PII fields are available then write NA.
	Information on which of the common IDs is collected in this data set?	A drop-down list of common IDs: Aadhar, Ration Card #, State Family ID, Health ID [needs to be populated]
	List of other key fields (if available)	
	The scale of data (# obs, size) as on the last update	Enter the # of observations in the data set. If not known then enter the data size in MB/GB/ TB.

Category of information	Indicator	Description
DATA COLLECTION	First point of data collection/entry	Specify who is collecting the information and populating the dataset at the point of entry. Dropdown - Frontline workers/eSeva Kendras/ Surveyors/Village officials/Block officials Other (specify)
	Mode of data collection at the point of entry (digital/paper)	
	If collected on paper at the first point of entry, then specify who digitises the data (OPTIONAL)	ONLY IF data are COLLECTED ON PAPER - Drop-down: digitised by entry into Excel/ Google Sheets; digitised by entry into MIS software
	The data universe/ population on which data are collected	Specify here the population on which data are being collected i.e. eligibility criteria for inclusion of the individual/entity into the database. E.g., if the data are on workers, specify which type of workers and their eligibility for inclusion. If the data are collected at a facility level e.g. PHC, then the population will be those accessing services at the facility.
	Frequency of data collection	
	If real-time/period, specify the intended frequency of update of data entry/capture [OPTIONAL]	List the minimum frequency of desired updates - e.g. daily, at the point of transaction/ record generation, weekly, monthly, quarterly, bi-annually, annually etc.
	If real-time/periodic entry, specify the actual frequency of updates in the past 3 months [OPTIONAL]	Describe update frequency for the past 3 months - not updated at all/ discontinued (only for historical data), daily, weekly, quarterly, annually etc.
	Earliest year of data collection	Specify the year in which the data was collected the first time.
	Latest year of data collection	Specify when the data was updated last. In the case of one-time data collection, it will be the same as the earliest year of data.

Category of information	Indicator	Description
DATA DOCUMENTATION, QUALITY & USE	Is there a codebook for the data set? [A codebook is a detailed documentation with definitions of all the fields and the composition of the dataset and table structures]	
	Describe data validation checks or measures undertaken	Describe all types of validation and data quality measures used - at the field level (audits/ checks etc.) and dataset level (logical validations, data cleaning)
	Describe any known data quality issues [OPTIONAL]	Describe issues in terms of missing data, populations not covered, unused fields etc.
	Are Dashboards available for this data set?	
	Brief description of reports generated (OPTIONAL)	

Category of information	Indicator	Description
DATA ACCESS	Data access type for line listing - open, closed or restricted for public	
	If openly available, specify a link to the data source (raw, aggregate etc.)	If the data are openly available, specify the link to the data source
	Data access type for aggregate reports or dashboards - open, closed or restricted for public	
	If openly available, specify the link to the data source for aggregate reports/ dashboards	If the data are openly available, specify the link to the data source
	If neither line listing nor aggregate data are openly available, please specify the reason	Explain if it is on a negative list/ restricted use/ has sensitive information/ is not feasible because of the size/ nature of data etc.
	Is access to line listing raw data for all units available to the state department/agency?	
	Formats in which the data are available	Please enter all formats in which the data are available: .xls, .xlsx, .csv, .txt., json,.pdf others specify
	Are APIs available for this data set?	
	Name & designation of nodal officer for this dataset/MIS/scheme	Please specify the name of the nodal officer responsible for this data set or MIS. This person can be contacted for more information on access and other aspects

Category of information	Indicator	Description
DATA ARCHITECTURE	What is the database management system being used?	
	What is the database hosting location?	
	List all the stakeholders involved in data management at the state/district level	List other stakeholders involved in managing, maintaining and updating the dataset. Some examples are IT vendors, NIC, in-house MIS staff, Project Management Unit etc. These agencies may or may not have ownership but are primarily responsible for managing and maintaining the dataset. Also e.g., the ownership may be the central ministry, but maintained by the state dept.
OTHER INFORMATION	Information accurate as of date	
	Name of Officer filling MDC	
	Designation of Officer Filling MDC	
	Contact details of Officer filling MDC	
	ANY OTHER COMMENTS ABOUT THE DATA SET/ PROCESS OF FILLING MDC	

सत्यमेव जयते

NITI Aayog