# Data Handling

## 3.1: 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 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 on 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 unauthorized users.

1. Assigning a personnel or nodal officer from within the department or agency.
2. Creating a data collection format (while keeping data use in mind)
3. Listing down the existing datasets and prioritizing the ones that need to be catalogued
4. Compiling the information on the data catalog comprehensively with clear definitions

### *Step 1: Understanding the medium and longevity of data storage*

Deciding how long the data or dataset for a particular scheme or program should be kept for 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 digitized or microfilmed. Institutional guidelines may require that data be retained for longer periods. Considering the longevity along with the medium of storage is the first step before handling any data. The storage of the data or dataset 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, In addition to this, 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 the 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. Examples of sensitive data include data with any personally identifiable information such as Aadhaar ID, mobile number, residential address etc. Examples of non sensitive data include data with any other characteristic that may not pose any risk of identification, such as number of household members, schemes enrolled in, asset profile etc. 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 unauthorized access is high. Several years of valuable data can be compromised or lost, if any intruder or unauthorized 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 needs 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 analyzed? 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 |

The functions and responsibilities for 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 to ensure 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. Authorised 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 and correct the data to match with the predefined standards and handle enquiries about data. Usually input and track the data and not edit the data retrospectively. Are able to 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 unauthorized 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 analyze the data periodically to ascertain that the information captured is valid and reliable.

## 3.2: Data Sharing Standards

There are multiple forms of data exchange possible between the government and other partners. For data exchange between government departments, standardizing 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 own personal data on scheme eligibility, benefits received, entitlements, certificates etc. Or, the government may want to, for accountability, set-up a website, dashboard to allow citizens in general and civil society to monitor, understand how government programs are being implemented, expenditure and utilization 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 from enrolment 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 exchange of data with tech vendors to populate MIS other tech solutions for program delivery. Alternatively, data could be shared with academics, researchers, think tanks and civil society organizations 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 with respect to 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.

## 3.2.1: 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 or 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 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?

• Format of exchange: The format that will be used to transfer the data, such as:
CSV, comma separated file
TXT, plain text file
SQL, query for the relational database
Data Interchange Format
Open Document Format
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?

## 3.2.1: 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 from time to time, in order 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. As 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 its description, current ownership of these datasets, storage platforms 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 group of dataset 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/ processing (access 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 transfer of relevant data through physical hard drives or on-line (web access). Clauses: The data providing party would state how relevant data as requested and agreed upon in the scope of work will be shared with the data requesting 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 transparency 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 regulation.
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 the 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 unauthorized 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 ensuring that the data are appropriately protected, aggregated and represented.

Reference


• Data handling also translates into reduction of the data security threat-level a priori by acquiring and handling only the minimum amount of sensitive data strictly needed for the analysis. (Data Security Procedures_2 (NIL)*
• Handbook on Using Administrative Data for Research and Evidence-based Policy: Model Data Use Agreements and Sample Text for Agreement Components*

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


Back to the main site

Designed by Jpal