



NATIONAL **DATA** **QUALITY** AND STANDARDS **GUIDELINES**

November 2025

TABLE OF CONTENT

Acknowledgment

Introduction _____ 4

Background and Rationale _____ 5

Importance of Data quality in Governance _____ 5

Purpose, Scope and Applicability _____ 7

Purpose _____ 7

Scope and Applicability _____ 7

Data Quality Principles

Data Quality Dimensions _____ 11

Completeness _____ 12

Uniqueness _____ 13

Consistency _____ 14

Timeliness _____ 15

Validity _____ 16

Accuracy _____ 17

Integrity _____ 18

Relevance _____ 19

Accessibility _____ 20

Metadata, Documentation & Standard—21

Role of Metadata in Data Quality _____ 21

Documentation and Standard Operating Procedures (SOPs) _____ 21

1

2

3

4

5

1

TABLE OF CONTENT

Integration of GSBPM in Data Quality Management	22
Use of Standard Classifications and Coding Systems	29
Data Interoperability and Open Standards	29

Data Quality Roles & Institutionalization 31

Quality Assurance Rules & Data Standard 33

Core Data Quality Management Processes	33
Data Collection and Entry Standards	34
Data Processing, Cleaning, and Validation Procedures	35
Mandatory Fields and Data Standards	36

Annex Table 39

Data Quality Table	39
Definition and Key Concepts	40
Example of Data Quality KPIs by Dimension	41

6

7

8

Acknowledgment

The development of the National Data Quality and Standards Guidelines follows the approval of the National Data Sharing Policy and the National Data Governance Framework. These guidelines provide a structured approach to ensure that data across government institutions is collected, managed, and used according to consistent quality principles and recognized standards. They aim to enhance the reliability, comparability, and integrity of data that support evidence-based policymaking and effective public service delivery.

This work was coordinated by the National Institute of Statistics of Rwanda (NISR), with valuable technical input from the Rwanda Information Society Authority (RISA), Cenfri, and the Ministry of ICT and Innovation (MINICT). Their contributions were instrumental in aligning these guidelines with Rwanda's Digital Transformation Agenda and the broader goals of the National Statistical System (NSS).

NISR acknowledges the active engagement of sector institutions, data producers, and other key stakeholders whose insights and feedback strengthened the relevance and practicality of these guidelines.

The completion of the National Data Quality and Standards Guidelines marks an important milestone toward strengthening data governance, promoting adherence to quality standards, and ensuring that data serves as a trusted and strategic asset for Rwanda's sustainable development.

1 Introduction

High-quality data is essential for Rwanda's transition to a knowledge-based economy. Reliable information underpins sound policies, strengthens public service delivery, and drives sustainable development. Data today is recognized globally as a strategic national asset, fueling innovation, informing governance, and shaping economic transformation.

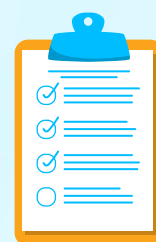
Despite its importance, many government institutions face challenges related to unknown, inconsistent, or poor-quality data. This undermines evidence-based decision-making, erodes public trust, reduces operational efficiency, and limits the potential value of information. To fully harness the power of data, it is imperative to ensure its quality from the outset.

Law No. 45/2013 of 16 June 2013 governing statistical activities in Rwanda, along with the United Nations Fundamental Principles of Official Statistics, emphasize the importance of data quality, defined as the extent to which data is fit for its intended use. Managing data quality involves more than correcting errors; it is about preventing them through proactive planning, robust systems, and continuous quality checks. However, there is currently no standardized approach to data quality management across all public institutions in Rwanda.

To strengthen national consistency, the National Institute of Statistics of Rwanda (NISR) has developed the National Data Quality and Standards Guideline. This guideline provides practical measures for public institutions to improve how data is collected, managed, safeguarded, and shared. While grounded in international standards, it is tailored to Rwanda's institutional context and development priorities.

The guidelines seek to:

- Promote a culture of quality throughout the data lifecycle.
- Address quality issues at the source rather than at the output stage.
- Establish regular mechanisms for monitoring and reporting.
- Concentrate institutional efforts on areas with the greatest impact.
- Define clear responsibilities for data stewardship and accountability.



Adopting these measures, government institutions will improve the credibility, comparability, and usability of their data. This will enhance efficiency, enable interoperability across systems, and support evidence-based governance, ensuring that Rwanda's data ecosystem contributes fully to national development and international commitments.

1.1 Background and Rationale

Over the past decade, Rwanda has modernized its data environment through initiatives such as strengthening the national statistical system through a national strategy for the development of Statistics, developing sectoral Management Information Systems, and expanding digital data platforms. These efforts have improved coverage and accessibility, but challenges persist:

- Fragmented systems with limited interoperability.
- Inconsistent methodologies and classifications across institutions.
- Incomplete or inaccurate administrative records.
- Weak accountability for data stewardship.
- Quality assurance processes that detect errors too late.
- Capacity gaps in data management and control.

Such issues reduce the effectiveness of government operations and limit the availability of reliable statistics for monitoring the National and international development agendas and other obligations.

This guideline responds to those challenges by:

- Providing a unified framework for data quality management.
- Ensuring consistent practices across public institutions.
- Aligning with international standards such as ISO 8000, GSBPM, GSIM and DAMA DMBOK 2
- Reinforcing Rwanda's commitment to evidence-based decision-making and efficient service delivery.



1.2 Importance of Data Quality in Governance

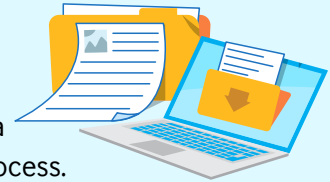
Data is only valuable when it is accurate, timely, and fit for purpose. Poor data quality undermines evidence-based decision-making, leads to inefficient resource allocation, and erodes public trust in institutions. Conversely, strong data quality management enables government institutions to design better policies, allocate resources effectively, and monitor progress toward national and global development goals.

High-quality data is the cornerstone of good governance. It supports informed policymaking, improves service delivery, enhances accountability, and fosters innovation by providing the private sector with reliable information for planning and investment.

To ensure that data effectively supports these functions, public institutions must implement structured data quality rules, standards, and management processes within their institutional data governance frameworks. These practices must be embedded throughout the entire data lifecycle, from collection and processing to dissemination and use.

The overarching aim of data quality management is to:

- Ensure data is fit for purpose and meets key quality dimensions.
- Promote standardized practices across institutions to facilitate comparability, integration, and reuse of data.
- Embed quality assurance controls into daily operations and information systems rather than treating them as a separate or reactive process.



Ultimately, investing in data quality is both a technical requirement and a strategic priority for effective governance. Reliable data strengthens institutional efficiency, promotes transparency, and ensures that Rwanda's development is guided by sound, credible evidence.

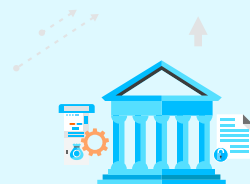
2 Purpose, scope, and applicability

2.1 Purpose

The purpose of this Guideline is to establish a consistent and comprehensive framework for managing data quality across all Government of Rwanda (GoR)¹ institutions. It defines the fundamental principles, minimum requirements, and roles and responsibilities necessary to ensure that data collected, processed, and used is accurate, reliable, and fit for its intended use. It aligns with Rwanda's data governance framework, and it provides practical guidance for embedding data quality into routine operations as well as long-term strategic planning.

Specifically, this Guideline enables GoR institutions to:

- Develop internal systems and practices for continuous data quality improvement.
- Comply with national and institutional data management obligations.
- Integrate quality assurance measures at every stage of the data lifecycle.
- Promote coordination, standardization, and interoperability across the national data ecosystem.



2.2 Scope and Applicability

This Guideline applies to all GoR institutions that generate, manage, or use data. It covers all data types and addresses every stage of the data lifecycle. All information systems and platforms handling data are within its scope. While institutions may maintain internal procedures for their operations, they must adhere to the minimum standards defined in this Guideline. It is intended for all personnel involved with data who are responsible for understanding, implementing, and ensuring compliance with these standards in their respective roles.

¹Government of Rwanda (GoR) in this framework means all public institutions and government-affiliated entities in Rwanda, including ministries, departments, agencies, local governments, and state-owned enterprises.

3 Data Quality Principles

The following principles are designed to foster a strong culture of data quality within Rwanda's public institutions. They serve as foundational guidelines to support consistent, accurate, and purposeful data management, ensuring that data is fit for its intended use. Each principle is accompanied by practical actions that institutions and individuals can adopt. These principles should guide how data is managed across government and be embedded in everyday practices.

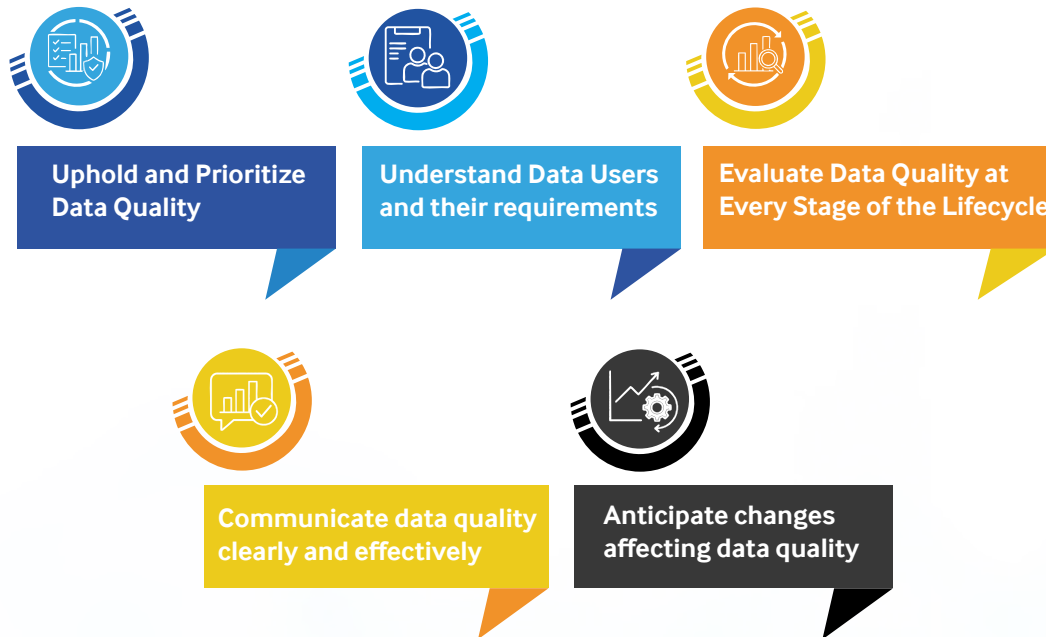


Figure1: Data Quality Guiding Principles

Principle 01

Uphold and Prioritize Data Quality

A strong commitment to data quality begins with clear accountability at all levels of government. Every institution must take responsibility for continuously assessing, improving, and reporting on the quality of the data it produces or manages. This commitment requires the integration of sound data governance and management practices into institutional workflows.

Institutions should adopt formal data governance structures, adhere to national data principles, and apply standardized data formats to promote interoperability and reuse. Leadership must champion these efforts by ensuring staff understand the importance of data quality and that adequate resources are allocated for its management.

Capacity building is essential. Institutions should invest in training staff on best practices in data quality management, including the application of quality dimensions. Continuous improvement should also be a core focus; institutions are encouraged to assess baseline data quality, monitor progress over time, and implement data quality action plans that prioritize areas with the highest impact.

Principle 02

Understand Data Users and their requirements

Understanding who uses your data, and for what purpose, is essential to ensure it is fit for use. In Rwanda's public sector, data users are not limited to the general public or external stakeholders; they often include other government ministries, departments, and institutions that rely on shared data for planning, monitoring, service delivery, and reporting. Users may also include policymakers, analysts, development partners, civil society organizations, private sector actors, and citizens seeking localized or thematic insights.

Each user group may have different expectations and data quality requirements depending on how they intend to use the data. For example, one institution may prioritize timeliness for operational decisions, while another may value completeness for long-term planning. Conflicting needs may arise, especially when data is used for multiple purposes.

Public institutions should therefore engage proactively with their users, both internal and external, through consultations, feedback mechanisms, inter-agency coordination platforms, or targeted research. Understanding user priorities helps institutions align data production and quality assurance with real-world needs.

Additionally, user needs are not static. They may evolve due to changes in policy, emerging technologies, or shifting national priorities. Institutions should establish regular communication channels to monitor these evolving needs and adjust their data processes accordingly to maintain relevance and usability.

Evaluate Data Quality at Every Stage of the Lifecycle

Data quality must be monitored across all stages of the data lifecycle, from design and collection through to processing, analysis, dissemination, and eventual archiving or destruction. At each stage, different quality risks may emerge, and these must be proactively addressed.

Quality assurance should be integrated into every phase of data handling. Institutions should adopt appropriate quality checks at each stage, rather than applying a single approach to all situations. Early detection and resolution of quality issues, especially at the point of collection, significantly reduce downstream errors and improve overall data reliability.

Principle 04

Communicate data quality clearly and effectively

Transparency about data quality builds trust and ensures that users interpret and apply data correctly. Institutions must provide clear, accessible information about the quality of the data they publish, including explanations of any trade-offs or known limitations.

Communication should be tailored to the needs and understanding of the intended audience. This means using plain language, avoiding technical jargon, and ensuring documentation is thorough and well-organized. Institutions should also provide metadata that clearly defines data elements, describes quality assurance processes, and highlights known data quality issues.

Where data is sourced from external providers, it is critical to maintain strong relationships and work collaboratively to identify and resolve quality concerns at the source. When changes are made to data processes, such as system upgrades or changes in methodology, users should be informed in advance and supported in understanding the implications.

Principle 05

Anticipate changes affecting data quality

While not all data quality issues can be predicted, institutions should aim to anticipate and manage changes that could impact data quality. This includes planning for system upgrades, policy changes, or new data collection methods that could affect existing datasets.

To minimize risk, institutions should integrate data quality considerations into the design of new systems and processes from the outset. Root cause analysis should be used to resolve recurring quality issues at their source, rather than relying on short-term fixes. Metadata and support documentation should be updated regularly to reflect changes and ensure continued clarity.

Regular communication with users can also help institutions stay informed about shifting expectations and quality requirements, allowing for proactive adjustments and sustained relevance of government data.

4 Data quality dimensions

Data quality dimensions are the measurable characteristics that describe how good your data is. The Data Management Association (DAMA) defines them as the “measurable features or characteristics of data” that can be used to assess quality and detect issues.

These dimensions help GoR institutions evaluate whether their data is fit for its intended purpose. Each dimension focuses on a specific aspect of data quality, such as completeness, accuracy, or timeliness, that directly affects the reliability, usability, and trustworthiness of the data.

By regularly measuring and monitoring these dimensions, institutions can:

- Detect and correct problems early.
- Ensure that data meets national standards.
- Improve decision-making and service delivery.

The following nine core data quality dimensions are recommended for use across GoR to guide data quality assessment and improvement. Each dimension includes a definition, good practices or applicable rules, and practical examples to help understand.

Completeness

Uniqueness

Consistency

Timeliness

Validity

Accuracy

Accessibility

Relevance

Integrity

Figure2: Data Quality dimensions

4.1 Completeness

Completeness refers to the extent to which all required data is present in a dataset. A dataset is considered complete when all expected records are included (record-level completeness). All essential fields within each record are filled (attribute-level completeness).

Important note: a dataset may be fully complete but still contain incorrect values.

Example:

The Ministry of Education conducts a school census. If a school has 500 students but only reports data for 470 students, the dataset is incomplete. Likewise, if key fields such as a student's age, sex, or village are left blank, the data is also incomplete. Such gaps can affect planning, for example, determining the number of classrooms, teachers, or textbooks needed.

Good Practices for Ensuring Completeness

1 Mandatory Fields:

Define and enforce completion of all critical fields before records can be saved.

3 Avoid Placeholder Values:

Do not use "N/A," "Unknown," or default numbers where actual data is available.

5 Regular Completeness Checks:

Run periodic (e.g., monthly) checks to identify missing or incomplete data.

7 Clear Data Standards:

Assign "mandatory" and "optional" labels to each field or variable in a dataset, based on its importance.

9 Metadata for Context:

Include metadata describing the dataset's purpose, scope, and acquisition process. Metadata should note any privacy, confidentiality, or accuracy constraints affecting completeness.

2 Full Record Coverage:

Ensure all expected records for the reporting period are submitted.

4 Automate completeness checks:

Integrate alerts or messages when data is missing or incomplete.

6 Data Entry Monitoring:

Supervisors should review data entry during collection and reporting periods to ensure completeness.

8 Up-to-date Standards and Classifications:

Keep values, definitions, classifications, and methodologies current to ensure completeness aligns with the latest requirements.



4.2 Uniqueness

Uniqueness measures the extent to which each record in a dataset represents only one distinct entity, with no duplicates. A dataset is unique when everyone, household, or entity appears only once, and each unique identifier is stored only once.

Duplicates can lead to double-counting, misreporting, and inefficient resource allocation, for example, providing the same service multiple times to the same person.

Example:

In a national vaccination tracking system, the Ministry of Health expects 10,000 patient records but finds 50 people recorded twice. This creates 50 extra records, making the total 10,050.

Uniqueness rate calculation: $10,000/10,050 \times 100 = 99.5\%$.

Good Practices for Ensuring Uniqueness

1 Unique Identifier :

Assign each record a permanent, unique ID. All Rwandan citizens have a unique National Identification Number (NIN) in the National Identification Agency (NIDA)'s database.

2 Duplicate Detection at Entry :

Implement automated duplicate checks before a record is saved.

3 Regular Data Cleaning:

Conduct scheduled like monthly or quarterly reviews to merge or remove duplicates.

4 Cross-System Validation:

Compare identifiers against other authoritative systems to capture duplicates.

5 Standardized Naming Conventions:

Use consistent name formats to avoid creating duplicates caused by spelling variations. "Jean Claude" and "J. Claude". Where possible, use autofill from NIDA.

4.3 Consistency

Consistency measures the extent to which data does not contradict itself, either within a dataset or across different datasets. Data is consistent when the same entity (person, household, facility, etc.) has the same information in all relevant systems, and related fields logically align.

Inconsistencies reduce trust in data, create confusion, and make data integration across systems difficult.

Good Practices for Ensuring Consistency

1 Internal Logical Checks

Validate that related data fields make logical sense.

3 Standard Code Lists and Formats:

Use official, standardized code lists and consistent formatting for names, locations, and classifications. For example, always use “Gasabo” instead of “GSB” or “GASABO DIST.” And store all phone numbers in the +250 format.

5 Automated Validation Rules:

Implement automated processes to check for inconsistencies in real time. For example, a system blocks an entry if the employment start date is earlier than the recorded birth date.

2 Cross-System Alignment:

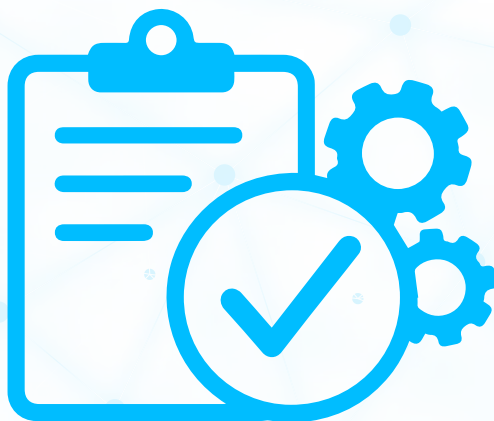
Ensure core identifiers and key attributes match across connected systems. For example, the Ministry of Health and NIDA should record the same date of birth for the same individual.

4 Regular Data Reconciliation:

Compare and align datasets from different institutions or systems on a routine basis.

6 Issue Tracking and Continuous Improvement:

Maintain a log of detected inconsistencies, review recurring issues, and update validation rules accordingly.



4.4 Timeliness

Timeliness describes the degree to which the data is an accurate reflection of the period that they represent, and that the data and its values are up to date. Some data, such as date of birth, may stay the same, whereas some, such as income, may not. Data is timely if the time lag between collection and availability is appropriate for the intended use.

Timeliness refers to how current the data in a system is and whether it is updated quickly enough to remain relevant for decision-making. The shorter the time between when data changes and when the system reflects those changes, the more timely it is.

If a dataset is not updated regularly, it becomes outdated and unreliable. This can result in incorrect reports, wrong planning decisions, or inefficient service delivery.

Good Practices for Ensuring Timeliness

1 Set Clear Update and Submission Deadlines:

Define the maximum allowable time between data collection and entry, and between reporting periods and submission deadlines.

3 Monitor and Follow Up on Submissions:

Track submission status daily or weekly and follow up with late submitters.

5 Performance Monitoring and Feedback:

Regularly review and share timeliness performance to recognize good performers and address delays.

7 Coordinate with Data Providers:

Work with providers to confirm capacity to meet timelines, address potential delays early, and keep users informed. Maintain a documented data release schedule that covers all production stages and accounts for potential bottlenecks.

2 Promptly Record Key Changes:

Update critical information

4 Automated Reminders and Alerts:

Use system notifications to remind data providers before and after deadlines.

6 Plan with User Needs in Mind:

Identify and align timelines with user requirements, including reference periods, legislative deadlines, and service standards.



4.5 Validity

Validity describes the degree to which the data is in the range and format expected. For example, the date of birth does not exceed the present day and is within a reasonable range. Valid data is stored in a data set in the appropriate format for that type of data. For example, a date of birth is stored in a date format rather than in plain text.

Validity means data is recorded in the correct format, and the values make sense. Each data element should follow the rules defined for it.

Good Practices for Ensuring Validity

1 Format Checks:

Enforce standard formats. Birth dates in all systems must be in YYYY /MM/DD format.

3 Drop-Down Menus

Use predefined lists to avoid invalid entries. Select “Male” or “Female” instead of free text for gender. Select Province, district, and other administrative location instead of writing them.

5 Pre-Loaded Reference Data:

Use existing lists to validate inputs. For example, facility codes in DHIS2 are validated against the national health facility list.

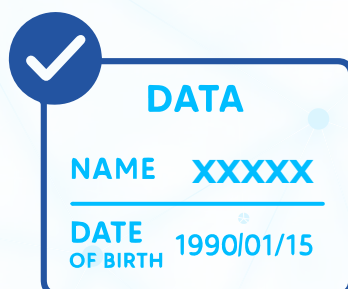
2 Range Checks:

Ensure numerical values fall within realistic ranges.

4 Automatic Rejection

Reject entries that fail validation rules. For example, reject a phone number with fewer than 12 digits.

VALIDITY



DATA

NAME	XXXXXX
DATE OF BIRTH	1990/01/15

FORMAT



4.6 Accuracy

Accuracy measures the extent to which data correctly reflects the real-world situation it represents. Data is accurate when recorded values match the true values, free from errors or distortions. Bias can undermine accuracy if the data systematically misrepresents certain groups or situations. Where bias exists, it should be documented and communicated to data users.

Accuracy checks can be performed at the record level (verifying individual entries) or dataset level (assessing overall correctness), depending on the data's purpose and use case.

Good Practices for Ensuring Accuracy

1 Source Verification:

Ensure recorded data matches original, authoritative source documents.

3 Spot Checks and Field Verification:

Periodically verify a sample of records against ground truth.

5 Double Data Entry for Critical Data:

For high-value datasets, use independent dual entry and compare results.

7 Use of Metadata for Context:

Document and share metadata that describes the data's source, collection method, processing steps, revisions, and coverage so users can assess accuracy.

9 Outlier Review:

Investigate extreme or unusual values and provide documented explanations to data users.

2 Validation and Error-Checking Rules:

Apply automated and manual rules to detect incorrect values, invalid ranges, or inconsistencies.

4 Error Reporting and Correction Mechanisms:

Provide channels for users to report errors and request corrections.

6 Feedback Loops to Data Providers:

Share error findings with collectors or institutions so they can correct inaccuracies and improve future submissions.

8 Representativeness Checks:

Confirm that the dataset adequately represents the geographic, demographic, or thematic domains it covers.



4.7 Integrity

Integrity measures the extent to which relationships between data elements are valid, complete, and preserved over time. In a high-integrity dataset, all records are properly linked to their related entities, and no relationships are missing or broken.

Example:

When integrity is compromised, for example, when a child's birth certificate in Irembo is not correctly linked to the parents' National ID numbers, systems cannot reliably match or verify records. This can lead to service delays, processing errors, or incorrect reporting.

Good Practices for Ensuring Integrity

1 Preserve Data Links:

Maintain correct relationships between connected records, for example, child and parent, households and members, or facility and district.

2 Prevent Orphan Records

Disallow saving records that require, but lack, a valid linked record. For example, no patient record can be created without a valid health facility code in DHIS2.

3 Regular Referential Integrity Checks:

Run scheduled checks to detect missing or broken relationships between linked data tables or systems.

4 Protect Linked Data from Accidental Deletion:

Prevent deletion of a record if dependent records still exist. For example, A school record cannot be deleted if student records are still linked to it.



4.8 Relevance

Relevance measures the extent to which data collected is useful, appropriate, and necessary for current policy, program delivery, or legal purposes. Data is relevant when it supports decision-making, meets stakeholder needs, and aligns with institutional or national priorities. Irrelevant data increases the burden of collection, wastes resources, and distracts from the real objectives of data programs.

Example:

In a population and housing census, a question like “What is your favorite type of music?” would be irrelevant, since it doesn’t support demographic, social, or housing indicators used for planning and policy.

Good Practices for Ensuring Relevance

1 Assess and Define Data Needs:

Conduct a needs assessment to determine exactly what data is required for decision-making, policy monitoring, or legal reporting.

3 Engage Stakeholders Early:

Consult end-users and key stakeholders before adding or modifying data fields. NISR consults with ministries before introducing new census/survey questions.

5 Eliminate Redundant Collection:

Avoid re-collecting data that already exists in authoritative systems. Use existing NIDA data instead of re-asking for National ID numbers.

7 Measure and Monitor Usefulness:

Track how datasets are used and reused to assess their continuing relevance and adjust collection plans accordingly.

2 Align with Policy and Strategic Goals:

Ensure all collected data supports national strategies, institutional mandates, and operational objectives.

4 Review and Update Regularly:

Conduct periodic reviews of data collection tools to remove outdated, unused, or low-value fields.

6 Leverage Existing Data Inventories:

Examine existing data catalogues to identify available sources and prevent unnecessary duplication.

8 Comply with Legal Authority:

Ensure the institution has legal authority to collect personal data and that it directly relates to an approved program or activity.

4.8 Accessibility

Accessibility measures the ease with which data can be discovered, obtained, and used by authorized users for legitimate purposes. Data is accessible when it is available through well-documented platforms, open or controlled channels, and in formats that allow efficient retrieval, interpretation, and integration. Limited accessibility restricts evidence-based decision-making, discourages data sharing, and undermines public trust in statistical systems.

Example:

Suppose administrative data on business registrations is stored in separate ministry databases without standardized formats or secure data-sharing protocols. In that case, statisticians cannot efficiently access or integrate the information for further analysis.

Good Practices for Ensuring Accessibility

1 Establish Clear Access Policies and Protocols:

By following national data classification and access guidelines, define transparent procedures governing who can access which data, under what conditions, and for what purposes, balancing openness with confidentiality and security.

2 Ensure Metadata and Documentation Availability:

Follow national metadata guidelines to provide comprehensive metadata that describes data sources, collection methods, formats, and usage rights, enabling users to understand and apply the data appropriately.

3 Adopt Open Standards and Machine-Readable Formats:

Store and disseminate data using non-proprietary, interoperable formats (e.g., CSV, JSON, XML) to facilitate reuse and integration across systems.

4 Promote Capacity Building and User Support:

Train data providers and users on access procedures, tools, and data stewardship practices to enhance utilization and ensure responsible use.

5 Metadata, Documentation, and Standards

Metadata, documentation, and standards ensure that data produced and managed across government institutions is consistent, interoperable, and fit for purpose. This chapter sets out the key requirements for metadata management, documentation, and the integration of international standards, while providing a structured framework for aligning all processes to the Generic Statistical Business Process Model (GSBPM). Detailed technical guidance is provided separately in the National Metadata Management Guideline.

5.1 Role of Metadata in Data Quality

Metadata provides the necessary context, structure, and meaning of data, enabling users to interpret and apply it correctly. Good metadata strengthens transparency, comparability, and quality assurance across government institutions.

Requirements:

- All institutions shall adhere to the National Metadata Management Guideline published by NISR.
- Metadata must be maintained for all datasets, including administrative, survey, census, and geospatial data.
- Metadata shall be captured in both machine-readable and human-readable formats and integrated into the National Data Catalogue.
- Metadata should be updated whenever data is revised or released.

5.2 Documentation and Standard Operating Procedures (SOPs)

Proper documentation ensures clarity, reproducibility, and accountability, while SOPs standardize operations across institutions. Together, they serve as the operational link between metadata and data quality.

Requirements:

- Each institution shall develop and maintain SOPs covering data collection, validation, processing, analysis, storage, dissemination, and confidentiality.
- SOPs must align with the National Data Governance Framework and explicitly map to relevant GSBPM phases.
- Documentation shall reference metadata elements to ensure coherence between institutional practices and the Metadata Guideline.
- SOPs must be reviewed at least every three years and updated to reflect new technologies, methods, and standards.

5.3 Integration of GSBPM in Data Quality Management

The Generic Statistical Business Process Model (GSBPM), developed by UNECE, provides a comprehensive framework for structuring statistical and administrative data processes. It standardizes the production cycle into eight phases, which are Specify Needs, Design, Build, Collect, Process, Analyse, Disseminate, and Evaluate. By adopting GSBPM, government institutions ensure a common language for collaboration and interoperability. Systematic quality assurance at every phase of the data lifecycle. Harmonization of processes across institutions and with international practice. Each phase contains sub-processes that guide the transformation of raw administrative records into statistical outputs that are relevant, accurate, consistent, accessible, and timely.



Figure3:GSBPM Processes

5.3.1 GSBPM Phases and Subprocesses

The framework adapts the Generic Statistical Business Process Model (GSBPM) to administrative data production. It divides the statistical production process into eight key phases and sub-processes as summarized in the table below:

Core Phases	Processes
<p>Specify Needs</p> <p>For producers of administrative data, Phase 1 ensures that system design and improvement are guided by both institutional needs and statistical needs.</p>	<p>1. Identifying operational and statistical requirements:</p> <p>Before establishing or upgrading any administrative data collection system, institutions must first identify their needs by assessing what data is required both for operational purposes and for producing statistics, while also reviewing best practices from other agencies or countries with modern systems.</p> <p>2. Consult and Confirm Needs: Institutions should engage both internal stakeholders (such as IT, operations, planning, and policy units) and external stakeholders (<i>including NISR and RISA</i>) to clearly define the purposes of data collection, the key variables and identifiers required, the desired frequency, and format. These consultations should also establish agreed-upon quality measures, such as timeliness, completeness, accuracy, and confidentiality, while ensuring compliance with relevant legal and policy frameworks, including the Statistics Law and the Personal Data Protection Law.</p> <p>3. Setting clear output objectives: The institution should define the expected output of the new system, covering both operational outputs and statistical outputs. Clearly specifying these objectives ensures that the system is designed to meet both administrative functions and statistical needs.</p> <p>4. Identify Concepts: The Institution should ensure clarity on the concepts and standards that will underpin system development, covering both operational definitions and statistical concepts. The concept should be aligned with international standards such as ISIC, ISCO, ISCED, ILO definitions, etc, where applicable.</p> <p>5. Check Data Suitability: step involves evaluating their compatibility with technologies such as databases, APIs, or mobile platforms, and identifying requirements for ICT infrastructure, staff capacity, and data security. It also reviews the legal basis for data sharing across agencies and highlights gaps, proposing solutions such as linking with the national ID system to improve accuracy and coverage.</p> <p>6. Prepare and Submit Business Case: The institution develops and presents a project document describing the current situation without the system (manual records, fragmented sources, or data gaps) and the proposed new system (integrated, standardized, interoperable) or comparing the current system with the proposed upgraded version, outlining benefits, required resources, and potential risks, and submits it for visa approval to NISR and RISA.</p>

Core Phases	Processes
<p>Design phase</p> <p>The Design phase describes the activities needed to ensure that administrative systems produce high-quality data that meet operational and statistical needs. It includes the design of outputs, variables, collection methods, processing, and workflows, with strong consideration of national and international standards.</p>	<p>1. Design Outputs and Data Sharing: Institution specifies both operational outputs (eg: Birth certificates) and statistical outputs (eg, Annual Birth rate), It also establishes procedures for secure data sharing with authorized users while ensuring confidentiality, aligns outputs with national and international standards, and documents metadata detailing their purpose, scope, and dissemination format.</p> <p>2. Design Variables: Institution defines the specific variables to be collected, within the administrative data system, ensuring alignment with national and international standards and classifications, such as ISIC for economic activity, ISCO for occupation, and ISCED for education. This step also involves designing or selecting unique and stable identifiers to enable linkage across different administrative systems and documenting comprehensive metadata for both collected and derived variables, including definitions, formats, coding schemes, and reference periods.</p> <p>3. Design Collection: Institution establishes the mechanisms for capturing data, such as electronic forms, system-to-system transfers, APIs, or machine-readable templates, while embedding validation rules at the point of entry to reduce errors. The design also incorporates metadata collection, including timestamps, source system details, and reporting unit characteristics, and defines protocols for periodic data sharing with authorized institutions.</p> <p>4. Design Processing and Analysis: Institutions establish rules and procedures for validating, editing, and correcting data, including range checks, logical consistency checks, and duplication checks, as well as methods for handling missing or late data through imputation or estimation. The design also covers processes for integrating data with other administrative or statistical sources, applying statistical disclosure control to protect confidentiality, and developing routines for producing consistent and reproducible statistical aggregates and indicators.</p> <p>5. Design production systems and workflows: Institutions map the end-to-end workflow from data collection to dissemination, focusing on efficiency, and automation such as automated validation, system-generated reports, and API-based data sharing where possible. The design ensures interoperable databases and IT systems that follow open standards and reusability principles, clearly defines user roles and access rights, and embeds security and data protection standards.</p>

Core Phases	Processes
<p>Build phase</p> <p>The Build phase focuses on implementing the design specifications into functioning systems, tools, and workflows for administrative data collection, processing, storage, and dissemination. The goal is to ensure that the system meets both operational needs and the quality requirements for statistical use.</p>	<p>1. Build collection instruments incorporating validation rules and metadata</p> <p>2. Build processing and analysis components with quality checks (range, consistency, duplication, completeness) and routines for data integration where applicable</p> <p>3. Build dissemination components with modules that allow secure sharing and support multiple dissemination formats (dashboards, APIs, open data portals, reports).</p> <p>4. Configure workflows ensuring smooth integration across the system with mechanisms for receiving and validating data from reporting entities, storing both raw and processed data, running automated quality checks, generating operational as well as statistical outputs and ensuring that clear roles and responsibilities are assigned to staff.</p> <p>5. Test Production Systems verifying the functionality and integration of all components, interactions of different modules, data transfer processes, formats, metadata, and integrity, as well as security and access controls. Findings are documented, and necessary corrections are made before full implementation.</p> <p>6. Test the Administrative Business Process by simulating real-world data collection, processing, and dissemination. Assesses whether outputs meet both operational and statistical requirements, while evaluating data timeliness, completeness, accuracy, and comparability with other sources. Based on pilot results, refine the system, with iterative testing carried out until the process is robust and reliable.</p> <p>7. Finalise Production Systems by documenting technical manuals, process descriptions, and quality assurance guidelines, and by providing training for staff and reporting stakeholders. Deploy the system into the live environment, supported by monitoring mechanisms to track performance and data quality. The key quality dimension for this stage is Accuracy & Validity, and accessibility.</p>
<p>Collect phase</p> <p>The Collect phase involves systematically acquiring, receiving, and storing administrative records to ensure their quality, consistency, security, and readiness for future statistical use.</p>	<p>1. Run Administrative Data Collection by capturing administrative events in real time and ensuring accurate recording. Data are transmitted on agreed schedules and formats, and regularly monitored for completeness, structure, and correctness. Ensure that initial checks confirm files contain expected variables and formats. A feedback loop should be established to inform providers of quality issues, promoting improvements at the source.</p> <p>2. Finalise Administrative Data Collection for a given period by integrating administrative records into a secure, structured repository while maintaining version control and achieving raw data for audit purposes. Pseudonymized personal identifiers were required, and document metadata and paradata,</p>

Since administrative data already exist for operational purposes, this phase focuses on establishing structured and sustainable systems for its capture and management.

including information on data providers, timing, systems used, and response completeness. Document system versions, APIs, forms, and hand over the structured dataset along with accompanying metadata to the processing phase for cleaning, integration, and analysis.

Core Phases

Processes

Process phase

This phase describes how administrative data are processed and prepared for use in producing official statistics.

It involves sub-processes that integrate, classify, check, clean, and transform data received from administrative sources, so that they can be analysed and disseminated as statistical outputs.

1. Integrate data: Administrative data often come from multiple registers, institutions, or IT systems. This sub-process harmonizes and consolidates these diverse datasets into a single coherent and consistent dataset, resolving differences formats or coding. Integration may involve linking records across systems, combining geospatial or statistical information with register data, and de-identifying personal information to ensure confidentiality. Proper integration ensures that the combined dataset is complete, consistent, and ready for processing and analysis.

Key data quality dimension here is Coherence/Comparability, as the focus is on ensuring that data from different sources are consistent, standardized, and can be meaningfully combined.

2. Classify and code: Incoming administrative records are coded and classified to align with official statistical classifications (e.g., economic activity, occupation, geographic codes) if necessary. Coding can be automatic (using algorithms or AI) or clerical. This ensures comparability across sources and over time. The key quality dimension for this stage is Coherence/Comparability.

3. Review and validate: Data are checked to identify missing items, inconsistencies, duplicates, or values outside expected ranges. Validation rules may be applied iteratively and can include cross-checks with other registers. Administrative providers may also run validation checks before transmission, but statistical producers conduct additional scrutiny to ensure reliability. The key quality dimension for this stage is Accuracy & Reliability.

4. Edit and impute: Where administrative data are incomplete, outdated, inconsistent, or contain errors, editing and imputation procedures are applied. These may involve replacing missing values, resolving duplicates, correcting logical inconsistencies, or applying statistical models to estimate unknowns. All changes are documented through metadata.

5. Derive new variables and units: Some statistical needs require variables or units not directly available from administrative systems. These may be created by applying transformations, aggregations, or modelling. For example, deriving household units from population registers or enterprise units from legal or tax records.

Core Phases	Processes
	<p>6. Calculate aggregates: Aggregates are created by summing data for records sharing certain characteristics (e.g. aggregation of data by demographic or geographic classifications). Measures of totals, averages, and distributions are derived. In cases where administrative data are combined with sample survey data, standard error or confidence interval measures may also be estimated.</p> <p>7. Finalise data files: All processed data are consolidated into a final dataset (microdata or macrodata), which becomes the input for the Analyse phase. Depending on user needs, both provisional and final files may be produced, with metadata documenting the processing steps and quality considerations.</p>
Core Phases	Processes
<p>Analyse phase:</p> <p>In this phase, administrative data that have been processed are transformed into statistical outputs and examined in detail.</p> <p>It includes preparing statistical content (tables, indicators, commentary, and technical notes) and ensuring the outputs are “fit for purpose” prior to dissemination to users. Analysts also build an understanding of the data by comparing results with previous periods, with other sources, or with known benchmarks.</p> <p>For regular administrative statistics, this phase occurs in every cycle.</p>	<p>1. Prepare draft outputs: Processed administrative data are transformed into statistical outputs. These may include indicators (e.g., enrolment rates, tax revenues, health service coverage), indices, geo-referenced outputs, or microdata extracts. Draft tables, dashboards, and visualisations are prepared, and explanatory methodological notes are added. When new techniques such as data linkage or machine learning are used, transparent explanations are included.</p> <p>2. Validate Outputs: Administrative data outputs are assessed against established quality frameworks and expectations. Validation involves comparing current outputs with previous administrative cycles, checking consistency with metadata and administrative rules, benchmarking against survey data or other external sources, assessing geospatial consistency, and detecting unusual patterns that may result from system or policy changes.</p> <p>3. Interpret and explain outputs: Validated outputs are analysed to explain observed patterns and trends. Analysts interpret differences across time, regions, or groups, while considering legislative changes or reporting practices that may affect administrative data. Commentary and insight are added to help users understand results.</p> <p>4. Apply disclosure control: Institution ensures that the data (and metadata) to be disseminated, shared, or internally stored for future use do not breach the appropriate rules on confidentiality. Methods may include suppression, aggregation, perturbation, or advanced statistical disclosure control techniques. Special care is needed for geospatial or linked administrative datasets.</p>

Core Phases	Processes
<p>Dissemination phase</p> <p>This phase manages the release of statistical data and products derived from administrative sources to users.</p> <p>It includes activities to assemble and release products via multiple channels and to support users in accessing and using them.</p> <p>For regular administrative statistics (e.g., vital statistics, tax data, education records), this phase is repeated in each cycle.</p>	<p>1. Update Output Systems: Administrative statistics are loaded into dissemination platforms such as databases, data warehouses, or statistical portals, ensuring that both data and metadata are formatted according to dissemination standards (e.g., SDMX, DDI). This process includes linking data with metadata to provide transparency on definitions, sources, and methods, performing final consistency checks, and clearly indicating any changes in the underlying administrative system, such as new reporting obligations or changes in definitions that could affect interpretation.</p> <p>2. Produce Dissemination Products: Administrative data are transformed into outputs tailored to user needs, including statistical reports, thematic briefs, press releases, dashboards, interactive tools, data tables, open datasets, geo-enabled outputs, and public- or restricted-use microdata files with appropriate anonymisation. The administrative source context, such as legislation, registration practices, and reporting delays, is clearly documented to support correct interpretation. Confidentiality risks are reassessed, and additional disclosure controls are applied if needed before release.</p> <p>3. Release Dissemination Products: Administrative statistics are officially released in line with scheduled timelines and release calendars. This involves coordinating with ministries or partner institutions when statistics are co-produced, managing advance access under clear protocols, and providing outputs to authorized users, including access to anonymized microdata.</p> <p>4. Promote Dissemination Products: Institutions ensure that administrative statistics reach the widest relevant audience through channels such as institutional websites, national data portals, open data platforms, social media, newsletters, and stakeholder workshops. Coordination with partner institutions ensures consistent messaging. Promotion emphasizes both the statistical value and the administrative context to prevent misinterpretation of changes resulting from reforms or administrative rules.</p> <p>5. Provide User Support: Institutions respond to user queries and service requests, such as access to administrative microdata, within agreed service standards. This includes maintaining a helpdesk or knowledge base with FAQs, reviewing user requests to identify evolving data needs and inform quality management, supporting external researchers or partner institutions with data access protocols, and ensuring transparency through the publication of metadata, quality reports, and methodological notes. This process strengthens trust in administrative statistics by enhancing user engagement, transparency, and responsiveness.</p>

Core Phases	Processes
<p>Evaluate phase</p> <p>The evaluation phase for producers of administrative statistics focuses on assessing the quality, efficiency, and relevance of the administrative data production process.</p> <p>Unlike traditional survey-based processes, evaluation here emphasizes governance, interoperability, timeliness, user satisfaction, and the sustainability of the data supply chain.</p> <p>It ensures that administrative data meet national and international statistical standards and remain fit for purpose over time.</p>	<p>1. Gather Evaluation Inputs: Producers of administrative statistics collect materials to assess the quality and performance of their data systems throughout the production process. Key inputs include feedback from data suppliers on reporting burdens and system compatibility, feedback from data users on accessibility, timeliness, comparability, and usefulness, process metadata such as timeliness of delivery and frequency of corrections, quality indicators like accuracy, consistency, coherence, and compliance with international frameworks, system performance metrics including downtimes and interoperability, internal staff assessments of workflows and capacity, and results from external audits or peer reviews.</p> <p>2. Conduct Evaluation: This step involves analyzing collected inputs to assess the performance of the administrative data system against agreed targets and benchmarks. This includes comparing delivery timelines with Service Level Agreements, reviewing data consistency across time and sources, identifying gaps and inconsistencies, checking compliance with metadata standards, and considering user satisfaction and stakeholder feedback. Findings are synthesized into an evaluation report or dashboard that highlights strengths, weaknesses, risks, and opportunities for improvement.</p> <p>3. Agree on an Action Plan: Decision-makers establish a structured plan to improve the administrative data system based on evaluation findings. The plan includes corrective actions such as updating data transfer protocols or legal agreements, capacity-building measures like staff training and IT upgrades, monitoring mechanisms to track implementation, communication strategies to share findings with data suppliers and users, and decisions on whether any phase of the process should be repeated.</p>

Use of Standard Classifications and Coding Systems

Standard classifications provide a common language for data, ensuring comparability across institutions and alignment with international practices.

Requirements:

- All institutions shall use nationally adopted classifications and coding systems, as coordinated by NISR.
- Priority standards include:
 - Rwanda customized International Standard Classification of Education (ISCED 97) (<https://statistics.gov.rw/documents/ISCED>)
 - Rwanda Customized International Standard classification of Occupations (ISCO-08) (<https://statistics.gov.rw/documents/ISCO>)
 - Rwanda Customized International Standard Industrial Classification of all Economic Activities (ISIC, Rev.4) (<https://statistics.gov.rw/documents/ISIC>)
 - Rwanda Customized Classification of Individual Consumption according to Purpose (COICOP) (<https://statistics.gov.rw/documents/customized-classification-individual-consumption-according-purpose-coicop>)
 - Customized Central Product Classification (CPC) (<https://statistics.gov.rw/documents/customized-central-product-classification-cpc>)
 - ISO country, language, and currency codes (<https://simplelocalize.io/data/locale-code/rw-RW/>)
 - Harmonized national geographic codes for administrative areas available on the NISR website
 - Institutions may propose adaptations, but these must be reviewed and endorsed by NISR to maintain national harmonization.

Data Interoperability and Open Standards

Data interoperability ensures that government systems can exchange, integrate, and reuse data seamlessly, while open standards promote sustainability, efficiency, and innovation. Interoperability and openness are essential for building a cohesive national data ecosystem that reduces duplication, enhances trust, and maximizes the value of public sector data.

Requirements:

- All ICT systems handling government data shall comply with the National Interoperability Framework (ICT Sector Strategic Plan (2024 – 2029)). Interoperability requirements must be considered during system design, procurement, and upgrades.
- Institutions must use open, non-proprietary standards (e.g., SDMX, DDI, JSON, XML, RDF, and ISO 19115 for geospatial data). Proprietary or closed formats that limit data exchange should be avoided or phased out.
- Metadata and datasets must be API-ready for secure and efficient data exchange across institutions. APIs should be documented, version-controlled, and integrated into the National Data Catalogue
- All institutional data catalogues must be connected to the National Metadata Repository.
- Implement Master Data Management (MDM) principles for core entities (e.g., individuals, households, businesses, institutions, geographic areas).
- Apply unique and consistent identifiers and reference data (e.g., NID for persons, TIN for businesses, standardized geographic codes).
- Promote linked data practices to connect datasets across institutions.

6 Data Quality Roles and Institutionalization

To ensure high-quality, reliable, and trustworthy data, public institutions in Rwanda must embed data quality responsibilities into their existing data governance structure. Data quality is not an isolated task; it is part of the full chain of command from top-level strategic leadership to operational data entry.

The national data governance framework provides clear guidance on roles and responsibilities as a foundation for effective data quality and standards. This ensures ownership, accountability, and decision-making across all data functions. While institutional arrangements may differ depending on mandate, size, and capacity, GoR institutions should integrate these roles into their operational structures and document them accordingly. Functions may be combined if needed, provided they remain well-defined and effectively managed.

The table below presents a summary of the key roles and their responsibilities as recommended in the national framework.

Table 1: Roles and responsibilities in data quality and standards

Role	Key Responsibilities
Data team	<ul style="list-style-type: none">• Approves the institution's data quality strategy and ensures alignment with national standards set by NISR.• Sets data quality objectives, priorities, and performance targets in line with the institution's strategic goals.• Resolves cross-departmental issues that affect data quality.• Reviews institutional performance reports on data quality and mandates corrective action where necessary.
Chief Data Officer (CDO)	<ul style="list-style-type: none">• Provides overall leadership for data governance across the institution• Coordinates data quality activities across departments, ensuring consistent application of standards.
Data Owner(s)	<ul style="list-style-type: none">• Own the quality of specific datasets• Approve dataset-specific data quality rules and quality assessment criteria.• Ensure that completeness, timeliness, accuracy, and other quality dimensions are met before data is shared or published.• Provide direction to Data Stewards and collaborate with Data Custodians to ensure systems support data quality needs.• Approve data quality reports and sign off on the publication of official statistics or management reports.

Role	Key Responsibilities
Data Steward(s) Data Governance Officer	<ul style="list-style-type: none"> • Monitor day-to-day data quality, running checks for completeness, accuracy, uniqueness, and consistency. • Maintain metadata and ensure datasets comply with agreed definitions, formats, and standards. • Work closely with Data Custodians to configure system validation rules. • Engage with Data Producers to address recurring quality issues at the point of data entry. • Prepare regular data quality dashboards and reports for Data Owners and the CDO.
Data Custodian(s) / IT Admin	<ul style="list-style-type: none"> • Implement and maintain system-level validation checks to prevent invalid or duplicate records. • Ensure that databases and applications store data in approved formats and enforce the use of unique identifiers. • Maintain secure and reliable systems that support timely updates and data sharing. • Work with Data Stewards to ensure systems meet functional requirements for data quality.

7 Quality Assurance Rules and Data Standards

7.1 Core Data Quality Management Processes

Public institutions must define structured processes that underpin the application of data quality rules and standards.

Key Processes:



7.2 Data Collection and Entry Standards

High-quality data is realized at the point of capture. Public institutions must implement rigorous and standardized collection and entry practices to ensure reliability, consistency, and comparability.

1 Standardized Data Collection Instruments

- All questionnaires, administrative forms, and digital tools must be standardized, pre-tested, and validated before use.
- Data definitions, classifications, and codes must follow national metadata standards (NISR metadata guideline) and, where applicable, international classifications such as ISIC, ISCO, and ICD.
- Instruments should minimize ambiguity and use clear, simple language to avoid misinterpretation.

2 Enumerator Training and Guidelines

- Enumerators and data entry staff must receive structured training covering definitions, coding standards, use of digital devices, and confidentiality protocols.
- Detailed fieldwork manuals should be developed to ensure consistent practices across all teams.
- Training should include mock interviews and pilot tests to strengthen familiarity with tools.

3 Use of Technology in Data Collection

- Computer-Assisted Personal Interviewing (CAPI) or web/mobile-based systems should be used wherever possible to reduce manual errors.
- Built-in validation rules (range checks, logical checks, skip patterns) must be implemented at the point of data entry.
- GPS tagging, timestamps, and interviewer IDs should be captured to enhance data integrity and accountability.

4 Data Entry Protocols

- Where manual data entry is unavoidable, institutions must adopt double data entry or verification sampling methods to detect and correct errors.
- Data entry staff must be trained on the correct handling of missing values, codes, and special entries (e.g., “Don’t know,” “Refused”).
- Unique identifiers must be consistently applied to ensure proper linkage across datasets.

5 Standards for Handling Metadata and Coding

- All collected data must be linked to metadata describing concepts, definitions, classifications, and coding rules.
- Metadata must comply with international standards.
- Institutions should maintain version control of instruments and metadata to document any changes across time.

6 Confidentiality and Ethical Standards

- Respondent confidentiality must be safeguarded in line with Rwanda’s data protection law and international principles.
- Informed consent procedures must be clearly documented and explained to respondents before data collection.
- Sensitive information must be encrypted or anonymized during collection and entry.

7 Quality Control during Data Collection

- Supervisors must conduct spot checks, re-interviews, and consistency checks during fieldwork.
- Institutions should apply real-time monitoring dashboards (where digital tools are used) to detect and address errors early.
- Any identified errors should be documented, corrected, and reported systematically.

7.3 Data Processing, Cleaning, and Validation Procedures

Data must be processed, cleaned, and validated systematically to ensure it is trustworthy, reliable, and ready for use.

1. Data Transformation

- Institutions should define and document all transformation rules, including coding, recording, aggregation, and creation of derived variables.
- Transformation processes must be consistent across datasets and follow agreed metadata standards
- Common transformations (e.g., conversion of text into codes, standardization of names and addresses, harmonization of dates and times) should be automated where possible to reduce human error.

2. Validation Checks

- Automated Checks: systems must run validation rules at the point of data entry to minimize errors (e.g., logical checks, numeric ranges, mandatory field completion).
- Manual Checks: The institution shall conduct secondary checks on suspicious records flagged by the system.

3. Consistency and Integrity Rules

- Data across related fields must be logically consistent (e.g., age vs. education level, gender vs. occupation, household composition vs. marital status).
- Cross-dataset validation should be applied to ensure comparability with administrative registers and previous survey rounds.
- Referential integrity (e.g., unique IDs in registries, linkages between households and individuals) must be enforced.

4. Handling Missing, Duplicate, and Outlier Data

- Missing Data: Institutions must document imputation rules (mean substitution, hot-deck imputation, model-based imputation) and justify their choice.
- Duplicates: Duplicate records must be identified and eliminated using unique identifiers or probabilistic matching methods.
- Outliers: Statistical methods must be applied to detect and treat outliers. Any corrections must be documented.

5. Data Lineage and Audit Trails

- All processing activities (editing, transformation, imputation, deletion) must be logged and stored in audit trails.
- Data lineage documentation must describe how raw data was transformed into final datasets.
- Audit logs should be accessible to authorized personnel for verification and accountability.

6. Documentation and Transparency

- Metadata must include details of data cleaning and validation rules applied.

- All imputation and editing methods must be documented in survey or administrative system reports.
- End-users must be informed about the level of adjustments applied to the dataset.

7.4 Mandatory Fields and Data Standards

To ensure accuracy, consistency, and comparability, all Government of Rwanda (GoR) institutions that collect or manage information on individuals shall capture the following mandatory data fields, in accordance with the prescribed standards, validation rules, and consistency checks. Additional or domain-specific data standards not explicitly covered in these guidelines are available on the National Institute of Statistics of Rwanda (NISR) website.

1. National ID (NIN)

- **Description :** Unique national identification.
- **Type :** Numbers.
- **Standard:** 16-digit numeric.
- **Universe:** All individuals.
- **Question should be asked:** Enter National ID number or what is your national ID
- **Validation Rules:** Must be validated in real time against the NIDA registry to prevent duplication.
- **Notes:** ID serves as a unique identifier for individuals. Systems should auto-fill other fields (e.g., Names, Sex, Date of Birth).

2. Full Legal Name

- **Description:** All names of the person in the order:
- **Type:** String.
- **Standards:** Start with Surname/family name, then followed by Other Names.
- **Universe:** All individuals.
- **Question should be asked:** What is your surname? What are your other names?
- **Validation Rules:** It must match official NIDA records. Names must not contain numbers or symbols. The order Surname + post-name must strictly follow the legal format (Law N° 32/2016).
- **Notes:** Names should be in proper case (e.g., Kamana (*Surname*) Owen (*Other names*) or Akaliza (*Surname*) Mwiza Diane (*Other names*). Auto-fill from NIDA is recommended.

3. Sex

- **Description:** Sex of the person, and it is the one recorded in his or her birth record.
- **Type:** Categorical.
- **Standard options:** Male and Female
- **Universe:** All individuals.
- **Question should be asked:** What is your sex?
- **Valid Codes:** Always in Capital
 - 1 = Male
 - 2 = Female.

Validation Rules: It must strictly follow the legal options (Law n° 71/2024 of 26/06/2024 governing persons and family)

4. Date of Birth

- **Description:** Date of birth of the individual.
- **Type:** date
- **Standard of date:** YYYY-MM-DD format, start with Years, Months, and then Days
- **Universe:** All individuals.
- **The question should be asked:** What is your date of birth?" or "Enter your date of birth in YYYY-MM-DD format.
- **Valid Range:** 1900-01-01 to current date.
- **Notes:** Age is automatically calculated in the system.

5. Marital Status

- **Description:** Current marital status of the person.
- **Type:** Categorical.
- **Standard options:** Single, living together, married, separated, divorced, and widowed
- **Universe:** Individuals.
- **The question should be asked:** What is your present marital status?
- **Valid Codes:**
 - 1 = Single (not living in a union)
 - 2 = Living together
 - 3 = Married
 - 4 = Separated
 - 5 = Divorced
 - 6 = Widowed
- **Consistency Checks:**
 - A spouse cannot be single, divorced, separated, or widowed.
 - Household head and spouse must have the same marital status.
 - Individuals who are married, divorced, or widowed must be at least 18 years old.

6. Highest Level of Education Attained

- **Description:** Highest level of education completed.
- **Type:** Categorical.
- **Universe:** All individuals.
- **Question should be asked:** What is the highest level of education you have attended or are currently attending?
- **Valid Codes:**
 - 1 = None
 - 2 = Pre-primary
 - 3 = Primary
 - 4 = Lower secondary
 - 5 = Upper secondary
 - 6 = University
- **Validation:** Must match Ministry of Education / UNESCO-approved codes.

7. Area of Residence

- **Description:** Administrative location where the person resides.
- **Type:** String / Code.
- **Universe:** All individuals.
- **The question should be asked:** Enter Province, District, Sector, Cell, and Village.
- **Validation:** Must follow official location codes.
- **Notes:** Required for geographic analysis and service delivery.

8. Primary Contact Number

- **Description:** The Main phone number of the person.
- **Type:** String.
- **Standard:** +250 plus 9 digits
- **Universe:** All individuals.
- **The question should be asked:** Enter primary contact number.
- **Valid Format:** +250XXXXXXXXX (12 digits).
- **Validation:** Must comply with a valid Rwandan mobile format. The +250 should be auto-filled, then enter the remaining numbers.

Implementation Notes:

- Systems should automate validation wherever possible (NIN, age checks, marital status rules).
- Mandatory fields must be enforced before saving records.
- Use dropdowns or coded values for standardized fields (e.g., Sex, Marital Status, Education, etc.)
- Ensure all fields comply with the National Metadata Management Guideline and support interoperability.

Data quality table

Dimension	Applicable Rules
1. Completeness	<ol style="list-style-type: none"> 1. All required fields must be filled before saving. 2. All expected records for the reporting period must be submitted. 3. Avoid placeholder values like "N/A" or "0000000000". 4. Run monthly missing data checks. 5. Supervisors must monitor completeness during data entry.
2. Uniqueness	<ol style="list-style-type: none"> 1. Assign a unique ID for each record. 2. Check for duplicates before saving. 3. Merge/delete duplicates quarterly. 4. Cross-check IDs with official systems. 5. Use standardized naming to avoid accidental duplicates.
3. Consistency	<ol style="list-style-type: none"> 1. Related fields must match logically. 2. Use the same core data across linked systems. 3. Apply official code lists. 4. Reconcile datasets regularly. 5. Enforce standard formats.
4. Timeliness	<ol style="list-style-type: none"> 1. Define update deadlines for all datasets. 2. Record changes within set days (e.g., 5 working days). 3. Track timely submissions. 4. Send reminders before deadlines. 5. Publish timeliness performance reports.
5. Validity	<ol style="list-style-type: none"> 1. Enforce correct formats (e.g., YYYY-MM-DD for dates). 2. Check numeric ranges. 3. Use drop-down lists. 4. Reject invalid values automatically. 5. Validate against reference lists.
6. Accuracy	<ol style="list-style-type: none"> 1. Match data to source documents. 2. Randomly verify samples monthly. 3. Provide a way to report errors. 4. Use double-entry for critical data. 5. Share error findings with data collectors.
7. Integrity	<ol style="list-style-type: none"> 1. Maintain correct links between related records. 2. Prevent "orphan" records. 3. Check for broken links. 4. Restrict deletion of linked records. 5. Use centralized master lists.
8. Relevance	<ol style="list-style-type: none"> 1. Collect only needed data. 2. Review datasets yearly. 3. Consult users before adding fields. 4. Avoid redundant data collection. 5. Align with policy needs.
9. Accessibility	<ol style="list-style-type: none"> 1. Use role-based access control. 2. Store in usable formats (CSV, Excel). 3. Follow official data-sharing protocols. 4. Publish non-confidential data openly. 5. Include metadata and definitions.



Definitions and Key Concepts

1. **Data:** Structured or unstructured information collected, stored, processed, and used to support decision-making, reporting, or service delivery.
2. **Data Quality:** The degree to which data is fit for its intended purpose, meeting dimensions such as accuracy, completeness, consistency, timeliness, relevance, and accessibility.
3. **Data Lifecycle:** The complete sequence of stages that data passes through, from specification and collection to processing, analysis, dissemination, archiving, and disposal.
4. **Data Governance:** The framework of policies, standards, processes, and roles that ensure accountability, consistency, and security in the management of data across an organization or system.
5. **Data Stewardship:** The practice of managing and overseeing data assets, ensuring compliance with standards, protecting data integrity, and facilitating proper use.
6. **Metadata:** Structured information describing the content, context, structure, quality, and provenance of data, enabling users to understand, interpret, and use it appropriately.
7. **Data Integrity:** The accuracy, completeness, and consistency of data over its lifecycle, ensuring that information remains trustworthy and unaltered except through authorized processes.
8. **Data Interoperability:** The ability of data systems and institutions to exchange, access, and use data seamlessly, often through adherence to common standards and classifications.
9. **Administrative Data:** Information collected primarily for administrative purposes (e.g., records, transactions, registries) that can be leveraged for statistical or policy analysis.
10. **Statistical Data:** Data collected or compiled according to recognized statistical methodologies for producing official statistics, including censuses, surveys, and derived indicators.
11. **Data Custodian:** An individual or unit responsible for the technical management, storage, security, and maintenance of data within an organization.
12. **Data Owner:** A senior official accountable for the quality, use, and governance of a particular data-set or data domain.
13. **Data Producer:** An individual or institution that collects, generates, or compiles data, ensuring it meets prescribed standards and quality requirements.
14. **Data User:** Any individual or entity that accesses, analyzes, or applies data to inform decisions, policies, or research.
15. **Data Validation:** The process of checking data for accuracy, completeness, consistency, and conformance with predefined rules or standards.
16. **Fit-for-Purpose:** The extent to which data meets the needs of users for a specific application, policy, or decision-making context.
17. **Sensitive Data:** Data whose disclosure, misuse, or unauthorized access could compromise privacy, security, or institutional integrity, including personal, financial, or confidential records.
18. **Data Standard:** A formally agreed rule, specification, or methodology that ensures consistency, comparability, and interoperability of data across systems and institutions.
19. **Data Quality Dimensions:** Attributes used to measure and evaluate data quality, including but not limited to accuracy, completeness, consistency, timeliness, validity, relevance, accessibility, interpretability, coherence, and integrity.
20. **Data Ecosystem:** The interconnected network of institutions, systems, standards, and practices involved in collecting, managing, sharing, and using data across a country or sector.



Example of Data Quality KPIs by Dimension

Dimension	KPI Name	Definition / Purpose	Formula / Measurement	Interpretation	Frequency
Completeness	Data Completeness Rate	Measures how much required data is available	$(\text{Filled mandatory fields} \div \text{Total mandatory fields}) \times 100$	Below 90% indicates missing or uncollected information, often due to poor data entry controls	Monthly
	Record-Level Completeness	Tracks how many records are fully complete	$(\# \text{ of complete records} \div \text{Total records}) \times 100$	Helps identify datasets requiring additional validation or imputation	Quarterly
Uniqueness	Duplicate Record Rate	Evaluates redundancy in records	$(\text{Duplicate entries} \div \text{Total records}) \times 100$	Duplicates distort aggregates, analytics, and reporting	Quarterly
	Entity Resolution Rate	Tracks matched entities across systems	$(\text{Resolved duplicates} \div \text{Total duplicates identified}) \times 100$	Shows progress in master data management (MDM) or entity resolution projects	Quarterly
Consistency	Cross-System Consistency Index	Measures alignment between key data elements across systems	$(\text{Consistent values} \div \text{Total cross-system comparisons}) \times 100$	High inconsistency implies a lack of interoperability or version control	Quarterly
	Referential Consistency	Measures harmony between datasets (e.g., codes, classifications)	$(\text{Aligned code values} \div \text{Total code values}) \times 100$	Detects use of outdated code lists or inconsistent standards	Quarterly
Timeliness	Data Submission Timeliness	Measures the time lag between the event and data availability	Average (days between event and record update)	Long lag impacts decision timeliness	Monthly
	Update Frequency Compliance	Measures adherence to the update schedule	$(\text{Datasets updated as per SLA} \div \text{Total datasets}) \times 100$	Indicates process discipline	Monthly

Dimension	KPI Name	Definition / Purpose	Formula / Measurement	Interpretation	Frequency
Validity	Data Rule Conformance	% of records adhering to business or validation rules	$(\text{Records passing validation} \div \text{Total records}) \times 100$	Captures structural quality; low rates signal rule enforcement gaps	Monthly
	Outlier Detection Rate	% of numeric records within acceptable range	$(\text{Records within range} \div \text{Total records}) \times 100$	Detects potential entry or process anomalies	Quarterly
Accuracy	Data Accuracy Rate	% of records matching trusted source or ground truth	$(\text{Correct records} \div \text{Total checked records}) \times 100$	Indicates the precision of captured values	Quarterly
	Verification Audit Score	% of random samples validated without correction	$(\text{Valid samples} \div \text{Samples audited}) \times 100$	Supports assurance audits	Annual
Integrity	Referential Integrity Score	Measures the correctness of linked records	$(\text{Valid foreign key links} \div \text{Total foreign key links}) \times 100$	Critical for relational databases	Quarterly
	Data Relationship Completeness	Tracks missing relationships between entities	$(\text{Valid relationships} \div \text{Expected relationships}) \times 100$	Reveals structural issues	Quarterly
Relevance	Data Usage Rate	% of datasets used in reports, dashboards, or policies	$(\text{Datasets used} \div \text{Total datasets available}) \times 100$	High use = relevance; low use = redundancy	Bi-annual
	User Satisfaction Index	% of users rating data as useful/reliable	$(\text{Positive feedback} \div \text{Total responses}) \times 100$	Obtained via surveys or feedback portals	Annual
Accessibility	Data Availability	% of datasets accessible via authorized platforms (catalog/API)	$(\text{Accessible datasets} \div \text{Total datasets}) \times 100$	Indicates how well data infrastructure supports openness	Quarterly
	Access Request Fulfillment Rate	% of approved requests fulfilled within SLA	$(\text{Requests fulfilled on time} \div \text{Total requests}) \times 100$	Reflects efficiency and transparency	Quarterly

Dimension	KPI Name	Definition / Purpose	Formula / Measurement	Interpretation	Frequency
Validity	Data Rule Conformance	% of records adhering to business or validation rules	$(\text{Records passing validation} \div \text{Total records}) \times 100$	Captures structural quality; low rates signal rule enforcement gaps	Monthly
	Outlier Detection Rate	% of numeric records within acceptable range	$(\text{Records within range} \div \text{Total records}) \times 100$	Detects potential entry or process anomalies	Quarterly
Accuracy	Data Accuracy Rate	% of records matching trusted source or ground truth	$(\text{Correct records} \div \text{Total checked records}) \times 100$	Indicates the precision of captured values	Quarterly
	Verification Audit Score	% of random samples validated without correction	$(\text{Valid samples} \div \text{Samples audited}) \times 100$	Supports assurance audits	Annual
Integrity	Referential Integrity Score	Measures the correctness of linked records	$(\text{Valid foreign key links} \div \text{Total foreign key links}) \times 100$	Critical for relational databases	Quarterly
	Data Relationship Completeness	Tracks missing relationships between entities	$(\text{Valid relationships} \div \text{Expected relationships}) \times 100$	Reveals structural issues	Quarterly
Relevance	Data Usage Rate	% of datasets used in reports, dashboards, or policies	$(\text{Datasets used} \div \text{Total datasets available}) \times 100$	High use = relevance; low use = redundancy	Bi-annual
	User Satisfaction Index	% of users rating data as useful/reliable	$(\text{Positive feedback} \div \text{Total responses}) \times 100$	Obtained via surveys or feedback portals	Annual
Accessibility	Data Availability	% of datasets accessible via authorized platforms (catalog/API)	$(\text{Accessible datasets} \div \text{Total datasets}) \times 100$	Indicates how well data infrastructure supports openness	Quarterly
	Access Request Fulfillment Rate	% of approved requests fulfilled within SLA	$(\text{Requests fulfilled on time} \div \text{Total requests}) \times 100$	Reflects efficiency and transparency	Quarterly

