# Guideline on Assessing Quality of Administrative Data for Producing Official Statistics

*Version 1*

NATIONAL STATISTICS BUREAU
SEPTEMBER 2022

United Nations | Department of Economic and Social Affairs

# Guideline on Assessing Quality of Administrative Data for Producing Official Statistics

*Version 1*

**NATIONAL STATISTICS BUREAU**
SEPTEMBER 2022

United Nations | Department of Economic and Social Affairs

Layout & Design National Statistics Bureau

# Contents

# Introduction[1]

The use of administrative data as an additional source to produce official statistics has been increasing in recent years. Administrative data provides valuable information on the Bhutanese society and can replace expensive surveys as well as provide additional information on areas where surveys are not being conducted.

It is important to remember that administrative data is collected for administrative purposes of the public sector, not necessarily to produce official statistics. However, administrative data is a great data source for producing official statistics and should be exploited as much as possible. As with other data sources being used to produce official statistics such as surveys, we need to consider the quality of the administrative data and how well the data fits the purpose we want to use it for. Essentially, we want to know how suitable the administrative data is for producing official statistics and whether it is fit for purpose. This is the key in assessing the quality of administrative data for producing official statistics.

In 2021/2022, the National Statistics Bureau (NSB) was tasked by the Government Performance Management Division (GPMD), Office of the Prime Minister (OPM) to assess the quality of official statistics produced by using administrative data as a mandatory indicator for all government agencies in the Annual Performance Agreement (APA) 2021-2022. This report and the accompanying mapping tool and questionnaire compose the "Guide of assessing the quality of administrative data for producing official statistics".

### Quality principles in official statistics

The international statistics community has agreed on a set of quality principles, which are relevant for producing official statistics. These principles are equally important when we produce official statistics based on administrative data sources. The quality principles, or dimensions, can be seen in the table below:

| **Quality dimensions for producing official statistics[2]** | |
|---|---|
| **Relevance** | Extent to which the data satisfy the needs of users |
| **Accuracy** | Closeness of the data to the exact value that the statistics were intended to measure i.e., how complete is the data knowing that statistics always come with a level of uncertainty. |

*1 This guide and the accompanying tools are developed with inspiration from international experience as well as quality guides and assessment tools from other countries. The list of resources used can be found in annex 1.*

*2 UN DESA: "United Nations National Quality Assurance Frameworks Manual for Official Statistics"*

| | |
|---|---|
| **Timeliness** | Data should be up-to-date, and thus be as relevant as possible to the user i.e., the length of time between the date of the event / the reference period and the dissemination of the statistics. |
| **Punctuality** | Statistics should be released on time and made available to users when expected i.e., the time lag between the actual release date and the planned date by which the statistics should have been delivered |
| **Accessibility** | Ease and conditions with which the data can be obtained by those producing statistics as well as its users |
| **Clarity** | Availability of appropriate documentation relating to the data (also known as metadata) |
| **Consistency/coherence** | Ability to combine data sets in different ways and for various uses |
| **Comparability** | Extent to which the data can be compared over geographical areas (e.g., Dzongkhags), over time (e.g., a period of 5 years) or across countries |

In addition to above, completeness is important when we talk about administrative data sources. Completeness is the extent to which the data source includes all the essential values needed to get the full picture.

To learn more about how NSB is working with ensuring quality of official statistics, please refer to the "Bhutan Statistics Quality Assurance Framework"[3].

**Challenges with using administrative data for producing official statistics**

Many countries around the world are already using and further exploring administrative data sources. Many lessons learned have been made and some of the typical challenges with using administrative data sources for producing official statistics have been identified. Some of the challenges include:

- Definitions, concepts and classifications can differ from those needed for statistical purposes as the administrative data source is made for operational purposes, not for statistical purposes.
- Administrative data is in theory a complete set of data on those population units of the study area concerns. However, there can be coverage issues with both under- and over-reporting certain population units or even relevant sub-population groups.
- Even if all relevant population units are included in the administrative data source, some cases or variables can be missing. The data is thus incomplete.
- There is often little information (metadata) available on the administrative data source which makes it difficult to understand the data and thus complicates the statistical production.

3 https://nsb.gov.bt/wp-content/uploads/dlm_uploads/2020/12/BSQAF-2020-v1_web.pdf

- Population units are not always linked to a unique identifier, or the data includes only aggregate data, not individual data. This can cause difficulties in linking data with other relevant sources.
- Since the administrative data is owned by a public institution, they can change the administrative data system without thinking about the consequences for producing official statistics.

**Using unique ID throughout the entire administrative data system**

Establishing a unique ID on respectively persons, businesses and housing units is a crucial step in ensuring interoperability across the public sector. The unique IDs on persons, businesses and housing units should be used throughout the entire administrative data system and will create vast opportunities for producing official statistics as data can be linked across all sectors.

Since all citizens in Bhutan already has a CID, this unique ID on persons should be used by all public agencies when collecting administrative data on citizens. Bhutan needs to explore the opportunity for establishing unique IDs for businesses and housing units as well as issuing CID to children below 15 years, which likewise should be used by all public agencies, thus creating the foundation for a strong administrative data system, which can be used to produce high quality official statistics.

**How to assess the quality of administrative data**

***A. Build an overview of available administrative data sources***

The first step in assessing the quality of administrative data is to identify which data sources exists. Some administrative data sources are already being used for producing official statistics, while other administrative data sources could be a potential source for producing official statistics – either replacing partly or entire surveys or producing new information on society which is not yet published.

A mapping tool[4] can be used to identify available sources. Such a tool provides a template for detailing information about the source such as information about the data holder, which variables the data source contains, the target population, process of data collection, data validation procedures, which statistics it can produce etc. The mapping can be done sector by sector and will thus give an overview of how different administrative data sources within the same sector can possibly be linked. This overview is useful as the different public institutions within the same sector might not themselves have this overview.

Once the mapping is done within a sector, it is possible to identify those administrative data

4 *The Mapping tool to identify available admin data within a sector is available at www.nsb.gov.bt*

already being used for official statistics, and which administrative data sources could be further investigated for their potential.

### *B. Assess the quality of administrative data for producing official statistics*

A questionnaire[5] has been developed to assess the quality of administrative data being used to produce official statistics. The results of the questionnaire will help both producers and users of the statistics to better understand its quality and whether it is fit for purpose. The questionnaire will also be used to score the quality of official statistics.

The first time an administrative data source is assessed, it is recommended that the assessment is done in collaboration between NSB and the administrative data holder to establish a common understanding of the quality dimensions and the quality of the official statistics. Thereafter, the assessment will be conducted once a year on those statistics of highest priority and importance. Assessment on other official statistics should also be conducted regularly but with a shorter frequency. The frequency will be agreed with NSB. The subsequent assessments will be conducted more like a self-assessment where the data holder updates the questionnaire based on the current situation. Once the self-assessment is complete, the data holder sends the questionnaire to NSB, and a meeting will be conducted to discuss the results. A short work plan should be produced after each assessment with point of actions on how to improve identified quality issues.

Official statistics based on administrative data which is produced by NSB will likewise be assessed with NSB conducting a self-assessment.

The four sections below describe the different phases in assessing the quality of administrative data and are comparable with the four sections in the “Questionnaire for assessing the quality of administrative data for producing official statistics”.

5 *The Questionnaire for assessing quality of administrative data is available at www.nsb.gov.bt*

# 1. Institutional arrangements

Institutional arrangements concern the structures around the administrative data.

### 1.1 Mandate

A mandate to collect the administrative data and to produce statistics for official use will contribute to an effective and well-functioning national statistics system.

A national statistical law is specific to the production of official statistics and gives a clear mandate to the national statistics system and in particular the national statistics office to collect data and to compile and disseminate statistics. In Bhutan, there is no statistical law as such, however, the central authority to produce official statistics is mandated to NSB in the "Executive order" from 2006[6].

Besides a national statistical law[7], legal provisions or other types of mandates of the administrative data holder to collect administrative data on citizens and businesses is equally important.

### 1.2 Confidentiality and security

Confidentiality concerns the protection of data and in particular data at individual level, which can be tracked to specific persons and organisations. Protection of administrative information is engrained in the 2022 law "Rules for Administrative Disciplinary Actions" issued by the Royal Civil Service Commission. It is therefore important to know whether the administrative data holder has systems in place which keep the individual data confidential so information about persons or organisations is not misused or disclosed without authorisation. Mechanism to protect confidentiality could include staff awareness of the importance of confidentiality, anonymisation procedure and log system to track who access data.

Security concerns how data is stored and protected e.g., from hackers or disasters that could either delete data entirely or prevent temporary access.

Storing data on a centralised server helps protecting data by avoiding random storage of confidential information on e.g., one computer's hard disk. A centralized server gives the possibility to decide on security measures which can be centrally managed.

Effective data transmission and storage security provisions seek to ensure that data is stored

---

6 *The "Executive Order" from 2006 is available on the website of NSB: https://nsb.gov.bt/wp-content/uploads/dlm_uploads/2021/01/Executive_Order2006.pdf*

7 *If you want to learn more about the importance of a legal framework, you can read the United Nations Statistics Wiki on the topic. The importance of the legal framework - MSITS 2010 Compilers Guide - UN Statistics Wiki*

and transmitted in a secure way, so data is not leaked nor lost. Further, putting in place proper back-up and disaster recovery procedures will reduce the risk of data being jeopardized by hackers and disasters.

Protecting confidentiality and ensuring security help increase trust and goodwill from the persons and organisations whose data we are dependent on.

### 1.3 Staffing at agency level

A key quality principle is the independence of the official statistics production from political interference. One way of checking this is to ask questions on the independence of the unit at the administrative data holder and whether they can work independently from the often more political institution, which they are part of. Furthermore, having sufficient and skilled staff to produce official statistics is important to ensure the statistical products can be checked for its completeness, validity etc. in due time.

### 1.4 Data Exchange Agreements

Administrative data can be shared within the Bhutan Statistical System with the purpose of linking different data sources to produce official statistics. Administrative data is also shared with NSB either because NSB is responsible for producing and publishing the relevant statistics or because NSB stores the data in the Bhutan Statistical Database System (BSDS) for statistical purposes. In those cases, it is important that a Data Exchange Agreement is signed between the administrative data holder and the agency receiving the data.

A Data Exchange Agreement is a way of formalising the details of collaboration and exchange between the administrative data holder and the data receiver. Next to a general commitment to collaboration, the Data Exchange Agreement should contain technical details about the data exchange. The technical details include issues such as:

- Type of data to transfer (e.g., information on data fields, metadata structure, variables)
- Format of delivery (e.g., API, CSV etc.)
- Means of transmission (e.g., email, secured transmission system etc.)
- Frequency of data transfer (monthly, quarterly, annually etc.)

If the data transfer concerns microdata (data at individual level), the Data Exchange Agreement can also include information on how data is to be kept confidential and secure at the agency receiving the data, so the administrative data holder is confident with transferring their data to another authority.

### 1.5 Technical Checks

Ensuring the correct transmission of data will reduce the risk of errors and time spent on understanding the data. Transmission of data should comply with the agreed format as describe in the Data Exchange Agreement. Furthermore, metadata (i.e., data that defines or describes the

administrative data such as concepts, definitions, and classifications) should be included with the transfer of data.

# 2. Quality of input data

This section refers to the phase of data entry from the respondent to when data is received by the administrative data holder i.e., the data collection itself. The data might go through several hands before it reaches the administrative data holder. Data collection is a crucial step in ensuring high quality of administrative data.

### 2.1 Accuracy

Data accuracy refers to how well the data matches reality and how well the collected data capture what we are trying to measure. The methods used for collecting data can influence the quality.

Errors in data can occur already from the very start if the reporting form/system, which the data holder has designed, is e.g., unclear, give cause to confusion, complex in its set-up, too lengthy or has too much overlapping data requests with other reporting forms.

Data errors can occur when the respondent enters data into the reporting form/system e.g. poor internet facilities, lack of capacity or appreciation of the importance of data reporting, and difficulty in capturing all the data needed for reporting.

Data errors can occur when data is transmitted and validated by the districts level e.g. errors resulting from accidentally deleting some data point. Finally, errors can occur once data is received by the administrative data holder and registered and handled in its repository.

Once possible errors in the data collection have been identified, it will be useful to set up measures to mitigate the identified risks to increase the quality of data entry in the future.

### 2.2 Validity

Validity refers to the extent to which data conforms to the expected format, type, and range e.g. dates are written in a certain format. When data is valid it is easier to link it with other datasets and to run automated processes.

From a data collection point of view, the more clarity the respondent has on the expected format, type and range, the more correct data s/he enters into the system. Therefore, it is useful to offer the respondent information and documentation on what standard the data should be reported. The reporting forms might also be developed with automatic data

validation features which could help the respondent to report data in the expected format, type, and range e.g., the reporting form only allows dates to be entered a certain way.

### 2.3 Unique ID

Unique ID refers to whether the data can be tracked to an individual population unit. To identify a unique population unit, the data system needs to work with unique identifiers. The CID is an example of a unique identifier for citizens in Bhutan, which some administrative data systems are using. Using a nationally generated unique identifier such as the CID, provides opportunities to link data from different data sources and to conduct more advanced analysis.

### 2.4 Timeliness

Timeliness refers to how well the data reflects the period they are supposed to represent and how up to date the data and its values are.

Data entry can influence the timeliness of data e.g., the time lag between the time of the event and when it is recorded in the system. Sometimes events are registered weeks or months later than they occurred for reasons that can either be institutional or legal (e.g., the law allows families to report births and deaths up to one year after the event occurred).

Data might also be collected with a different frequency than originally planned which might influence the quality of the data as data might be delayed and have spill-over effect on the timely production of statistics.

### 2.5 Respondent burden

Reporting data is a burden on respondents as they spend time on collecting and reporting data. It is therefore crucial to identify whether systematic efforts to reduce the burden on respondents are put in place. Efforts to reduce the burden could include reducing the volume of data collected to the minimum and streamlining different reporting forms so they do not ask for the same type of data more than once.

# 3. Quality of data processing and analysis

This section refers to the phase of processing the data once it has been collected. Data processing is an important step in assessing quality as during this phase errors to the data is detected and if possible corrected, and metadata is clarified.

### 3.1 Accuracy and completeness

Completeness refers to the degree to which all essential values are present. The more complete the data is, the higher the quality. Completeness refers to

- Under-coverage i.e., whether objects are missing in the data source, which should be present.
- Over-coverage i.e., whether the data source contains objects, which should not be present
- Selectivity i.e., whether the data source only contains information on selected part of the target population, which could lead to bias as whole sub-population groups might be missing
- Redundancy i.e., whether the data source includes duplications of the exact same objects

Conducting analysis on the completeness of the data is an essential way to establish an understanding of the quality, especially for users since they need to know how accurate the official statistics is. Trying to identify and understand the reasons for incomplete data, e.g., are missing data random or is there a system in certain type of missing data, will provide an opportunity to correct its causes in the future. However, a 100 % complete dataset is unrealistic and will also require enormous amounts of resources.

You also need to check your data for invalid values e.g., outliers (data point which are either way too high or way too low as could be expected for the type of data) or wrongly entered format (e.g., comma instead of point, dates etc.).

Once you know the reasons for incomplete and invalid data, it would increase the quality if adjustments are put in place. Adjustments could include imputation, correction of invalid data etc.

### 3.2 Clarity

Clarity during data processing refers to the availability of appropriate documentation relating to the data, in particular information about the concepts, definitions and classifications used. Concepts, definitions, and classifications related to administrative data can differ from the

ones needed for statistical purposes and, if they differ, influence the level of consistency and comparability. As a first step we need to ensure that concepts, definitions, and classifications used for the data is available, and follows international and national standards.

A glossary will provide clarity over the concepts used and their definitions. This knowledge is important to understand what the data describes e.g., if the term “school” is used without any details about the level we might not know whether the school is at primary or secondary level, or a school providing vocational training. Using the same definitions across different source is useful to ensure that we handle the same type of information in a similar way. Similarly, when international and national classifications are used e.g., the International Classification of Diseases or International Standard Classification of Education, we can compare our data across countries. Furthermore, operating procedures or guidelines which define how the administrative data is processed give clarity of the data we are working with.

### 3.3 Consistency

Consistency refers to the ability to reliably combine statistics and data sets in different ways and for various uses. As mentioned above, deviations from international and national concepts, definitions and classifications will influence both consistency and comparability. If deviations exist, they need to be described and justified to create an understanding of the differences and why these deviations are important to keep.

Besides concepts, definitions and classifications, there might be other sources which could cause lack of comparability over time and regions/countries e.g., changes made in the geographical coverage of the regions or changes made to which type of administrative data is collected.

# 4. Quality of statistical output

This section refers to the phase where the statistical output is ready and made available to users.

### 4.1 Relevance

Relevance refers to the degree to which the statistical output meets user needs. To understand the needs of users, they need to be consulted and asked. Regular contact with main users e.g., via a user group that meets twice a year on a regular basis, is a good way to learn about their needs and uses of the statistical output. Meetings with users also provide the opportunity to inform them about upcoming changes to the statistical output e.g., in case of changes to international standards and methodologies.

Another way to learn about the needs of users is to investigate how satisfied they are with the statistical output. This could be done through an anonymous user satisfaction survey e.g., published on the website or sent directly to specific users but also through user needs assessments or user consultations. The results of such studies should be analysed and discussed with relevant stakeholders, and appropriate measures should be implemented to seek and meet user needs, if deemed realistic and sound.

### 4.2 Clarity

Clarity of statistical output refers to the availability of appropriate documentation on the data where users can learn more about the methodological choices and limitations of the statistical output. Such documentation will help the user understand the statistical output and is useful for interpretation and comparison. Therefore, published official statistics should always be accompanied by metadata, which describe the content, sources, and methods. Also, if such documentation includes a description of the quality of the statistics, the user will have a much better understanding on any caution they need to take when using the data. Changes to statistical methodologies should be communicated to users.

Also, if errors have been detected in published statistical output and when those errors are corrected, it is a good practice to point out updates e.g., on the website where the publication is available.

Finally, clarity includes aspects of whether the data is available at different type of disaggregation by e.g., Dzongkhag, gender and age as disaggregation provides additional value to the statistical output.

### 4.3 Comparability

Comparability refers to the degree to which the statistical output can be compared, both over time and region or domain. To compare statistical output over time, data needs to be

available for an adequate period of time. It depends on the type of statistical output what an adequate period of time is. For annual statistics, 5 years could be a good time period, while for quarterly statistics 5 quarters might be more relevant.

Comparability is also relevant in geographical terms. E.g., district level administrations are always interested in statistics on their district and find it useful to compare with neighbouring or similar districts. Comparing official statistics across countries is important for national benchmarking. However, to compare across countries, the statistical output needs to be produced according to international standards.

### 4.4 Timeliness and punctuality

Timeliness refers to the length of time between the date of publication, and the reference period (the time), which the statistical results are collected and calculated. The longer the time, the less relevant the data is for the user. Therefore, a way to pay appropriate attention to the issue of timeliness is to set up a timeline for the production of official statistics. It is often found that a long production time improves accuracy but reduces timeliness, while a shorter production time reduces accuracy but improve timeliness. Different users will have different preferences for the trade-off between timeliness and accuracy.

Punctuality refers to the time lag between the actual date of publication and planned date of publication of the statistical output. To meet the quality dimension of punctuality, an up-to-date time schedule (i.e., publication calendar) on upcoming publications should be made publicly available to users e.g., on the website where the statistical output is published. It is internationally recommended that a publication calendar contains information on all publication for the coming year. Changes to the expected release data should be announced well in advance so that users are informed about any delays and can plan accordingly. At the end, the aim is to publish the statistical output in accordance with the planned release date.

### 4.5 Accessibility

Accessibility refers to the ease with which users can access the statistical output. It is also about the format in which data is available and the availability of supporting information.

Since users are different and have different needs (there is a difference between the needs of a student, a policy maker, and a researcher), statistical output can with benefits be disseminated through different means and formats. First and foremost, the published statistics should be available in a format which is easy for users to access but also to download and use.

To understand the statistical output, it is a good idea to explain terms and concepts together with the numbers, and to illustrate the data with figures and graphs to facilitate an understanding of the statistics. Also, users might find it necessary to explore older publications and thus previous releases should be made easily available and accessible. If the responsibility of producing the official statistics shifts from one authority to another (or from one department to another), links should be made so users can easily find older publications.

**Annex: Resources used as inspiration to develop the Guide on assessing quality of administrative data for producing official statistics**

- BLUE-ETS/ European Union Working Group: “List of quality groups and indicators identified for administrative data sources”
- ESSnet KOMUSO/European Union Working Group: “Checklist for Evaluating the Quality of Input Data”
- Istat, Italy: “Evaluating administrative data quality as input of the statistical production process”
- Office of National Statistics (ONS), UK: “Framework on Quality of Admin Data in Statistics” Home best-practice-and-impact.github.io
- Statistical Network/SN-MIAD: “Methodologies for an integrated use of administrative data in the statistical process” (national statistical offices of Australia, Canada, Italy, New Zealand)
- Statistics Denmark: “Guidelines for Official Statistics - Self-Assessment Scheme”
  Statistics Ethiopia: “Ethiopian Data Quality Assessment Framework: Assessment Tool”
- Statistics Netherlands: “Checklist for the quality evaluation of administrative data sources”
- UNECE: “Using Administrative and Secondary Sources for Official Statistics: A handbook of Principles and Practices”
- UN DESA: “United Nations National Quality Assurance Frameworks Manual for Official Statistics”
- UNECLAD: “Questionnaire to Evaluate the Quality of Administrative Records”
- UNESCO: “Data Quality Assessment Framework for Administrative Data”

**NATIONAL STATISTICS BUREAU**
Royal Government of Bhutan
Post Box No.: 338
Thimphu : Bhutan

Telephone No.: +975-02-333296/ 335848
Fax: +975 -02-323069