Chapter 7
Quality assurance for statistics compiled from different data sources

Introduction

7.1. Chapter 7 addresses quality assurance when different data sources are used for the production of statistics. It distinguishes among statistical data sources, administrative data sources and other data sources. The UN-NQAF, set out in chapter 3, applies to all data and statistics regardless of the source, but the challenges to obtain compliance can be different depending on the data source. This chapter aims to highlight certain aspects of quality assurance that are specific, or are of special importance, to statistics that are produced (completely or partially) using specific data sources.

7.2. Section 7.A. provides definitions of the different data sources and highlights the role of the GSBPM for quality assurance. Sections 7.B. to 7.E. discuss the potential benefits and challenges when using statistical, administrative, other or multiple data sources. Table 7.1 in section 7.F. provides examples of specific elements to be assured when different data sources are used in order to mitigate the challenges that have been identified. Section 7.G. provides selected references regarding the use of different data sources.

7.A. Use of different data sources

7.3. Statistics producers, data providers and the use of different data sources in the production of statistics. Three types of statistics producers at the national level can be distinguished: the NSO, other producers of official statistics and other statistics producers outside of the NSS. In the context of this chapter, data providers are entities that own or hold the data used in the production of statistics (source data). All types of statistics producers can use any type or combination of data sources, be they statistical, administrative, or other data sources.

7.4. Definition of different data sources for the production of statistics. For the purposes of this Manual it is suggested that data sources be distinguished by their purpose and by the entity responsible for data compilation. While it is acknowledged that others may want to define and distinguish data sources differently according to their respective needs, the following definitions and classification of data sources are proposed for the purposes of this Manual:

(a) Statistical data sources are data collections created primarily for official statistical purposes by government agencies or other entities working on behalf of the government. Statistical data sources include statistical sample surveys, censuses and statistical registers. There are different types of censuses, such as population and housing censuses, business censuses and agriculture censuses, among other types. Sample surveys and statistical registers can cover different units, for example individuals, households...
of statistics. For statistics producers, “data source” is the source of the data used for the production of statistics. For statistics users, the term may refer to the place where they obtain their statistics or data. From the perspective of a user, a primary data source is the entity that produces the statistics, while a secondary data source is an entity that disseminates statistics produced by someone else.

and businesses. Statistical registers can themselves be derived from different sources. For example, statistical business registers are often based on administrative data sources;

(b) Administrative data sources are data sets created primarily for administrative purposes by government agencies or other entities working on behalf of the government. Administrative data sources include administrative registers of persons and legal entities and the records of ministries, departments and specialized agencies, such as tax returns, social services records and customs data, or data of regional or local administrations. In contrast to statistical data sources, administrative data sources are not created in response to the need for statistical data but as a part of a government function, such as the provision of services or taxation. In some cases, statistical agencies participate in the design and/or collection of administrative data. In addition, statistical agencies may be involved at different stages of the production process of administrative data, with the aim of ensuring that the data will be usable for statistical purposes;

(c) Other data sources include all data sets that are not created primarily for official statistical or administrative purposes but rather for commercial or other private purposes. Other sources include data sets created by providers of communications, media and e-commerce services, providers of services based on Earth observation and remote sensing, and private insurance companies, but also include traditional sample surveys conducted by companies for their own purposes, such as market research. In general, other data sources include data sources associated with the term “big data” unless already included, in some instances, in statistical or administrative data sources.62

7.5. Other data sources and new data sources. Often the term “new data sources” is used when referring to other data sources, and both terms may be perceived as largely interchangeable at the time of the drafting of this Manual. However, this use of the term “new data sources” is misleading in multiple ways. New data sources can emerge from all three categories of data sources, be they statistical, administrative or other data sources. Furthermore, some of these data sources have been used in some countries for many years, and the notion of what is considered new changes over time.

7.6. List of other data sources. The following list of other data sources is provided only for illustrative purposes and is an attempt to reflect other major sources of data. It does not aim to provide an exhaustive list nor a classification of such sources.63

(a) Cross-country sample surveys by supranational organizations or international enterprises;

(b) Data compiled and maintained by private professional organizations or business associations, or by non-profit institutions in general;

(c) Data and records compiled and maintained and/or owned by enterprises that cover large parts of the population of statistical units, in particular e-commerce, media and telecommunications providers, but also other enterprises that provide services directly to individuals or businesses, such as insurance companies, banks and airlines;

(d) Earth observation and remote sensing;

(e) Thematic mapping and monitoring systems (e.g., field-monitoring stations for water quality, air pollution, etc.);

(f) Research/scientific and pilot studies;

(g) Citizen-generated data.

62 The Global Working Group on Big Data appears to distinguish between big data, administrative data and traditional statistical data sources (see https://unstats.un.org/bigdata/).

7.7. Data sources, the statistical production process and quality assurance. The statistical production process consists of several phases. The GSBPM addresses the following phases: the specification of needs, design, building, data collection, processing, analysis, the dissemination of the products and the evaluation of the process.\(^64\) The GSBPM provides a universal basis for the identification of pertinent quality characteristics and the formulation of quality indicators by analysing the statistical production process. The use of the GSBPM, if properly applied and interpreted, can pinpoint quality issues in the use of specific data sources.\(^65\) In addition, the selection of the appropriate data source itself is part of the statistical production process. However, overarching processes such as statistical infrastructure and management and support functions also need to be considered when identifying the quality principles and indicators that are most relevant for statistics compiled from a particular data source, or when selecting the data source.

---

**Box 7.1**

Elements to be assured related to the selection of the data source during the specification of needs phase

Subprocesses 1.1, Identify needs; 1.2, Consult and confirm needs; and 1.3, Establish output objectives
- To what extent does the data source satisfy information demand? (principle 14: Assuring relevance)

Subprocess 1.4, Identify concepts
- Is there metadata for administrative and other sources to determine whether relevant variables are available? (principle 14: Assuring relevance, and principle 18: Assuring coherence and comparability)

Subprocess 1.5, Check data availability
- To what extent have legal constraints regarding data collection, acquisition and use been assessed and any necessary changes been proposed? (principle 7: Assuring statistical confidentiality and data security, and principle 2: Managing relationships with data users, data providers and other stakeholders)
- To what extent do current data sources meet user requirements, taking into consideration the conditions under which they would be available and any restrictions on their use? (principle 14: Assuring relevance)
- If current data sources do not fully meet user requirements, to what extent has a strategy been proposed to fully meet user requirements? (principle 14: Assuring relevance)
- Is there an advance notification plan about the forthcoming changes to the data source? Is a contingency plan for changes to the data or data source in place? (principle 14: Assuring relevance)
- Has the completeness of data source(s) been evaluated, such as the percentage of units not belonging to the target population; the percentage of units missing from the target population; the coverage of the data; the absence of values for key variables; missing values in the source; and the total percentage of empty cells? (principle 15: Assuring accuracy and reliability)

Subprocess 1.6, Prepare and submit business case
- Has the data source been evaluated in terms of its cost-effectiveness? (principle 11: Assuring cost-effectiveness)

---

\(^64\) See GSBPM (version 5.1, January 2019). Available at [https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1](https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1).

\(^65\) It must be noted that for some sources, such as big data, the statistical paradigm is different. The design phase of the GSBPM, especially subprocess 2.3, Design collection, and subprocess 2.4, Design frame and sample, is being substituted by the study on how to extract, interpret, transform and store the data so as to make them manageable and structure them according to statistical needs into units and variables.
7.8. As also indicated in chapter 4, the GSBPM can help countries understand where to start and how to implement the NQAF by analysing the existing statistical production processes. Box 7.1 is based on the work of the Economic Commission for Europe (ECE) on quality indicators for the GSBPM and shows elements to be assured (indicators) related to the selection of the data source during the specification of needs phase, the first phase of the statistical production process.\textsuperscript{66} Box 7.1 also shows the link of the indicator to the UN-NQAF principle being addressed.

7.B. Using statistical data sources: potential benefits and challenges

7.9. Potential benefits in using statistical data sources. The main advantage of statistical data sources is that they allow data to be obtained according to specified needs and predefined statistical concepts.

7.10. Challenges in using statistical data sources. Statistical data sources impose challenges that directly affect the quality of the statistics produced from such sources. The major quality considerations that concern statistics produced from statistical sources are:

(a) The high cost of production (principle 11: Assuring cost-effectiveness);
(b) The availability of resources (principle 9: Assuring adequacy of resources);
(c) The low frequency of conducting sample surveys and censuses (principle 14: Assuring relevance);
(d) The respondent burden and the willingness of respondents to provide information (principle 13: Managing the respondent burden);
(e) Sampling and non-sampling errors (principle 15: Assuring accuracy and reliability);
(f) The need for complex sampling designs (principle 10: Assuring methodological soundness);
(g) The need for careful planning, implementation of instruments, training and supervision of staff and rigorous evaluation (principle 12: Assuring appropriate statistical procedures).

7.11. These quality issues may arise at different phases of the production process, and the GSBPM can be a useful tool for analysing these challenges.

7.C. Using administrative data sources: potential benefits and challenges

7.12. Potential benefits in using administrative data sources. The use of administrative data sources offers many potential advantages, such as cost-effectiveness, the reduction of the respondent burden, improved timeliness and improved relevance, accuracy and reliability due to their ability to obtain highly disaggregated data. In many statistical domains administrative data are indispensable to the production of statistics.

7.13. Challenges in using administrative data sources. There are multiple concerns and limitations when using administrative data sources, which should be addressed to better realize the potential benefits. The main concerns and limitations of administrative data sources in the context of quality assurance (quality threats) include:

(a) Insufficient cooperation with the providers/holders of data, a frequent lack of clarity with regard to the roles and responsibilities of different stake-


\textsuperscript{67} This will not apply to population censuses based on information from civil registers as already performed in some countries and as planned in the European Union. However, register-based population censuses require a fully functioning and complete civil registration system, which many developing countries do not have.
Quality assurance for statistics compiled from different data sources

7.D. Using other data sources: potential benefits and challenges

7.14. Potential benefits in using other data sources. The use of other data sources is an opportunity to rethink the elements that constitute the institutional environment and statistical processes and outputs. For example, the use of other data sources can offer the opportunity to overcome resource limitations, allow much more frequent and timely reporting, provide more objective information and, most importantly, generate data on phenomena and their aspects (disaggregations) that are difficult or impossible to capture with traditional statistical and administrative data sources. This leads to improved relevance.

7.15. Challenges in using other data sources. The full realization of the potential benefits of other sources is only possible if the identified concerns and quality limitations (quality threats) are addressed. Many of these concerns and limitations relate to the fact that entities outside the NSS own, hold or have responsibility for the other data sources. The major quality issues associated with the statistics compiled from other sources of data include:

(a) Limited access to other data sources and legal challenges regarding their access (as can also be the case for administrative sources) may require arrangements with the data providers (e.g., government agencies and private sector and research institutions); a lack of knowledge about the exist-
ence of such data; and the sustainability of the source over time (principle 2: Managing relationships with data users, data providers and other stakeholders);

(b) Incoherent use or lack of use of statistical standard concepts, definitions and classifications (principle 3: Managing statistical standards) may occur that put the accuracy, reliability, coherence and comparability of the resulting statistics in question;

(c) Providers of data (which may be the owner or holder of the data) are not subject to and may not adhere to the Fundamental Principles of Official Statistics and associated statistical quality principles such as professional independence (principle 4) and commitment to quality (principle 8);

(d) Utilizing data for statistical purpose may potentially put the confidentiality and privacy of individuals, households and businesses at risk, depending on how detailed the data being published are (principle 7: Assuring statistical confidentiality and data security);

(e) Data from sources such as mobile phones or social media are not representative of the entire population and may cause serious selection bias when used for statistical purposes (principle 10: Assuring methodological soundness, principle 12: Assuring appropriate statistical procedures, and principle 15: Accuracy and reliability);

(f) There may be interference and bias in the statistical production process and a lack of information about how the data are being produced (principle 4: Assuring professional independence, principle 5: Assuring impartiality and objectivity, and principle 6: Assuring transparency);

(g) The data collection is not designed for statistical purposes and may not provide the information required by users (principle 14: Assuring relevance) or may not correctly describe the phenomena that are to be measured (principle 15: Assuring accuracy and reliability);

(h) Technical access to and use of the data, in particular large amounts of data, can be very challenging and may require significant IT resources and expertise (principle 12: Assuring appropriate statistical procedures);

(i) Under- or over-coverage of the target population, population changes over time, misalignment of the reference period compared to what is desired and inherent bias as a result of the original purpose of the data set may occur (principle 15: Assuring accuracy and reliability, and principle 18: Assuring coherence and comparability);

(j) There may be an inability to quantify uncertainty when there are only non-sampling errors (principle 15: Assuring accuracy and reliability).

7.16. Conclusions of the United Nations Statistical Commission on the use of new data sources. Interest in the use of other data sources (including the data sources covered by the term “big data”) for the compilation of official statistics has been growing over the years. With the adoption of the 2030 Agenda for Sustainable Development and in view of the global and national data requirement to measure progress towards the SDGs, the use of other “new” data sources was identified as one of the top priorities for NSSs. Considering both the benefits and concerns/limitations of the use of the new data sources in official statistics, the Commission promoted their use and, at the same time, emphasized “the importance of ensuring the quality of data derived from new sources and new data providers, including those outside the official statistical system”. 68 Specifically, the Commission:

68 See E/2017/24, decision 48/106.


71 There are extensive methodologies for quality assurance and verification of the headline figures of national accounts such as gross domestic product and gross national income (GNI) owing to their widespread use for policymaking and administrative purposes, such as the determination of countries’ financial contributions to regional and international organizations. Please see Eurostat, “Monitoring GNI for own resource purposes”, available at https://ec.europa.eu/eurostat/statistics-explained/index.php/Monitoring_GNI_for_own_resource_purposes.
7.17. Other data sources should be used for the purposes of compiling official statistics if the quality of the statistical process and outputs are sufficiently assured. Quality assurance activities may be undertaken by the NSO and other NSS members on their own and/or as part of the partnership agreements with data providers.

7.E. Using multiple data sources: potential benefits and challenges

7.18. Potential benefits of using multiple data sources. The use of multiple data sources is increasingly common in the compilation of official statistics, especially for the monitoring of the SDGs. The use of one single source may not allow for the required disaggregation of data. However, disaggregated data may be obtained through integration with data from a different source.70 For example, combining data sources may provide a higher coverage of the target population and consequently the compilation of statistics with the required disaggregation level. Moreover, the integration of data sources can provide a better understanding of investigated phenomena due to the availability of more information. Data integration can include:

(a) Combining data from multiple sources as part of the creation of integrated statistics, such as national accounts, for which the quality depends on the quality of the basic statistics that are being used for its compilation and the quality of the integration and estimation processes;71
(b) Pooling data, for example, by merging sample survey data from different collection cycles in one data set (pool of data) with the aim of increasing the effective number of observations of a phenomenon, which allows more accurate estimates of the target population and, potentially, of the estimates pertaining to small domains, to be obtained;72
(c) Statistical matching (data fusion) and record linkage routines that link microdata from different sources;73
(d) Prioritization, when two or more sources contain data for the same variable but have potentially different values.

7.19. Major challenges in using data from multiple sources are:

(a) Assuring methodological soundness (principle 10), as it relates to different coverage and the use of different concepts and definitions, among other things;
(b) Assuring appropriate statistical procedures (principle 12), as they relate to the appropriate procedures, skills and knowledge required to link and integrate data;
(c) Assuring confidentiality and data security (principle 7);
(d) Assuring coherence and comparability over time (principle 18); the quality of the entire integration process should be assessed by calculating proper indicators.74


The methods can be differentiated as follows: (a) if the units of observation in the survey samples are exactly matched using units’ identifiers such as social security numbers or names and addresses that are error free, it is called “merging” or “exact matching”; (b) “record linkage” refers to integration procedures where units’ identifiers are not error free or, in the absence of identifiers, key variables (name, surname, birthday, gender, etc.) are used to link units; and (c) statistical matching (also called data fusion) typically aims to achieve a complete data file using data from different sources that contain the same units of observation but either have different identifiers or do not contain the same units at all. Statistical matching or data fusion investigates the relationship between variables that are not jointly observed. This can be done by creating synthetic records (statistical matching or data fusion at the micro level) or by estimating parameters of a model such as a regression (statistical matching or data fusion at the macro level). Traditionally, statistical matching is done with regard to the variables common to all data files used in the production of a particular statistical output.

See, for example, Statistics Canada, “Record linkage project process model”, 2017. Available at www.statcan.gc.ca/pub/12-605-x/12-605-x2017001-eng.htm.
7.F. Elements to be assured for different types of data sources

7.20. Table 7.1 lists examples of specific elements to be assured when statistical, administrative, other or multiple sources of data are used. These elements can help mitigate the challenges that have been identified for each of the data sources.

Table 7.1
Examples of some specific elements to be assured when different sources of data are used

<table>
<thead>
<tr>
<th>UN-NQAF principles</th>
<th>Data source</th>
<th>Requirements/elements to be assured*</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Managing the statistical system</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Principle 1: Coordinating the national statistical system</td>
<td>Statistical</td>
<td>There is a body that ensures coordination of sample surveys and their methodological soundness throughout the NSS (1.2, 10.1)</td>
<td>Sample surveys should be coordinated and integrated for cost-effectiveness and methodological soundness</td>
</tr>
<tr>
<td></td>
<td>Administrative</td>
<td>Administrative records are systematically linked with records from other relevant administrative data systems, as permitted by applicable laws (1.2, 10.1, 11.5)</td>
<td>The linking of administrative data will allow the best possible use of the information already available</td>
</tr>
<tr>
<td></td>
<td>Administrative</td>
<td>There is a unit that discusses and provides support for the use of administrative data sources and other data sources within the NSS (1.2, 10.1, 11.5)</td>
<td>The use of administrative and other data sources poses unique challenges that should be addressed by sharing experiences and best practices</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Principle 2: Managing relationships with data users, data providers and other stakeholders</td>
<td>Administrative Other</td>
<td>Laws and regulations are in place to allow the required access to administrative and other data sources (2.5, 2.6)</td>
<td>Limited access is a frequent obstacle for the use of administrative data sources and other data sources</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>Partnership agreements with data providers are in place (2.6)</td>
<td>The use of other data sources benefits from a partnership with the data provider</td>
</tr>
<tr>
<td>Principle 3: Managing statistical standards</td>
<td>Statistical Administrative Other Multiple</td>
<td>The national statistical agency monitors the use of statistical standards (3.2)</td>
<td>The use of standard concepts, definitions and classifications facilitates the use and integration of data</td>
</tr>
<tr>
<td>Managing the institutional environment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Principle 5: Assuring impartiality and objectivity</td>
<td>Other</td>
<td>Data sources and methodologies are chosen on an objective basis (5.3)</td>
<td>Data from other sources should only be used if they are selected and compiled according to professional standards</td>
</tr>
<tr>
<td>Principle 6: Assuring transparency</td>
<td>Other</td>
<td>The terms and conditions under which other data sources are being used are made public (6.1)</td>
<td>Users must have the ability to know what other data sources are being used and how they are being used</td>
</tr>
<tr>
<td>UN-NQAF principles</td>
<td>Data source</td>
<td>Requirements/elements to be assured*</td>
<td>Explanation</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------</td>
<td>-------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Principle 7: Assuring statistical confidentiality and data security</td>
<td>Other Multiple</td>
<td>Special procedures exist for keeping linked data secure (7.2, 7.3, 7.5)</td>
<td>The use of data from multiple sources frequently requires establishing record-linkage</td>
</tr>
</tbody>
</table>

### Managing statistical processes

<table>
<thead>
<tr>
<th>Principle 10: Assuring methodological soundness</th>
<th>Data source</th>
<th>Requirements/elements to be assured*</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principle 11: Assuring cost-effectiveness</td>
<td>Statistical</td>
<td>Different data sources are evaluated in terms of cost-effectiveness (11.2, 11.5)</td>
<td>Censuses and sample surveys are very expensive (therefore, some countries are moving to register-based censuses if they have a civil register)</td>
</tr>
<tr>
<td>Principle 12: Assuring appropriate statistical procedures</td>
<td>Statistical</td>
<td>Censuses and sample survey data are geocoded (12.2)</td>
<td>Geocoding of censuses and sample survey data facilitates data integration and allows disaggregation by location</td>
</tr>
<tr>
<td>Administrative Other</td>
<td>Procedures for coding, editing, imputing, etc., are documented (12.2)</td>
<td>The steps needed to use administrative data and other data sources for statistical purposes need to be documented</td>
<td></td>
</tr>
<tr>
<td>Multiple</td>
<td>Procedures for linking data from different sources are established and are documented and available (12.2)</td>
<td>Changes in procedures for linking data from multiple sources may threaten consistency of results over time</td>
<td></td>
</tr>
<tr>
<td>Principle 13: Managing the respondent burden</td>
<td>Statistical</td>
<td>There is a communication plan to explain the purpose of the sample survey and census (13.2)</td>
<td>The quality of the sample survey and census results depends on the willingness of the respondents to reply accurately to the questions</td>
</tr>
</tbody>
</table>

### Managing statistical outputs

<table>
<thead>
<tr>
<th>Principle 14: Assuring relevance</th>
<th>Data source</th>
<th>Requirements/elements to be assured*</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Administrative Other Multiple</td>
<td>Statistics based on new data sources are being developed in response to society's emerging information needs (14.3)</td>
<td>Taking new user needs into account is crucial for the use of all data sources</td>
<td></td>
</tr>
<tr>
<td>UN-NQAF principles</td>
<td>Data source</td>
<td>Requirements/elements to be assured*</td>
<td>Explanation</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------</td>
<td>--------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Principle 15: Assuring accuracy and reliability</td>
<td>Statistical</td>
<td>Sampling errors are measured, evaluated and systematically documented (15.1, 15.2)</td>
<td>Accuracy and reliability of sample survey results rely on the appropriate sample survey design</td>
</tr>
<tr>
<td></td>
<td>Administrative Other Multiple</td>
<td>Non-sampling errors are described and estimated when possible (15.1, 15.2)</td>
<td>Non-sampling errors could contribute significantly to uncertainty in statistical products, and are difficult to quantify</td>
</tr>
<tr>
<td>Principle 18: Assuring coherence and comparability</td>
<td>Administrative Other</td>
<td>Results are compared with other available information and over time (18.2)</td>
<td>The use of administrative data and other data sources requires additional attention to assess coherence and comparability, as they are not compiled primarily for statistical purposes</td>
</tr>
<tr>
<td>Principle 19: Managing metadata</td>
<td>Statistical Administrative Other Multiple</td>
<td>Metadata are provided and cover all relevant aspects of the use of different data sources (19.2)</td>
<td>Statistics must always be accompanied by metadata to allow the user to evaluate the statistics being provided</td>
</tr>
</tbody>
</table>

*References to the UN-NQAF requirements concerned (or elements to be assured under the UN-NQAF requirements) are indicated in parentheses (see chapter 3 and the annex for details).

### 7.G. Selected references on quality assurance

7.21. This section provides selected references on quality assurance when specific data sources are being used.

7.22. **Selected references on quality assurance for statistics obtained from statistical sources:**


7.23. **Selected references on quality assurance for statistics produced using administrative data sources:**


Section 7.24. Selected references on the use of other data sources. At this time, there are no specific references on quality assurance for the use of other, including “new”, data sources. Some references regarding the use of other data in general include:


7.25. Selected references on data integration. Good practices in assuring the quality of data obtained by the integration process from different sources are described in several documents:


(c) Statistics Canada, “Record linkage project process model”, 2017. Available at www.statcan.gc.ca/pub/12-605-x/12-605-x2017001-eng.htm;

