Data quality and metadata

Developing National Systems of Tourism Statistics: Challenges and Good Practices
Regional Workshop for South East Asian countries, 16 – 19 June 2009

United Nations Statistics Division
Statistics of International Trade in Services Section

Data quality

Available resources to collect, analyze and store tourism statistics will make an effect on the quality of the data.

Outline

Quality

- Relevance of statistical concepts
- Accuracy
- Timeliness
- Accessibility and clarity of information
- Comparability of statistics
- Coherence
- Completeness/coverage
- Cost and burden

Metadata

Data quality

Several statistical organizations and countries have developed definitions of quality, outlining the various dimensions (aspects) of quality and quality measurement and have integrated them into **quality assessment frameworks**

Examples of quality assessment frameworks:

European Statistical System (ESS) focuses on the statistical outputs and defines quality with reference to six criteria

IMF Data Quality Assessment Framework (DQAF) – holistic view of data quality, including governance of statistical system

OECD Quality Measurement Framework – takes the user's side to approach quality – uses seven dimensions

Data quality

No unique indicator of data quality – several criteria are used, like:

- Relevance of statistical concepts
- Accuracy
- Timeliness
- Accessibility and clarity of information
- Comparability of statistics
- Coherence
- Completeness/coverage
- Cost and burden

Relevance

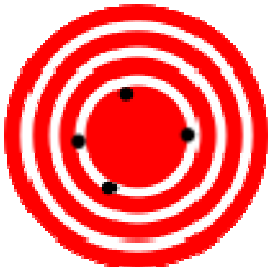
Relevance in statistics is assured when statistical concepts meet current and potential users' needs. Identification of the users and their expectations is a must.

Accuracy

Accuracy is defined as the closeness between the computations or estimates and the (unknown) true population value.

Assessing the accuracy of an estimate involves analysing the total error associated with the estimate: bias (+/-) and standard deviation (when possible).

High **accuracy** but low **precision** (large sample error?)



High **precision** but low **accuracy** (biased estimate?)

Accuracy (cont.)

- **Sampling errors**: lack of accuracy due to observing only a sample instead of the whole population (quantifiable by the *standard error*)

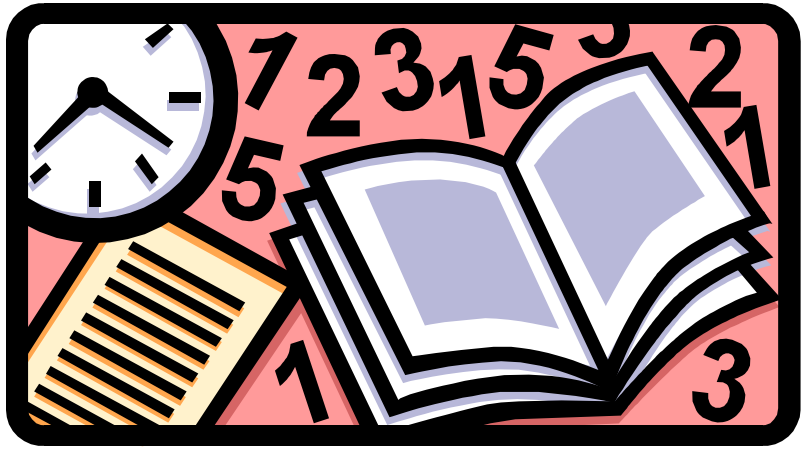
- **Non-sampling errors:**
  - Coverage errors (under- or over coverage)
  - Non-response errors (surveys)
  - Measurement errors
  - Processing errors
  - Model assumption errors

Timeliness

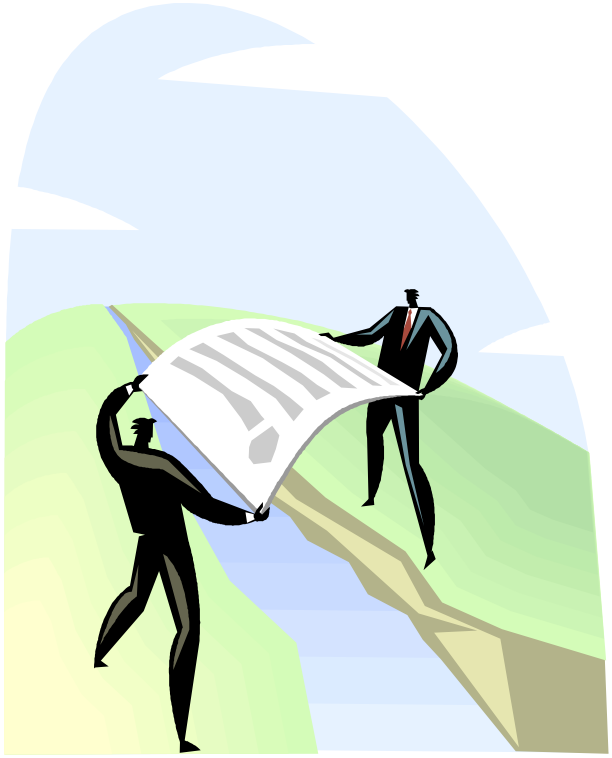
Users want the latest data that are published frequently and on time at pre-established dates.

Data

- ✓ Collection
- ✓ Editing
- ✓ Consolidation
- ✓ Dissemination



Accessibility and clarity of information



Statistical data are most valuable when they are:

- ✓ Easily accessible by users
- ✓ Available in the form users desire
- ✓ Adequately documented – accompanied by good **metadata**

Assistance in using and interpreting the statistics should also be forthcoming from the providers.

Comparability of statistics

Statistics for a given characteristic have the greatest usefulness when they enable reliable comparisons of values across geography and over time.

Providing comparable country data makes it possible for international organizations to publish regional and world totals.

Comparability of statistics (cont.)

For comparability the following are needed:

- ✓Common definitions
- ✓Common unit of measurement
- ✓Unified methodology
- ✓Timely submission of data to international organizations

Coherence

Coherence is the measure of the extent to which one set of statistical characteristics agrees with an other and can be used together (with each other) or as an alternative (to each other).

Completeness/coverage

The component of completeness reflects the extent to which the statistical system in place answers the users' needs and priorities by comparing all user demands with the availability of statistics.

Cost and burden

Although not measures of quality, they are positively correlated with quality.

Costs: office space, utility bills, staff-hours involved, funding of surveys, etc.

Response burden: simplest way to measure is the time spent by the respondent to provide information

A compromise between quality and cost and burden must be achieved

Metadata

- Statistical metadata facilitate sharing, querying and understanding of statistical data over the lifetime of the data. They also refer to any methodological descriptions on how data are collected and processed.

- Metadata is essential for the interpretation of statistical data.

Metadata

There is a bidirectional relationship between metadata and quality:

- ✓Metadata describe the quality of statistics
- ✓Metadata are themselves a quality component improving the availability and accessibility of statistical data

Metadata

As a minimum segmentation, the following two levels of metadata are recommended:

- Structural metadata presented as an integral part of the data tables – for example footnotes explaining the statistical output
- Reference metadata providing details on the content and quality of data – for example a description of data sources and statistical processes and estimations related to producing the statistics

Metadata

- Metadata provides a mechanism for comparing national practices in the compilation of statistics. This may help and encourage countries to implement international standards and to adopt the best practices.

- Better harmonization of approaches will also improve general quality of the data.

UNWTO metadata project

INVENTORY OF DATA SOURCES: BASIC FORMAT *

Country

The statistical procedure
Title
Data source
Description

Administration
Institutional framework
- Institution in charge
  - Unit responsible within this institution
  - Other units and/or institutions involved

Data dissemination
- Name of publication
- Published by
- Frequency
- Length of time series available
- Reference period for the latest available publication
- Timeliness
- Access to data via the Internet

Coverage
Geographical coverage
Frequency
Reference period for the first procedure
Reference period for the last procedure conducted

Concepts
Specific subjects
Variables of reference

Data collection, manipulation / accounting conventions, etc.
Frame
Type of enumeration
Sampling
Distribution of the sample over time
Data collection principles
Geographical breakdown
Reporting units
Other issues

Observations on its use

Additional documentation

- The UNWTO metadata project is based on IMF's data quality assessment frameworks: General Data Dissemination System (GDDS), Special Data Dissemination Standard (SDDS) as well as the Standard Data and Metadata eXchange project (SDMX).

- UNWTO suggests three stages:
  1. Inventorying data sources
  2. Inventorying available data related to tourism statistics
  3. Checking the quality of data related to a potential TSA exercise.

More information: http://www.unwto.org/statistics/metadata/metadata.html