Some key concepts related to metadata

– to be used in Statistics Norway's metadata systems

by Anne Gro Hustoft and Hans Viggo Sæbø

Content

Background

During our work on a metadata strategy for Statistics Norway, it became clear that while our statistics are based on a good common understanding of concepts, we lacked formal definitions for the most central metadata concepts. This was a potential source of misunderstandings amongst people working in different subject matter areas or on different types of tasks. This again could lead to misunderstandings and lower efficiency and quality when exchanging data and metadata and filling up of Statistics Norways metadata or metadata-driven systems such as StatBank (dissemination of tables), Stabas (classifications) and Vardok (definitions of variables).

The establishment and documentation of key concepts related to metadata is an important part of the implementation of our metadata strategy. This document is a start in this direction. It is made in close contact with those working in statistical methods, IT, production and dissemination of statistics. It was subject to a hearing round in all these departments. The document has also been discussed in our metadata forum, in the steering group for our metadata strategy and in our standards committee.

A useful, ongoing work is the discussion and recommendation on how these concepts are to be used in the practical, everyday work of filling up our metadata systems.

At a more detailed level there is also a need for precise definitions to be used for machine-to-machine communication. This work is being carried out in a web-services project "Service library for metadata systems". These definitions must also be consistent with the more general, key concepts.

People working on different tasks can misunderstand each other because their interpretation of key concepts depends on their purpose, which part of the organisation they belong to and the level of detail they are working on. Statistics is traditionally published in the form of tables and some of our key concepts are developed to describe the contents of these tables. The same terms can have a slightly different meaning at the micro-level i.e. when we describe individual data stored in our data archive. Another source of misunderstandings can be that terms developed within one domain e.g. social statistics can be interpreted differently from the same terms in another domain e.g. macroeconomic statistics (e.g. national accounts). Due to these problems some terms are not included in our list of key concepts and others have been given a specific context. In most cases, it is more productive to work on a precise definition of the concepts than to discuss the naming of these. The list of key concepts with their definitions is an attempt in this direction.

About the definitions

This work started by looking at international definitions. common practise in Statistics Norway and elsewhere. The SDMX-initiative (Statistical Data and Metadata Exchange) has collected together many definitions used by international organisations such as Eurostat, the European central bank, BIS, OECD, IMF, UN and the World Bank. When relevant for our definitions the source is listed as SDMX even if this is not the primary source.

Concepts listed in this document are grouped under statistics, metadata, quality and registers and thereafter concepts related to the units that statistics builds upon and their properties. Cross-references between concepts are underlined.

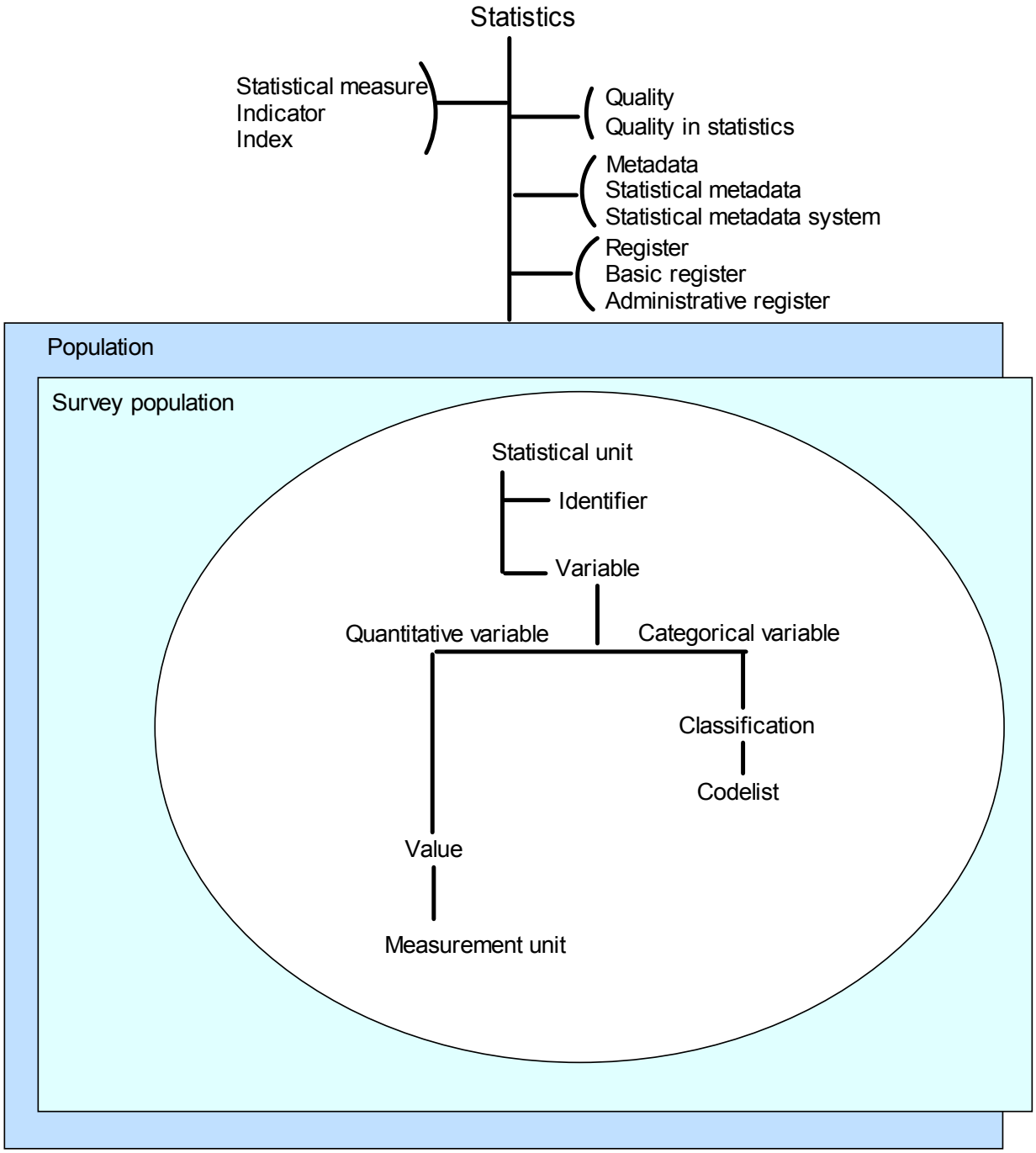
Most concepts are given as titles in bold-face font (e.g. the first concept 'statistics'). In some cases it is easier to define together concepts that are related to one another. In that case it may be intuitive to understand what is meant by the concept but difficult to make a precise definition for the concept alone. (e.g. the concept 'data' is only mentioned under the concept 'statistics') We have provided examples and synonyms where possible.

The concepts that are included are those where our experience has shown that misunderstandings can arise when metadata are put into our central metadata systems aswell as general definitions connected to statistics and metadata. Many other concepts could have been defined. The Metadata Common Vocabulary (part of the SDMX- initiative) lists over 300 definitions in English. We have also used several handbooks and documents in Norwegian available only within Statistics Norway. The first

version of this document was produced in Norwegian for use within Statistics Norway but it has later been translated to English in a reduced version more appropriate for an international audience.

The figure below shows the concepts in the order in which they are presented in this document.

Figure 1. Defined concepts



DEFINITIONS

Statistics
Statistics are numerical data concerning a group or a phenomenon which become apparent through comparing and processing information about the individual units in the group or a selection of these units, or through systematic observation of the phenomenon.
Source: Statistics Act

Statistical measure

Source: SDMX: *A summary (means, mode, total, index etc.) of the individual quantitative variable values for the statistical units in a specific group (study domains).*

Examples: average, median, total, error range, index.

Indicator

An indicator is a measure that is derived from data and/or statistics, which indicates the status or development within a specific area.

Examples: Indicators for sustainable development, social indicators

Index

Source: SDMX and ISI (2003):
A quantity that shows by its variations the changes of a magnitude over time or space

Examples: price index, production index.

Quality

Source: SDMX
Quality is defined as the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs.

Quality in statistics
Statistics Norway defines quality of statistics with reference to six criteria:

Source: SDMX og Eurostat (2003):
- *Relevance*
- *Accuracy*
- *Timeliness and punctuality*
- *Accessibility and clarity*
- *Comparability and coherence.*

Metadata

Source: SDMX and ISO/IEC FCD 11179-1: *Metadata is data that defines and describes other data.*

Synonyms: Documentation, information (about data)

Statistical metadata
Statistical metadata is structured information about statistics. This includes information used for producing, disseminating, understanding, finding and (re)using statistics.

Statistical metadatasystem

Source: SDMX og UNECE (2000): *A statistical metadata system is a data processing system that uses, stores and produces statistical metadata.*

Register
A register is (ideally) a complete inventory of the statistical units within a specific population. Different variables are used to describe these units. All statistical units in a register have an identifier which makes update possible.

Source: A combination of UNECE (2000) and a definition from Statistics Sweden (2004)

Basic register
A basic register is a register which defines and identifies basic statistical units..

- Example: Population register

Administrative register
An administrative register is a register established by a central government administration to support their activity.

Population

Source: SDMX and UN statistical office :

Population is the total membership or population or "universe" of a defined class of people, objects or events.

There are two types of population, viz, target population and survey population.

A target population is the population outlined in the survey objects about which information is to be sought and a survey population is the population from which information can be obtained in the survey.

The target population is also known as the scope of the survey and the survey population is also known as the coverage of the survey.

Statistical unit

Source: SDMX and UNECE (2000):
An object of statistical survey and the bearer of statistical characteristics. The statistical unit is the basic unit of statistical observation within a statistical survey.

There are three related concepts:

- *Observation units are those entities on which information is received and statistics are compiled.*
- *A reporting unit is a unit that supplies the data for a given survey instance*
- *Analytical units are real or artificially constructed units, for which statistics are compiled.*

Identifier

An identifier gives a unique identification of the statistical unit included in the statistics.

Example: Personal identification number

ISO/IEC FDIS 11179-1

A sequence of characters, capable of uniquely identifying that with which it is associated, within a specified context.

Variable

Source: SDMX and UN statistical office:
A variable is a characteristic of a unit being observed that may assume more than one of a set of values to which a numerical measure or a category from a classification can be assigned (e.g. income, age, weight, etc. and "occupation", "industry", "disease" etc).

Measurement unit
A measurement unit has a measurement type
(e.g. currency: NOK, Euro….) and provides the level of detail (e.g. NOK, 1000 NOK..) for the value of the variable

Synonym: Unit of measure
Source: SDMX:
The actual unit in which the associated values are measured.
Comment: *List of units of measure used for the data disseminated (e.g. Euro, %, number of persons) with a specified order of magnitude (e.g. thousand, million).*

Classification

Source: SDMX and UN statistical office:
A classification is a set of discrete, exhaustive and mutually exclusive observations, which can be assigned to one or more variables to be measured in the collation and/or presentation of data. The terms "classification" and "nomenclature" are often used interchangeably, despite the definition of a "nomenclature" being narrower than that of a "classification".

Code list

Source: SDMX:
A code list is a predefined list from which some statistical concepts (coded concepts) take their values.

Context: *Each code list has the following properties: a) identifier (it provides a unique identification within the set of code lists specified by a structural definitions maintenance agency); b) name (also unique); c) description (a description of the purpose of the code list); and d) code value length (either an exact or a maximum number of characters and a type, i.e. numeric or alphanumeric).*

References

Several handbooks and documents that are only available within Statistics Norway.

Eurostat (2003): *Assessment of Quality in Statistics: Definition of quality in statistics,* WG, Luxembourg, October 2003

Eurostat (2005): *European Statistics Code of Practice.* Adopted by the SPC 24 February 2005

FNs statistiske kontor: *United Nations Glossary of Classification Terms*, publisert på http://unstats.un.org/unsd/class/family/glossary_short.htm.

ISI (2003): *The Oxford Dictionary of Statistical Terms* - Oxford University Press 2003

SCB: Report 2004:2- Registerstatistikk - administrativa data för statistiska syften.

SDMX: *Metadata Common Vocabulary.* Edited by Marco Pellgrino (Eurostat) and Denis Ward (OECD), siste versjon oktober 2005

SSBs Metadatastrategi - Sæbø, Andersen, Hoel, Hustoft, Linnerud, Torvbråten: SSB-rapport 2005/2

Strategy 2002-, Statistics Norway

UNECE (2000): *Terminology on Statistical Metadata*, Conference of European Statisticians - Statistical Standards and Studies No 53.