Statistisk sentralbyrå
Statistics Norway

Steady States of Data in SSB

Version 2.0

TALL SOM FORTELLER

NOTATER / DOCUMENTS

2024/45

The Standards Committee

In the series Documents, documentation, method descriptions, model descriptions and standards are published.

Published: 28 October 2024

ISBN 978-82-587-2951-5 (electronic)
ISSN 2535-7271 (electronic)

Symbols in tables	Symbol
Category not applicable Figures do not exist at this time, because the category was not in use when the figures were collected.	.
Not available Figures have not been entered into our databases or are too unreliable to be published.	..
Confidential Figures are not published to avoid identifying persons or companies.	:
Decimal punctuation mark	.

Preface

There is a need for a common description of the steady states of data that are part of statistical production in Statistics Norway (SSB). A common understanding and terminology will help in our continuous work to improve the production of statistics, contribute to verifiability and facilitate increased reuse of data.

Statistics Norway, 1 October 2023

Arvid Olav Lysø

Abstract

The purpose of this document is to describe the five steady states of data in Statistics Norway, and how they should be documented.

Contents

1. Introduction

The aim of this document is to contribute to a clearer description of steady states of data in the production of statistics in Statistics Norway (SSB). The descriptions are specific enough to be of help in the continuous work to improve the production of statistics, contribute to verifiability and facilitate increased reuse of data.

The UNECE's process model, the Generic Statistical Business Process Model (GSBPM)[1], describes the production process for official statistics. It describes and defines all processes and sub-processes, from the work that is done before data is collected, until the final statistical product is disseminated.

Once a data set is available in SSB, in its least processed form, it will change as a result of the various steps in the production process. A steady state of data is the result of a data set having gone through given operations and processes. Well-defined steady states of data are important to achieve:

- standardized and recognizable production runs
- verifiable statistics
- internal and external reuse of data
- identify data that must be archived according to the Norwegian Archives Act.

Naturally, the focus of SSB's quality work[2], is on our statistics. At the same time, expectations and requirements for SSB's ability to share data, have increased significantly in recent years. This means that we in addition to producing statistics, have an increased focus on producing high-quality, reusable data sets. An important prerequisite for reuse is that those who want to reuse the data can understand the data and know what changes the data has undergone. Use of data in SSB and reuse within and outside SSB requires good metadata. Definitions of steady states of data and other statistical terms must therefore be harmonized to the greatest extent possible with international statistical frameworks and definitions.

The term verifiability is used several places in this document. Ideally, we should produce statistics so that posterity or an independent body with access to the data and our documentation will arrive at the same statistical results as ourselves.

The data states described are *source data*, *input data*, *processed data*, *statistics*, and *output data*. The first three steady states are mainly microdata that provide information about individual units, while statistics and output are mainly aggregated data. To describe the data states in a good way, we also need the term "start data". This is not a steady state of data but is more tied to the production process itself. Start data is the data with which to start processing. This means that start data may consist of one's own steady state of input data, as well as various data shared from others in SSB, e.g. someone else's processed data. Some statistics will have start data that only consists of their own input data, others will have start data that consists only of shared SSB data, while some will have start data that consists of both their own input data and shared SSB data. To make it easier to share and reuse data, this document suggests that the steady states of data are stable and follow from a **data set** satisfying given criteria, i.e. a steady state of data does not depend on which processes the data set is used in. A data set that has been granted status as (satisfying the

[1] GSBPM v5.1 - Generic Statistical Business Process Model - UNECE Statswiki

[2] Statistics Act Section 5 (2) The statistics shall be relevant, accurate, timely, punctual, accessible and clear, comparable and coherent. https://www.ssb.no/en/omssb/ssbs-virksomhet/styringsdokumenter/statistikkloven

requirements of) processed data[3] will continue to be processed data even if the data set is used as start data for another statistic.

In addition to the steady states, some important concepts such as statistical product and quality indicator are defined, which are closely linked to steady states and statistical production. The document describes which metadata are the most relevant for each steady state and gives instructions for when the data owner's variable name should be used and when SSB-defined variable names should be used.

1.1. Mandatory steady states

This document provides general descriptions of all the steady states of data that are part of statistical production in SSB. At the same time, it is determined that the steady states source data, processed data, statistics and output data are mandatory. This means that those responsible for a given statistic is also responsible that the associated and mandatory data sets are produced, documented[4], long-term stored and made available in accordance with the steady state descriptions in this document. In this connection, it is important to emphasize that "documented and long-term storage", especially for output data, does not mean that all documentation must necessarily be stored on Dapla[5]. For example, output data for assignment financed activities can be documented in SSB’s document archive. This output data does not then also need to be documented on Dapla.

It is recommended that the responsible for the statistics regularly assess whether there is a demand and need also for input data. Particularly for data sets that contain personal data, it may be appropriate to document and long-term store input data since this is the first steady state that contains pseudonymised data.

[3] See section 2.3, variable is calculated, and accuracy improved.

[4] The data sets will be documented in a separate solution. At the time of writing (2024), work is being done to implement such a solution in Dapla.

[5] Dapla is Statistics Norway’s cloud data platform.

2. Steady states of data

2.1. Source data

Source data is data stored as it was delivered to SSB from the data owner, i.e. in the data owner's data format and data model together with information about the time and delivery order. Source data may contain personally identifiable information and other sensitive information. They can also be unstructured or structured. Source data is part of the statistics' documentation and can be a necessary source for research and for creating new statistics. Unless the source data is available, it will not be possible to verify SSB's statistics. The original source data will often be compressed and encrypted for storage after relevant parts of the data set have been transformed into input data.

Source data, especially those collected via API, are checked for any data errors so that the data owner can be contacted. Relevant controls may be frequency (data do not arrive/come less frequently than expected), scope (fewer variables than assumed), volume (far fewer observations than expected) and content (e.g. proportion of zero values much larger than expected or the values obviously not logical).

Examples of source data are:

- Basic data: Data that changes relatively rarely and is often used in many different contexts. Population registers for people (Population register), businesses (Business register) and places (Land register) are examples of basic data.
- Transaction data: Data that reflects that an exchange of information or an event has taken place. Bank transactions, receipt data and payroll data are examples of transaction data.
- Administrative data: Data compiled, primarily by public authorities, for wider use, such as reporting or tax purposes. Administrative data can, like transactional data, reflect that an exchange or event has taken place. Customs declarations and the "a-ordning"[6] are examples of administrative data.
- Statistical data: Data collected with the main purpose of producing statistics. Surveys/ questionnaires are used where we have not established automated, electronic reporting or where data with the necessary frequency, timeliness, level of detail or delimitation that is necessary to produce statistics is not available. They are also used to comply with European regulations, which in some cases require data collection with a given questionnaire design.
- Aggregated data and reports: Data that the data owner has processed before SSB receives them. This means that parts of the production process are carried out outside of SSB, and that SSB often does not have insight into all the algorithms or processes used in the production of the data.

Data created by the data owner can often be in the form of a document (for example, a cashier receipt). To minimize the possibility of errors in data processing by the data owner, we should strive to capture this data

- as close as possible to the source where they occur
- as unprocessed as possible7 (copy of original document)
- as close as possible to the time when the data occurs.

[6] https://www.skatteetaten.no/en/business-and-organisation/employer/the-a-melding/about-the-a-ordning/about-a-ordningen/

[7] Unprocessed means that the data has not been subjected to extensive linkage, transformations or aggregations. Data where the data owner has corrected errors and made simple extracts for SSB will still be considered unprocessed in this context.

In cases where there is a need for data beyond the original document, it may be necessary to retrieve data from several sources or retrieve data that has already been processed by the data owner.

In cases where SSB receives processed data, such as administrative data from registers, it is important to cooperate with the data owners so that the data owners themselves take responsibility for quality assurance and error correction before SSB receives the data. Agreements have been made on delivery of data to SSB and cooperation on quality with several data owners.

Relevant metadata

Information on the data set level such as data owner, which area the data covers, and time information are relevant metadata for source data. Metadata about individual variables results from, and is limited to, the information that the data owner themself submits.

2.2. Input data

Input data is mainly source data that has been transformed into SSB's standard storage format[8] and possibly restructured[9]. Input data is thus data collected from external parties and adapted to the purpose of the statistics. These external data, together with shared data from others in SSB, will constitute a statistics' start data. However, some macro statistics will not have input, only start data consisting of shared data from others in SSB. Other macro statistics will be based on a mix of input (own collected data) and shared data from others.

Input data has been processed through various stages of data minimization, pseudonymisation, recoding, restructuring and reformatting. The input data must not contain personally identifiable information and must be structured and stored in standardized formats and systems.

For many statistics, there will also be validation[10] associated with input data. This means that variable names and contents in input data are unchanged from how they are in source data, except that

- The data are minimized so that only variables that are necessary in the further production process are included.
- Directly identifying variables (e.g. national identity number) are pseudonymized.
- Standard classifications and code lists are used wherever possible (e.g. Standard Industrial Classification, gender).
- The data can be restructured and adapted to the statistical purpose.
- Character set, date format, address, etc. have been changed into SSB's standard format.

Even if the data is minimized, more variables will often be included in the input data than those that appear in the final statistics. Reasons for this include

- we want to be able to carry out control, error correction, imputation, weighting, analysis and similar operations to improve quality, create statistics and other estimates (e.g. of uncertainty)
- there may be a need to share with other statistics and for research
- there is a need to further develop and manage code effectively in line with increased knowledge and changes in data sources.

[8] For example, character set UTF8, date format YYYY-MM-DD.

[9] It can often be appropriate to restructure the data according to the principles of "Tidy Data" (each observation is a row, each variable is a column, each cell is a value), see https://vita.had.co.nz/papers/tidy-data.pdf

[10] Machine validation where validation reports are sent back to the respondent who then resubmits the data.

The Statistics Act and the General Data Protection Regulation (GDPR) stipulate that all data collected and processed shall: "... be adequate, relevant and limited to what is necessary for the purposes for which they are processed (data minimisation)". To satisfy the requirement for adequate and relevant data, enough variables must be collected so that we can carry out the necessary checks and analyses so that we get good quality in our products. This will, as with other types of data, most often involve the need for more variables than those directly used in calculations to create the statistics. At the same time, it must be ensured that we do not collect data that is not necessary for the purpose for which the data will be used.

Relevant metadata

Basically, the same metadata as for source data, but if input data is to be shared, there may be a need for more or different metadata. In addition, it may be relevant to supplement with metadata that provides information about the processing carried out in SSB.

2.3. Processed data

Processed data results from input data when

- the input data is integrated with other data and new variables are calculated
- accuracy is improved through checking the validity of the data and corrective measures in the form of, for example, filtering, editing or imputation
- metadata with definitions of variables has been added.

As a rule, no aggregations have been made in processed data. This means that processed data mainly reflect individual observations, in the same way as source data. If the source data received by SSB are aggregates, the processed data will often be aggregated at the same level as the source data. In processed data sets, SSB stands for variable names and variable definitions.

It is a goal that processed data is, to the greatest extent possible, created automatically and using algorithms, without manual operations. The changes made to data must in any case be traceable and documented in such a way that the statistics are verifiable.

In processed data sets, specific choices have been made on how errors and deficiencies in the input data are to be handled, and how processed variables are established. The choices may be different for different statistics. When processed data is to be reused for purposes other than what the data was originally processed for, a concrete (new) assessment must therefore be made as to whether it is necessary to supplement with new data or make other changes to the data.

Note that no changes to input data should be made because of processing. This means that input data that has been corrected for errors and omissions during processing, is stored as processed data, and not as input data.

Data may also be processed for sharing purposes e.g. National Education Database, Social Security Database. Such data must satisfy the requirements in this document and are otherwise subject to the same quality requirements as processed data for statistical purposes[11].

[11] European statistics Code of Practice https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142

Processed data from other statistics

In addition to data that is collected directly for a given statistic, there will often be a need for various types of processed data from other statistics[12] and supporting data in the production process, for example

- population information from one of the population registers
- sampling information
- processed data from other statistics
- weights calculated from other statistics.

Even if these data have already been processed through population management or other statistical processes, there will often be a need for adaptation and transformation when they are to be integrated with other data in a given statistical production, typically in the process phase for the given statistics. Macro statistics are an example where several sources and/or processed data are put together to form a statistic.

Statisticians who use processed data from others are responsible for taking care of information about which data sets have been used so that the production can be verified.

Versions

Depending on whether the processing involves manual operations, or whether processed data can be recreated in its entirety from source data by running programs, there will be a need to create and store different versions of processed data. These must be named and versioned according to the "Naming standard and versioning of data sets on Statistics Norway's Data Platform (Dapla)"[13].

Relevant metadata

For processed data, these are definitions of variables, i.e. a description of each individual variable and how it has been calculated. In addition, documentation is needed of which accuracy-improving measures have been carried out and why they have been done.

2.4. Statistics

Statistics are data that fall under the definition of statistics in the Statistics Act[14]: "statistics: quantitative information about a group or a phenomenon, and which is obtained by aggregating and processing information about the different units of the group or a sample of these units, or through systematic observation of the phenomenon." (Statistics Act Section 3. Definitions a))

Statistics will usually be aggregated data or estimated sizes. The individual variables in statistics are defined by SSB and are named by SSB in line with current practice. These do not necessarily coincide with the variable names and definitions in processed data.

Sometimes the statistics follow the processed data by simple aggregations (unadjusted statistics). At other times, statistical methods will be used to estimate numbers based on sample data and indices, calendar-adjusted numbers, seasonally adjusted numbers and trend series (adjusted statistics) will be calculated in addition.

[12] Processed data from other statistics can come both from other statistics in SSB and from external sources such as Eurostat.

[13] Internal document (Navnestandard og versjonering av datasett på Dapla) only available in Norwegian

[14] https://www.ssb.no/en/omssb/ssbs-virksomhet/styringsdokumenter/statistikkloven

Statistics and processed data are shared internally by being included in other statistics' start data, especially those of macro statistics, and may then contain confidential and detailed data that are not published.

Relevant metadata

Variable definitions – that is, a description of each individual variable and how it is calculated. In addition, the methods and programs/code used to produce the statistics must be documented.

2.5. Output data

Output data are statistics where the requirements for confidentiality are taken care of. It is this steady state of data that is published by SSB itself or delivered to others. Rules and mechanisms for the release of, for example, source data or processed data fall outside the scope of this note.

The transformation from statistics to output data is characterized by the fact that

- confidential data is suppressed by filling those cells with a symbol (:), aggregation or other methods that safeguard confidentiality requirements
- the quality is considered good enough for publication in accordance with quality requirements in official statistics[15]
- any requirements for joint publication have been taken care of. This could, for example, be the simultaneous publication of unadjusted, calendar and seasonally adjusted figures16

Examples of output data are:

- tables in StatBank Norway[17]
- commissioned statistics[18]
- international reporting

Versions

If published figures are revised after publication, all versions must be preserved.

Relevant metadata

Current metadata for output data is the same as for statistics. In addition, any programs or code which have been used to create the product must be documented. As mentioned in 1.1 Mandatory steady states, documentation of all output data will not necessarily be stored on Dapla, but for example in SSB’s document archive.

2.6. Confidentiality

Note that many confidentiality methods take processed data as a starting point, and not statistics. Many statistical tables are often produced at the same time, so that a coordinated confidentiality is ensured. The result of such a method is confidential output data and unprotected statistical data. There is then no reason to produce output data in a separate step.

[15] https://www.ssb.no/en/omssb/kvalitet-i-offisiell-statistikk
[16] https://www.ssb.no/en/omssb/kvalitet-i-offisiell-statistikk
[17] https://www.ssb.no/en/statbank
[18] https://www.ssb.no/en/data-til-forskning/tabelloppdrag

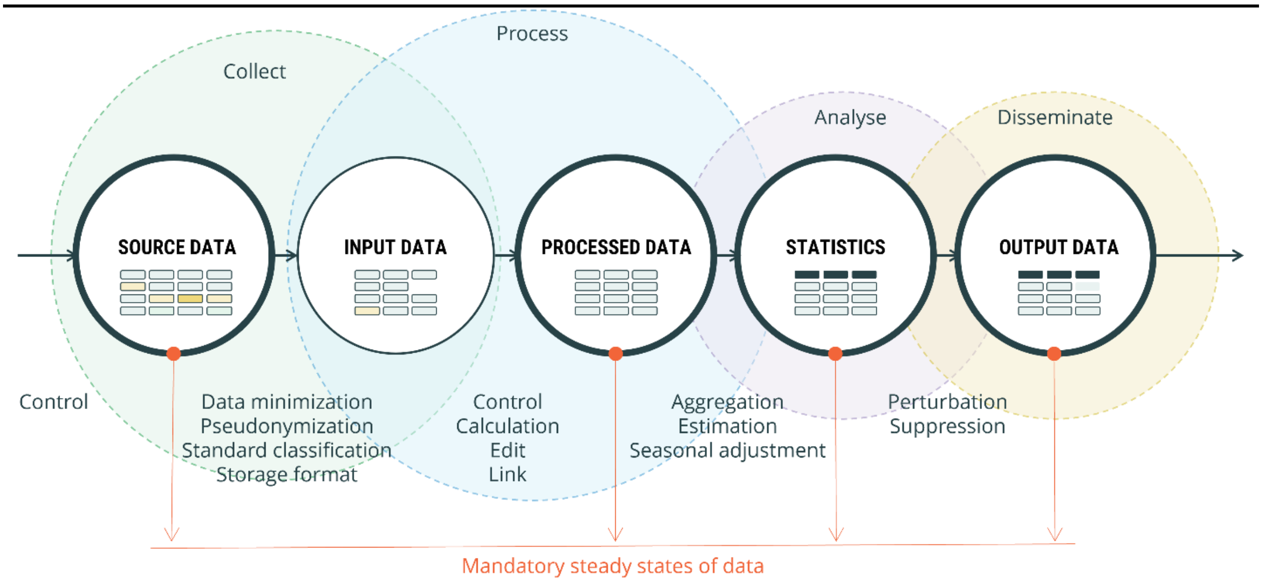
2.7. Long-term storage

When it comes to archive-worthy files/data sets and the delivery/deposit of these in the National Archives, this is still an open issue in SSB. In that context, both delivery/deposit[19], and the possibility that SSB could store the archive material[20] itself, is discussed. It is the steady state of processed data that is considered worthy of archival.

2.8. Steady states and the main transformations

Figure 2.1 summarizes the various steady states of data and gives examples of the main transformations that are performed to bring the data from one steady state to another.

Figure 2.1 Steady states of data and examples of main transformations



The steady states source data, processed data, statistics and output data are mandatory, which means that associated data sets are produced, documented, long-term stored and made available in accordance with their steady state descriptions.

However, some general comments are needed to nuance what are mandatory steady states of data. In "Naming standard and versioning of data sets on Statistics Norway's Data Platform (Dapla)"[21] we use the terms data products[22] and statistical products[23]. The documentation requirements for these will vary. For what we refer to as data products, for example the Population register and Business register, only source data and processed data will be mandatory steady states of data. The data

[19] Paragraph 16 of the Archives Act: Upon delivery, ownership of the archive passes to the recipient institution. When depositing, the depositor and later his heirs have the right of ownership to the archive. Ownership still passes to the receiving institution when the succession to the depositor is broken, or when a hundred years have passed since the deposit.
[20] Paragraph 10 of the Archives Act: State archives must be handed over to the National Archives in accordance with the provisions laid down in accordance this act. The National Archivist can nevertheless give consent for state archives to be handed over to institutions outside the National Archives, or for them to continue to be kept by the archive-creating body.
[21] An internal standard for naming and versioning of datasets.
[22] Not all data in SSB can be linked directly to a statistic in the Statistics Register. Data is processed for other uses and purposes, e.g. preparation of data for research, processing of data to be included in other statistics, and data included in population registers. These are called data products.
[23] All SSB's previous and current statistical products are included in the Statistics Register. Before publication on ssb.no, all statistical products must be registered in the Statistics Register with information on e.g. the name of the statistics, subject area, owner and publication time. In addition, the statistics are assigned a short name.

products are mainly microdata in the form of administrative/population registers that are processed for secondary use, i.e. they are included in other statistical products and/or research.

For statistical products that collect their own data (from, for example, registers or via surveys), source data, processed data, statistics and output data are mandatory.

For macro statistics such as national accounts, environmental accounts, etc., in many cases only processed data, statistics and output data are mandatory, in that they only use processed data (possibly statistics/output data) and do not collect source data (input data).

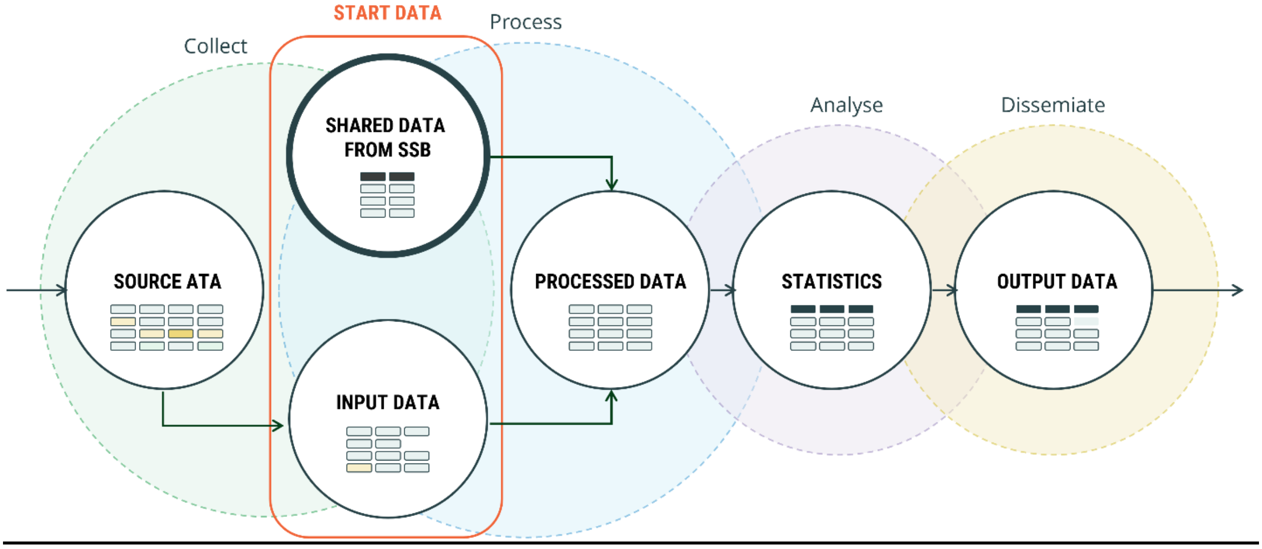
This means that only processed data is mandatory for all production, while input data is always optional. However, all responsible for statistics should consider whether input data should also be stored. Input data can be an important steady state when it comes to sharing because it is the first steady state with pseudonymized data. It is also the steady state that shows what data minimization has been done to the source data.

Recommended quality indicators must be established and followed up for all steady states of data. One must therefore be able to retrieve data that can be used to calculate quality indicators from all stored steady states of data. Examples of quality indicators are proportion of invalid data, unit non-response and editing proportion.

The figure below shows the relationship between start data and the steady states of data. As previously mentioned, start data is linked to process, and is the data with which one starts the processing. "Shared data from SSB" is data that is shared by others in SSB, be it their input data, processed data, statistics or output data.

For some statistics, start data will only consist of their own input data, others will have start data that only consists of shared data from others in SSB, while some will have start data consisting of both their own input data and shared SSB data.

Figure 2.2 Start data and the steady states of data



3. Other definitions

3.1. Statistical product

A statistical product is statistics that, alone or together with other statistics, describe a group or a phenomenon and are communicated in such a way that they make sense to the users. Different target groups have different needs. For some users this may mean a table to quickly look up the latest figures, for other users a graph showing both the latest figures and the development over time may provide more insight. For exploration, an interactive model may be most useful to users. Often, various forms of textual description and analysis will be part of a statistical product.

In the "Naming standard and versioning of data sets on Statistics Norway's Data Platform (Dapla)"[24], the statistical product is described as follows: All Statistics Norway's previous and current statistical products are included in the Statistics Register. Before publication on ssb.no, all statistical products must be registered in the Statistics Register with information on for example the name of the statistics, subject area, owner and publication time. In addition, the statistics are assigned a short name.

3.2. Process data

Process data is information about events on the data and occurs during the production process in all steady states of data. All events and changes in data are logged, with information on: When (time of the event / change), who changed the data or triggered the event, what (new, changed, deleted), how (manual, automatic) and preferably also why (recontact, professional judgment etc). Process data is often included, together with metadata and data, as a basis for creating quality indicators. An example of this is the quality indicator for editing proportion, divided into manual and automatic editing.

3.3. Quality indicator

Quality indicators are numerical quantities or statistics that help to illuminate the quality of data. Quality indicators can be simple counts, or the result of calculations or analyses. Different parts of the production process can have different quality indicators. Such indicators should be established, calculated and followed up for all steady states of data. An overview of recommended quality indicators has been prepared in SSB. There is a need to continue working with quality indicators in SSB. There are few statistics that have implemented such indicators in production, and there is a need to gather experience with the use of the recommended indicators.

Examples of quality indicators are:

- Unit non-response: The ratio between the number of units that are missing and the number of units that should have been included in the data set. Reported to the Ministry of Finance at an aggregated level.
- Editing proportion: The proportion of values edited in relation to the number of possible values.
- Potential timeliness: Number of days from when the statistics are ready for publication until they are published. The purpose of this indicator is to be able to give an indication of whether there is room for improvement in the timeliness of the statistics.

[24] Internal document (Navnestandard og versjonering av datasett på Dapla) only available in Norwegian

3.4. Pseudonymization

The Norwegian Data Protection Authority defines pseudonymisation as the de-identification of personal data so that they cannot be linked to a specific person without the use of additional data (for example a connection key) which is stored separately and sufficiently securely. Pseudonymised personal data is not anonymous.