

Efficient use of data in a new statistical production process

Susanne Gullberg Brannstrom, Statistics Sweden, susanne.gullbergbrannstrom@scb.se

Abstract.

To make statistics relevant and reliable in a world of decreasing response rates and expanding access to administrative data the NSI's in our countries has for some time changed the statistical production from survey collections to use more administrative data. The work has often been focused on converting the administrative data to statistical information and then use it in the statistical production. Doing so causes underutilization of the original data source, since you have one, or a few, statistical subject matters in mind when you transform the administrative information to statistical information. To be able to use all the information in a data source it is important that it is stored in its origin but easy to access and use.

Stat Sweden have made a strategic decision to use data that is already in house at first hand when making changes in or producing new statistics. This requires a statistical production process that allows to search and use data already collected for other purposes. Stat Sweden has introduced this new way of work and has built a technical environment to support the new production process that is automatized as far as possible.

This abstract focus on the aim to use the full capacity of the data collected for a broad variety of statistical purposes.

The outcome of the new way of work will give a lot of benefits. For example that different statistics use the same sources which gives harmonisation and consistency. The new technical environment will give faster statistical production process due to automatization. We will get more use of administrative data since it is easy to search and find inhouse. It will also give traceability in who is using what data and only collect a data source once.

Keywords: Efficient use, administrative data, Re-use, Automatization, Statistical process

1. Introduction

To make statistics relevant and reliable in a world of decreasing response rates and expanding access to administrative data the NSI's in our countries has for some time changed the statistical production from survey collections to use more administrative data. The work has often been focused on converting the administrative data to statistical information and then use it in the statistical production. It is an important work that is under constant development. But – only focusing on that task causes underutilization of the original data source, since you have one, or a few, statistical subject matters in mind when you transform the administrative information to statistical information. To be able to use all the information in a data source it is important that it is kept in its origin but easy to access and use. Stat Sweden has made a strategic decision to use data that is already in house at first hand when making changes in, or producing new, statistics. This requires a statistical production process that allows to search and use data already collected for other purposes. Stat Sweden has introduced this new way of work and has built a technical environment to support the new production process that is automatized as far as possible.

The overall aim is to reduce response burden, reduce production cost and modernize the production. This paper focuses on using the full capacity of the data collected for a broad variety of statistical purposes.

2. Benefits of a new way to work

The aim is to make full use of the data collected and, during the process of data refinement, enable a broad range of statistical purposes. When society changes rapidly, there is also a need to quickly produce new statistics that shed light on new phenomena and new populations. One example is Eurostat's urgent need to monitor population movements from Ukraine following the outbreak of the war. With a new way of working and new technical solutions, several additional advantages will also arise for Statistics Sweden. We gain the ability to collect more administrative data from external actors' systems through

machine-to-machine solutions instead of collection processes that require sending and receiving data manually.

We can make data searchable and subscribable internally for use across all statistical production, rather than only for a specific product as was often the case previously. That gives us the opportunity to reuse already collected and processed data to a much greater extent than today, so we only need to collect data once.

With the new IT-tools we achieve cost-effective and automated data management with reduced manual handling. We can manage our data securely and avoid duplicate storage and due to implementing „Archiving by design” we can preserve, archive, and dispose of our data efficiently and automatically throughout the entire production process.

With a shared production environment instead of customized production systems for each statistical product, we gain an infrastructure and a way of working that allows employees themselves to handle changes affecting data collection and statistical production—without requiring IT to rebuild individual production systems just to accommodate, for example, a delivery arriving in a new format.

3. Traditional statistical production

Previously, almost all data collection was carried out within each individual statistical product, except some larger administrative data, like VAT and Standardised Accounting Statements. During the design phase, it was decided whether direct data collection would be conducted and whether certain administrative data should be gathered to complement the collected responses. This meant that the information collected was tied to each specific product and, although it could be shared with others, no one could easily know that the information existed at SCB unless they were familiar with the statistics and

therefore knew that the data was being collected and could ask the responsible person for access.

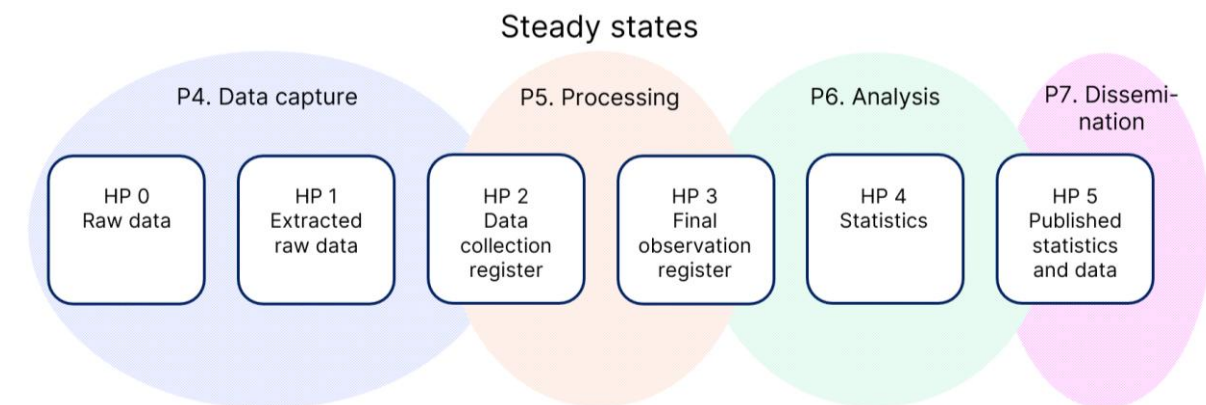
In most cases, people were unaware that someone else was already collecting the same information. Instead, they contacted the data providers themselves and ordered the data independently. In several cases, this resulted in the same data being collected multiple times in different parts of Statistics Sweden without anyone knowing about each other's efforts. It also meant that each statistical product had to handle administrative input data and prepare it for the specific statistics they were responsible for producing. While this work can certainly be enriching and enjoyable, it is not particularly efficient when several people are doing more or less the same thing every month, quarter, or year, depending on the periodicity of the statistics.

4. The future way to produce statistics

Statistics Sweden have made a strategic decision to use data that is already in house at first hand when making changes in or producing new statistics. This requires a statistical production process that allows to search and use data already collected for other purposes. The in data is presented to be easy to understand and use and has been processed to an extent so it benefits as many users as possible regardless statistical area to give consistency between different outputs. Statistics Sweden has introduced this new way of work and has built a technical environment to support the new production process that is automatized as far as possible.

The technical solutions being developed have a core of Steady states or handover points (HP). Steady states are points in the statistical production process with well-defined and metadata-described data sets. In these points, and in these points only, data can be (re)used for data exchange, within and between statistical products, in an, as far as possible, automated production process for the development of statistics.

Figure 1: Steady states of data in the statistical production process



Within a statistical production process for a specific statistical product, the various steps have always existed, but they have often been referred to as data under processing until a final observation register in steady state 3. The final observation register is the register designed for a statistical product based on what one wants, or is required, to report for each product. It is the final observation register that has been documented regarding variable content, quality, and production process, and it is the register that is shared with others. The final observation registers are used for many purposes, in addition to serving as the basis for published statistics for a specific statistical product. Internally, they are used as input for other statistics or analyses, while externally they are used for assignments, research, etc.

In the new production environment and according to the new way of working, it is possible to define and describe metadata for data to be shared with each other in all steady states throughout the process. One can view it as data being refined toward statistical information at each step. Although it is possible to share data in every steady state, there is no obligation if the data is not relevant to anyone else.

Steady state 0 refers to input data in a file that has been sent to Statistics Sweden or data held by a provider, such as another authority, which may be available via an API that one connects to. The data formats can vary greatly.

In steady state 1, data from steady state 0 has been imported into Statistics Sweden through a tool that reads the file and, in the same step, verifies the file format agreed upon with the provider. The content is documented in a newly developed metadata system, where classifications and code lists are also documented, so that the data at the time of import can be validated against classifications. Steady states 0 and 1 therefore mainly concern the technical ability to import data to Statistics Sweden. Further review and processing of the data occur in the next step.

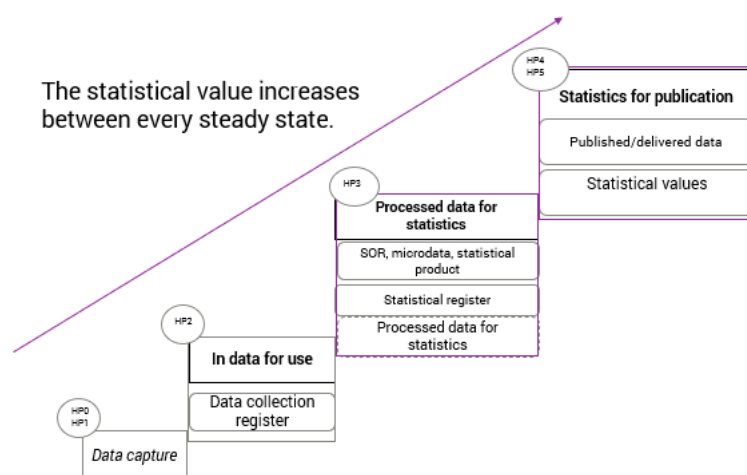
For the in data to be consumable by everyone at Statistics Sweden, it needs to be prepared in consumable formats and structured into logical datasets. The data also needs standardized formats and established identifiers so that datasets can be easily linked. This is done in steady state 2. Additionally, validation checks of cell values can be performed, and deviations can be flagged. In steady state 2, the data is also described at variable level in the metadata system, together with documentation about the original purpose of the data. The purpose of the documentation is to provide users with sufficient information to determine whether it is a data source suitable for their statistical production. It is therefore in steady state 2 that anyone at Statistics Sweden should be able to view metadata about the in data and thereby know what data exists in house Statistics Sweden.

The statistical process begins after steady state 2. Here, the datasets that form the basis for one or several statistical products are created with necessary variables and structures from already existing datasets in steady states.

When re-designing or creating new statistics, data from all datasets in every steady state can be used. Although data from steady state 2 and 3 are most common. A statistics producer uses a search tool to see what data exists in the

different steady states and select the variables needed from each dataset and create the datasets needed for their statistical purposes. They then expose the relevant datasets created in steady state 3 to be used by others. In steady state 3, several datasets may be made available. It can be anything from composite datasets that are to form the basis for further processing within a statistical area, to a final observation register for a specific statistical product. To be able to use data from different datasets in different stages gives the opportunity to use information from in data in other ways or in other combinations, to be able to develop new statistics to reflect new phenomena in society. When only using data that has been designed for a specific statistical purpose in mind some, or a lot, of the flexibility in indata have got lost.

Figure 2: Data management in a refinement process through steady states



5. Appendix: List of references

Andersson, H., Ohlsson, M., Strandell, G., Wallenberg, C. Guide för Statistikproduktion 4.0 (SCB 2025)

Andersson, H, Erikson, J. Architecture for multi-source statistics (UNECE 2026)

European Archives Group: Archiving by design (European Commission 2023)
https://commission.europa.eu/document/download/e3cf4d38-ee41-42f9-8994-f6589ffad458_en?filename=Whitepaper%20AbD_en.pdf

Support for statistical production (SCB) [SPS](#)