Architecture for multi-source statistics

Henrik Andersson, Statistics Sweden, henrik.andersson@scb.se

Johan Erikson, Statistics Sweden, johan.erikson@scb.se

Abstract.

A modernized statistical production will in nature be multi-sourced. Administrative registers and other digital data sources will be combined with traditional or complementary sample surveys to create qualitative statistical output while minimizing response burden. Furthermore, many new data sources will be useful for several statistical outputs. This situation calls for an architecture that gives the statistical office control over its data and data quality, and how data flows both inside the office and to and from other agents in various data eco systems. Over the last few years, Statistics Sweden has started implementing such an architecture. This paper describes the fundamental concepts in that architecture, and how we have used various standards in building it. The paper describes our data architecture using "steady states" of data throughout the production process, how data management, data access and data sharing is set up to allow for maximum re-use. An example of how this is used in practice at the moment is also presented.

Keywords: multi-source statistics, architecture, re-design

1. Multi-source statistics

With multi source statistics, we mean a statistical production process where statistical products are produced from the combined use of several data sources, (administrative registers, other types of digital data sources and traditional or complementary sample surveys with direct data collection). At Statistics Sweden we have called this movement “Statistical Production 4.0” (SP 4.0 in the rest of this paper) which is a synonym for multi-source statistics. When data sources are used in the production of several statistical products, there will be many and sometimes complex data flows within the office.

2. Designing multi-source statistics

In order to support the design and implementation of SP 4.0 in re-design of products or entire subject fields, Statistics Sweden has developed a user guideline for implementing multi-source statistics (Andersson et al 2025). The work can be divided into three main steps: In step 1 you compile the user needs of and the quality demands put on statistics and data. There may be legal demands (including EU regulations) that affect the different quality dimensions. In step 2 you conduct an analysis of the possible data sources to find if you can use use existing data sources more than today and if are there new or previously unused data sources that could be used as well. In step 3 you analyse the methodological challenges. Issues may include selection bias, coverage issues, measurement errors, definition of variables, matching and integration, and for large data sources computation possibilities.

3. The cornerstones of our architecture

Multi source statistics with many internal data flows calls for a data centric architecture that puts data itself in the centre. In our architecture we address this with Context management and Data management. A more detailed description of the contents here can be found in Andersson & Erikson (2026). With Context management we mean management of “hubs” on which we connect other characteristics that steer how the production process is done. With Data management we include both data and metadata.

4. Context - a hub for work, data and design

One of our architectural principles is that all data handling and work on statistical activities must be done within a context. The term context can be used for all types of work related to statistical production, both production of statistical products and registers and more temporary work like projects. The context also acts as a hub on which we connect other characteristics. Some of the most important of those are a) who owns and produces the context (a context is connected to a section in our line organisation, and the head of the section automatically becomes the information owner for data produced in the context); b)the production cycle for recurring production to define production rounds and reference times; c) who in the staff that work within the context (we use role based authorities); d)what the context produces (statistical products, data sets in steady states and other artefacts); e) what data from other contexts that the context needs for its production; f) what information classification the data the context handles and produces has and g) how the data lifecycle is to be treated (preservation, archiving or deletion after a certain amount of time, according to archiving by design, European Archives Group 2023).

5. Steady states of data

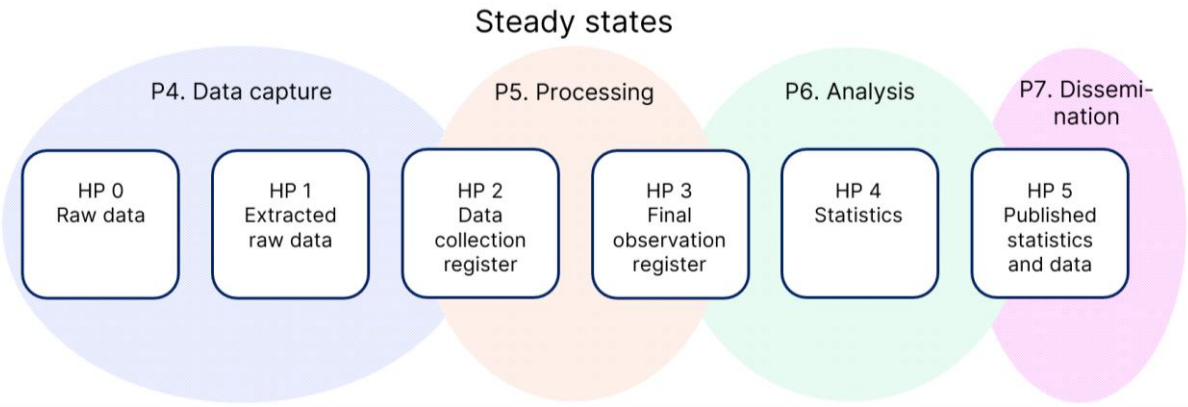
For our data management, we have defined Steady states of data (the abbreviation HP comes from the equivalent term in Swedish). A steady state is a standardised point of refinement in the production process, where the data is described by metadata and quality descriptions. Datasets in steady states are made available for internal or external users

The steady states used at Statistics Sweden are the following (see also figure 1):

- HP 0 Raw data: Captured data in the format given by the data provider.
- HP 1 Extracted raw data: Raw data that has been unpacked and put in a format adapted to Statistics Sweden.
- HP 2 Data collection register: A compiled register of one data source that can be used in further refinement in the production process. Data is not yet adapted or transformed for statistical purposes.

- HP 3 Final observation register: Despite the name, this steady state incorporates both the final observation registers connected to specific statistical aggregates **and** intermediate statistical registers created in the transformation from processed observation registers to final observation registers.
- HP 4 Statistics: Statistical aggregates before treatment of confidentiality.
- HP 5 Published statistics and data: Aggregates after treatment of confidentiality that have been published or delivered to a customer/user.

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



The content of datasets in steady states must be described with a data structure located in a central metadata system at Statistics Sweden.

6. Cataloguing steady state datasets in a dataset catalogue

The datasets produced in the steady states are all gathered in a common dataset catalogue, where the dataset contents, not the datapoints, can be accessed from all systems within Statistics Sweden. The location of the datapoints is also included and can therefore be addressed. A dataset can be produced several times and will thus be versioned.

7. Exchange of data between contexts

In order to use data from another context a provision agreement is set up between contexts where one context gets the right to use data sets from another context.

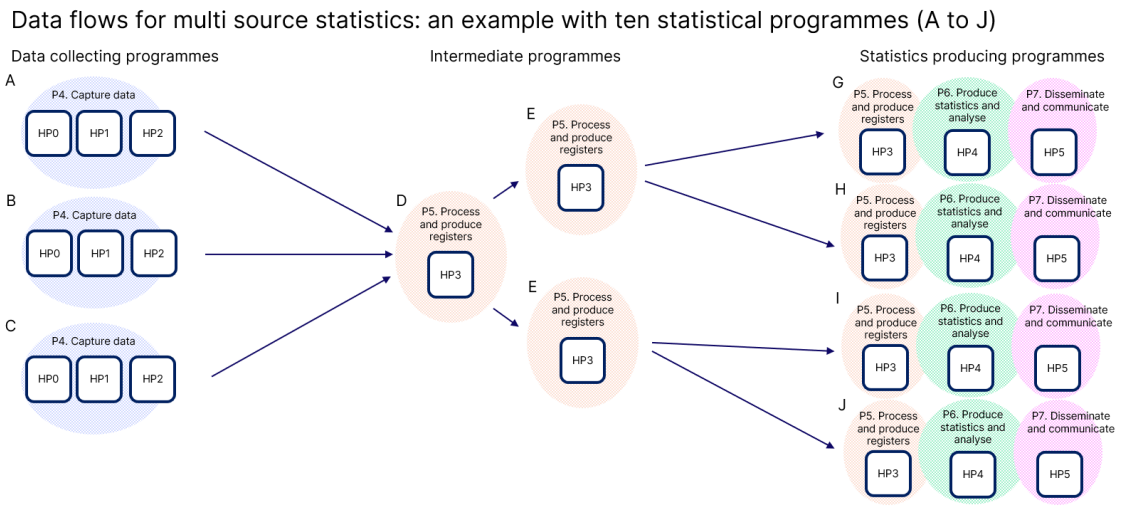
Anyone that has a role withing the consumer context can set up a contract, but the information owner of the producer context needs to approve the contract. Access to the data is given through the provision agreement via the dataset catalogue.

8. Data flows for producing multi-source statistics

In the redesign of a whole or a part of a statistical field, the first thing is to set up a production design for the whole field. This means that decide which data sources will be used, which statistical products will be produced and how the production flow will be throughout the whole field. We have three different types of contexts: Statistics producing programmes that produce output in the form of statistics or data for external users (data in steady states 3, 4 and 5), Data collecting programmes that collect data sources and arrange them in data collection registers (data in steady states 0, 1 and 2), and when necessary Intermediate programmes, that are used in complex statistical fields to take steps between data collection registers and final registers (data in steady state 3 only). There may be one or several intermediate programmes depending on how the design is set up.

Once you have set up the contexts you need, you can design the output of each context and the data sets in each relevant steady state. After that, you can set up provision agreements between contexts. An example of a design of multi-source statistics is shown in figure 2.

Figure 2: Data flows for multi-source statistics



9. An example: a new design for economic statistics

The design for the economic statistics is right now being re-drawn according to SP 4.0. New data sources, both administrative registers and new types of data sources like transaction data have emerged through the digitalisation of society. These sources can replace much of the traditional data collection and lower the response burden of businesses. Sometimes they have overlapping variables for the same reference time and need to be coherent when used together (Erikson et al. 2025). The new data sources will become gradually available over the next few years, so an overarching design is created that can be implemented as they become ready for use (Lennmalm et al, 2026). The proposed overarching design for the re-designed economic statistics fits well with the architecture described in this paper and figure 2 above. The overarching design divides the production into phases where each data source will be a data collecting programme, a number of intermediate programmes will be used to refine data and statistics producing programmes will produce primary and secondary statistics[1].

First, data collection and creation of data collection registers takes place. Each data source is treated separately, collected and prepared into data sets ready for consumption. At least some of the sources will be used both for economic statistics and in other statistical areas.

The intermediate stage includes matching and harmonising of data sources that have a common content. The matching will include a priority order implemented as a set of rules on which sources take precedence over others, and how data for some sources or variables should be recalculated to fit with others. A starting point could be to start with sources that are based on or connected to the standard chart of accounts. The core sources[2] are harmonised into an internally coherent profit and loss account and

[1] In Lennmalm et al (2026) the intermediate results are denoted as HP2 wheras they in this paper are denoted as HP3. This is because we changed how we denote those since the first paper was written.

[2] Digital Annual Reports (DIAR), Value Added Tax including OSS (VAT/OSS), Standardised Accounting Statement (SRU) and Monthly PAYE Return (AGI).

balance sheet. Other sources that can be linked to specific posts in the profit and loss account or balance sheet are adapted so that they are equal to or a logical part of those posts. Remaining direct data collection can also be designed to collect specifications or additions to the administrative data. After that, the sources are integrated into one single data layer with target variables The aim is to go from separate sources to a coherent coordinated data material, with only one value for each variable for each object for each reference period, and with common metadata. Following that, further processing converts observation variables into target variables. After that, a coherent transformation of microdata to fit the designed target population(s) is made. This involves adaptation of data to the target population (for example an annual master frame from the Statistical Business Register for annual statistics), treatment of non-response and coverage problems, derivation of statistical units and recalculation of data based on those, recalculation to the correct time period for units with a deviating observation time period and some additional micro data calculations based on model assumptions.

Finally, statistics are created from the micro data. The procedures can vary depending on whether sample data are included and which statistical output to create. In this phase, the re-design might come up with restructured or new output compared to the statistical products that are produced today. For secondary statistics (e.g. national accounts), there is no real difference compared to the situation today, statistics and data from the primary statistics is used to produce secondary statistics. The values that are used will be more coherent though.

Besides designing the different phases, the re-design project will also need to address how different frequencies are to be handled; both handling within each frequency (one design for annual statistics, one for quarterly statistics and one for monthly statistics, each based on what data sources are available to use for each frequency) and harmonisation between frequencies including adaptations in retrospect (for example adjustments of quarterly data when annual data is ready).

The steering group for the re-design project has decided to proceed with detailing a conceptual design in accordance with the ideas described here, with some priority tasks to take place already this year.

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)

Erikson, A-G., Lennartsson, D., Strandell, G., Svartengren, S. Ekstat 4.0 Searching for a new design of the economical-statistical system with access to new data sources (SCB 2025)

Lennmalm, A., Kain Wyatt, M., Landberg, O. Ekstat 4.0 Fortsättning – Produktionsflödet från data till statistik (SCB 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