# HOW STATISTICS SWEDEN USES THE DEPARTMENT OF DATA MANAGEMENT TO ENSURE ACCESS TO HIGH-QUALITY DATA

**Kristina Strandberg[1], Ulf Durnell[1]**

*[1]Statistics Sweden, Sweden*

## Abstract

In 2021 Statistics Sweden underwent a complete re-organization. At the wake of a pandemic a change of that magnitude could have been seen as a big risk, but it was the firm belief of management that some changes were crucial for Statistics Sweden to be successful in its mission – to keep supplying Sweden with accurate and reliable statistics. One of the core changes in the new organization was the establishment of a Department for Data Management – a department with the responsibility to secure and streamline the process of data management for the entire organization, and to ensure the continued access of high-quality data.

It might sound like a contradiction, but one of the biggest challenges facing statistical institutes today is the rapid growth of available digital data. From responders, there is pressure to submit data digitally, through machine-to-machine solutions. Swedish authorities also are tasked to make it easier for companies to respond. Digitalization of business systems and registers, constantly adds to the list of potential data sources, from both private actors and other government agencies. Add to that the possibility to use data from other administrative sources, such as mobile network data. Considerable development of data collection methods and data management processes is needed to stay on top of the digital transformation, and ultimately – to stay relevant in an ever-changing world.

During the first two years of operation, the work at the Department of Data Management has been concentrated to two things: developing automated and efficient processes for data collection and data management, and to enabling wider use of existing data. This involves creating technical solutions, as well as straightening out legal and data security related problems. Statistics Sweden now has standardized processes for Monitoring of Potential Data Sources, Approval of new Data Sources and a Process for Register Production. To support data management, a new IT platform with related metadata system has been designed and is taking form.

In conclusion, giving the Department of Data Management the mandate and responsibility to supply the organization with high quality data has helped Statistics Sweden to bring about several necessary changes in our processes for data collection and data management. In this paper, we present the new department, and highlight the important work that has been fundamental to achieve these changes.

**Keywords:** data management, data handling, data collection, coordination of data management

# 1. Introduction

This paper aims to show how Statistics Sweden uses its newly established Department of Data Management to achieve one of its strategic goals, to ensure access to high-quality data.

## 1.1 Vision and Strategy

Statistics Sweden's vision is to *Give Sweden Useful and Reliable Statistics*. While the vision describes the mission, the strategy (Statistics Sweden, 2022) serves as a road map to get there. Within the strategy, there are five overall goals: A) Useful Statistics and Data, B) Smart Data Capture and Data Handling, C) the Production of Statistics is Innovative, Efficient, and Secure, D) Staff at Statistics Sweden Develop Themselves and the Organization, and E) We Work Together. In this paper, we will focus on B) smart data capture and data handling, see figure 1.

Table 1: Key concepts of sub-goal B: Smart data capture and data handling

| Key Concepts |
|---|
| • National architecture – we have, together with data owners, established an overall architecture with common standards and processes for an efficient and quality-assures data access for statistical purposes0000<br>• Operational architecture (metadata, quality, security) - We have an operational architecture that supports flexible and quality-assured data management for all types of data sources. Concepts and definitions are coordinated, and all data has been classified with metadata. We have routines that ensure that the data is of high quality already at the source.<br>• What data to use – We use primarily data that is already available within the national architecture. Next, we use data that can be found in the data holder's business system. Direct data collection from individuals, households, enterprises, and organizations is the last option, and should be digital if possible. The response burden has been halved.<br>• Respondents trust – The respondents have confidence in Statistics Sweden and can easily provide information needed. |

## 1.2 The New Organization

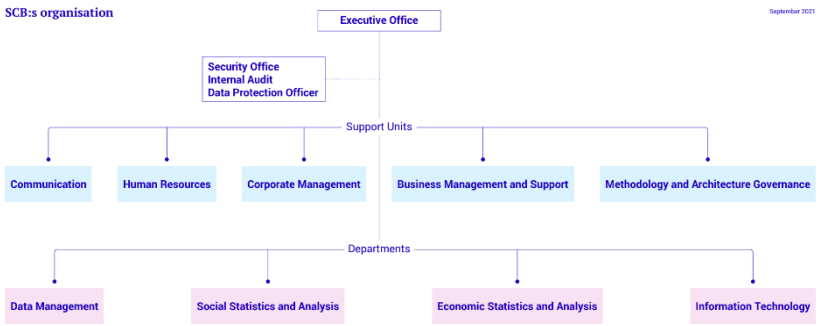
Digitalization and globalization are the driving forces behind the changing conditions for any producer of official statistics. From responders, there is pressure to submit data digitally, through machine-to-machine solutions. Swedish authorities also are tasked to make it easier for companies to respond. Digitalization of business systems and registers constantly adds to the list of potential data sources, from both private actors and other government agencies. Add to that the possibility to use data from other administrative sources, such as mobile network data. Considerable development of data collection methods and data management processes

is needed to stay on top of the digital transformation, and ultimately – to stay relevant in an ever-changing world.

In the wake of the latest big review of the strategy, Statistics Sweden's management saw that some changes were crucial for Statistics Sweden to be successful in its mission. As a result, Statistics Sweden underwent a complete re-organization in 2021 (fig. 1). One of the core changes in the new organization was the establishment of a Department for Data Management – a department with the responsibility to streamline the process of data management for the entire organization, and to ensure the continued access of high-quality data.

As important as the new data department was the structure of the subject matter departments. To better be able to find common ground between statistical products, regarding both design and technical solutions, all statistical production was concentrated in two subject matter departments. One department is centered around social statistics, while the other focuses on economic statistics. By centralizing rather than dividing products, Statistics Sweden is hoping to facilitate the development of common solutions. While design and production of statistics is very exciting, this paper will concentrate on the responsibilities of the Department of Data Management.

Figure 1: Organization at Statistics Sweden




## 2 Department of Data Management – Up and Running

Giving one department the responsibility – and mandate – to coordinate all data management has allowed Statistics Sweden to take a more cohesive approach to both day-to-day production and development within the field of data management. During the first year of the new organization, a lot of time and effort was invested into designing new processes for data

capture and data handling. This includes streamlining processes for register production, approval of new data sources and implementation of machine learning in statistical production.

### 2.1 The Prioritized Modes of Data Capture

One of the key concepts of smart data capture concerns what data to use. We use primarily data that is already available within the national architecture. Next, we use data that can be found in the data holder's business system. Direct data collection from individuals, households, enterprises, and organizations is the last option, and should be digital if possible. When this is completely implemented, the response burden has significantly decreased.

With this as the starting point, Statistics Sweden has adapted a list of preferred modes of data capture, table 2 (Statistics Sweden[1], 2023). The underlying theory is that cost per observation as well as respondent burden is lower in at the top of the list, and higher at the bottom of the list. As is noted in table 2, documentation in paramount to be able to use data wider. We need to know what data is available, which variables it contains, and all quality aspects of these variables. In short, we need a catalogue of data, with metadata and a complete description of all other characteristics important in the production of statistics.

Table 2: The Preferred Modes of Data Capture

| | |
|---|---|
| *1* | *Existing data*<br>Data that is already available at Statistics Sweden. Requires a complete catalogue of available data with its descriptive metadata. This data comes in two different shapes:<br>a) Final data – Existing data from other surveys at Statistics Sweden. This data is processed at Statistics Sweden for a specific survey.<br>b) Processed data – Existing data originating from other authorities, data originally held for purposes other than Statistics Sweden´s statistical production. |
| *2* | *Administrative data from the public sector*<br>a) Administrative information from authorities and other public actors, information about others. Existing data from other authorities, originally held for purposes other than Statistics Sweden's statistical production.<br>b) Administrative information from authorities and other public actors, information about their own activities. Existing data from a public actor, originally held for purposes other that Statistics Sweden's statistical production. |
| *3* | *Administrative data from private actors*<br>a) Administrative data from private data owners, data about others.<br>b) Administrative data from private owners, data about their own activities. |
| *4* | *Data collection from the internet*<br>a) API. We retrieve from a web page´s underlaying data source though an automatic data flow.<br>b) Web scraping. Data is downloaded from a web page via an automated solution.<br>c) Manual collection. Data is collected from a web page via manual browsing. |
| *5* | *Digital data collection*<br>a) Fully automated collection, through machine-to-machine (M2M).<br>b) Semi-automatic collection, sent via file or imported through Statistics Sweden´s data collection application (SIV) and completed manually in web forms.<br>c) Digital collection, a web form that is filled out manually and submitted via computer or via cell phone. |

6 *Direct data collection*
   a) Paper forms, questionnaires
   b) Interview
   c) Physical observation or visitor interview

## 2.2 The Process of Data Approval

To enable an efficient and quality assured evaluation of potential data sources, Statistics Sweden has developed a process for the approval of new data sources (Statistics Sweden[2], 2023).

It is the Department of Data Management that runs the approval process, but it is carried out in close cooperation with the subject departments. The approval process consists of four steps: 1. Initial Assessment of the Data Source, 2. Possibility to Access Data for Test, 3. Thorough Investigation of the Data Source, and finally 4. Decision to Collect Data for Production of Statistics.

Anybody within the organization can initiate the approval process by reporting the potential data source to the Function of New Data Sources (FUND). FUND is a permanent working group consisting of business developers, methodologists, and experts from the cognitive lab. FUND functions as gatekeepers of the process and takes potential data through the first step. Steps 2 and 3 are carried out by an experienced test team, with the same technical expertise as FUND, but put together specifically for each new potential data source. In the thorough investigation of the source, all aspects important to the statistical production process are evaluated, for example overall quality of data, legal aspects, and cost to access data. The final decision whether to take in data or not is made by the head of the Department of the Data Management, together with the head of subject matter department where data will be used. From start to finish, the entire process should not take longer than six months.

## 2.3 The Process of Register Production

An important task for the Register Unit at the Department of Data Management is to deliver on Statistics Sweden’s long-term goal, that all register production is carried out in the in-house register production tool, SCB-Fiber. Until this tool is fully in place, SQL or SAS scripts are used in its stead.

Furthermore, when SCB-Fiber fully developed, all datasets and their steady states will be registered in Statistics Sweden's Dataset Catalog and made available for all use within the organization. For the time being, database tables may be created to make steady states of registers and statistics available. Such database tables should be stored in a location separate from working tables, where users with assigned permission can retrieve the data that is

needed. In the Statistical Production Support, all steps in the process are thoroughly described (Statistics Sweden[4], 2023).

### 2.4 The Process for Machine Learning

Efforts to automate manual steps of the data handling processes in combination with the increasing access to big data set of high quality has led to the rise of machine learning (ML) methods. To ensure that Statistics Sweden keeps up with development regarding ML, the Department of Data Management put together a team to handle these questions specifically, Team ML. Team ML has been able to concentrate time and effort into exploring the use of ML-algorithms in data handling processes, such as editing, imputation and coding. To safe-guard quality in the final statistical product, Team ML has designed a process for implementing ML into the statistical production process (Statistics Sweden[3], 2023).

### 2.5 Challenges

In the first year directly following the re-organization, a lot of effort had to be put into helping everybody finding their new role and purpose. In some parts of the organization, for example in the subject matter departments, most things were the same as in the old organization. For the Department of Data Management on the other hand, everything was new. Since the department had to be built from the ground up, the strategy could be used as a blueprint to create a structure with the sole purpose of moving towards the strategic goals. Eventually, it became clear that this created a mis-match between the Department of Data Management and the subject matter departments. By working together to coordinate efforts connected to the key concepts in the strategy, difficulties have been mitigated.

Another challenge was the continuous coordination of rapid-moving process development. In the beginning, focus was to develop new processes supporting the new thinking about data collection and data handling. Once these processes were in place, adjustments had do be made for them to fit seamlessly together.

Challenges more specifically tied to data did also arise. In statistics, there is a clear way to describe the quality in data when it is connected to a specific statistical product. Now data quality had to be declared in general terms, to support wider use within Statistics Sweden. A whole new vocabulary to talk about data quality had to be “invented”. Statistics Sweden’s cognitive lab developed a model that describes the data generating process, Mätteknik 2.0 (Persson, 2022). Combined with the newly developed quality criteria for digital data (Japec & Jansson[1, 2], 2023), there is support to describe data quality in a way that enable its wider use.

It should also be mentioned that data quality in itself is an issue. Data is always collected for a specific purpose, and it is not necessarily true that is has good enough quality for other purposes. When data is held and used within the same organization (for example within Statistics Sweden), it is relatively easy to maintain a good dialogue regarding quality improvement and use for additional purposes. Issues arise when data is to be used between actors, in particular between public actors. This has sparked a conversation regarding the coordination of information needs and quality requirements within the public sector.

### 2.6 National Architecture

Ena is Sweden's digital infrastructure. It is a work of establishing a joint digital infrastructure for information exchange and Statistics Sweden is partcipaiting in this development. Basic data domains ensure that the data exchanged within the infrastructure is correct and accessible. Reusable building blocks make services more uniform, and enables faster and more efficient development.

Common European Data Spaces (CEDS) will make more data available for access and reuse. This will be done in a trustworthy and secure environment for the benefit of European businesses and citizens. It is important to investigate the coming CEDS including specific and clear aspects to be investigated such as data management, quality as well as opportunities, and risks for ESS and NSI associated with CEDS. It is also important for the ESS and NSI to consider the strategic and important impact of the CEDS in association with NSI potential roles and engagement in this area.

The Government has proposed Statistics Sweden be appointed as competent body as a result of the changes in the EU's Data Governance Act. The proposal means Statistics Sweden will support other authorities who have questions about privacy-promoting techniques, and secure processing environments so others can access their protected data.

Statistics Sweden has almost 300 years of experience in data management in its various forms. The strength is the operational process for the quality assurance of data for various purposes (where the current focus is the statistical purpose) where Statistics Sweden uses standardized processes (GSBPM), methods and tools. Statistics Sweden has a solid experience in managing data, data sharing and management in collaboration. The business is undergoing constant transformation as a result of changes in the outside world and the opportunities created by digital technology and the dataification of society. As quality-assured data becomes increasingly important, Statistics Sweden can contribute to an even greater societal benefit by handling data for purposes other than statistics in a role as a data steward.

### 2.7 Results and Conclusions

By creating a Department for Data Management, Statistics Sweden has concentrated the responsibility for (and mandate over) the processes involved in data management in one department. This enables coordination of these processes, - and of the data. While there have been challenges along the way, the new department has accomplished a lot, where new processes for register production, approval of new data sources and implementation of ML into statistical production should be mentioned.

In addition, the mandate to focus on data management has given the Department of Data Management time and resources to step on to the national scene to a bigger extent than before. Participation in national wide efforts such as Ena, the Swedish digital infrastructure, the development of national data domains that goes together with development of CEDS and implementing the role as a competent body as a result of changes in the EU's Data Governance Act. Statistics Sweden is a driving force, not only in the production of statistics, but also in the coordination and use of data.

## References


Lawn, M., & Nóvoa, A. (2013). The European Educational Space: New Fabrications. *Sisyphus – Journal of Education*, *1*(1), 11-17. https://doi.org/10.25749/sis.2827

Haldorson, M. (2022). SCB:s nya organization. Planering, genomförande och uppföljning av en ny organisation som infördes 1 september 2021. *Intern rapport*. SCB:s nya organisation

Japec, L. & Jansson, I. (2023), Kvalitetskriterier för statistik baserad på digitala data – bakgrund, *Government mandate, Uppdrag att främja delning och nyttigförande av data för smart statistik.* Uppdrag att främja delning och nyttiggörande av data för smart statistik - Regeringen.se

Japec, L. & Jansson, I. (2023), Kvalitetskriterier för statistik baserad på digitala data – vägledning, *Government mandate, Uppdrag att främja delning och nyttigförande av data för smart statistik.* Uppdrag att främja delning och nyttiggörande av data för smart statistik - Regeringen.se

Persson A. (2022), How to Develop the Cognitive Lab's Methods to Contribute to the Evaluation of New Data Sources, *paper to Statistics Sweden's Scientific Board*

Statistics Sweden (2022), SCB:s strategi, *Strategy documentation*, SCB:s strategi

Statistics Sweden[1] (2023), Dataförsörjningslistan, *Statistical Production Support*, Prioriteringslista för dataförsörjning.pdf

Statistics Sweden[2] (2023), Godkännandeprocessen för nya datakällor, *Statistical Production Support*, Godkännandeprocessen för nya datakällor.pdf

Statistis Sweden[3] (2023), Process för Maskininlärning, *Statistical Production Support*, SPS Övergripande process C

Statistis Sweden[4] (2023), Registergranskningsprocessen, *Statistical Production Support*, SPS Process 2.5


Please, rename the document according to the abstract number.

For example: 1234_Q2024