UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

ModernStats World Workshop **23 September 2024**

21-22 October 2024, Geneva Switzerland

A dataset catalogue as a tool for automated and metadata driven statistical production

Henrik Andersson, Patrik Wahlgren (*Statistics Sweden*)

henrik.andersson@scb.se patrik.wahlgren@scb.se

Abstract

To have better knowledge and control over the data in the statistical production, Statistics Sweden is implementing a dataset catalogue for the statistical production. The catalogue describes the datasets in the different phases, steady states, of the statistical production, with references to the in-house metadata systems together with references to the actual datapoints. Thus, the dataset catalogue will be a reference point between all metadata along with references to where the actual datapoints are stored, which is stored will be in each producing system. In order to automate the statistical process, the dataset catalogue is equipped with a notifying system. The consumer system can therefore set up a subscription for datasets of interests and fetch the datapoints from the producer system whenever new datasets has been entered in the catalogue. By using a json schema to describe GSIM datastructure, the datapoints can be retrieved through a generic api. For large datasets we can use other means of transport, but the datastructure schema remains the same.

1 Introduction

The production of statistics is changing. At Statistics Sweden we are having a new approach: to have “direct collection as a last resort”. Which means that we try other means to collect the data or minimize the data collected by human interaction. Different administrative registers and other data sources are harvested and internal *“data collection registers”* are created, which can be used as source for many Statistical Programmes.

Together with the harvesting of different sources, Statistics Sweden aims to automate the harvesting, collection, by the creation of internal *“data collection registers”* and data distribution between different Statistical Programmes. The main key issues when addressing automation has been identified as: context; meta-data; and generic design.

2 Pre-requisites for automation

In order to automate the statistical production, Statistics Sweden has chosen to use GSIM as the foundation, building up context and design using the Business Group, describing the contents through the meta data as in the Concept Group, structing the data as described in Structure Group and finally exchange the data using the information objects in the Exchange Group.

All activities and transformations are executed within processes described by the GSBPM, but updated to suit the Swedish approach to a more automated production.

There is a need to be able to work in a known context, where the design is pointing to data with relative paths, relative your own context. To achieve this, we use the metadata as pointers to where the data is located or will located, which is explained in more detail under section 2.6 in this document. By using relative pointers to the data in your design, the design can be valid independent of which year, month, or week it is executed.

Statistics Sweden has also implemented clear points where the datasets are documented within the production flow. These points, Steady States (but abbreviated HP in Swedish), are used to enable an overview over the data contents withing Statistics Sweden and as exchange points between different systems and or different Statistical Programmes.

2.1 Statistical programmes

Statistical Programmes are used as a container for statistical production for one or more statistical products, but it is within the Statistical Programme that the production is executed. Statistics Sweden also uses the programmes to create statistical registers. That means that we have Statistical Programmes that only collect data and will not produce any statistics, but create *data collection registers*. Other Statistical Programmes will then subscribe to data from different Statistical Programmes (*data collection registers*) in order to produce their statistics. In Figure 1 below you will find different Statistical Programmes and their interaction, together with their Steady States (Exchange points), also see section 2.3.

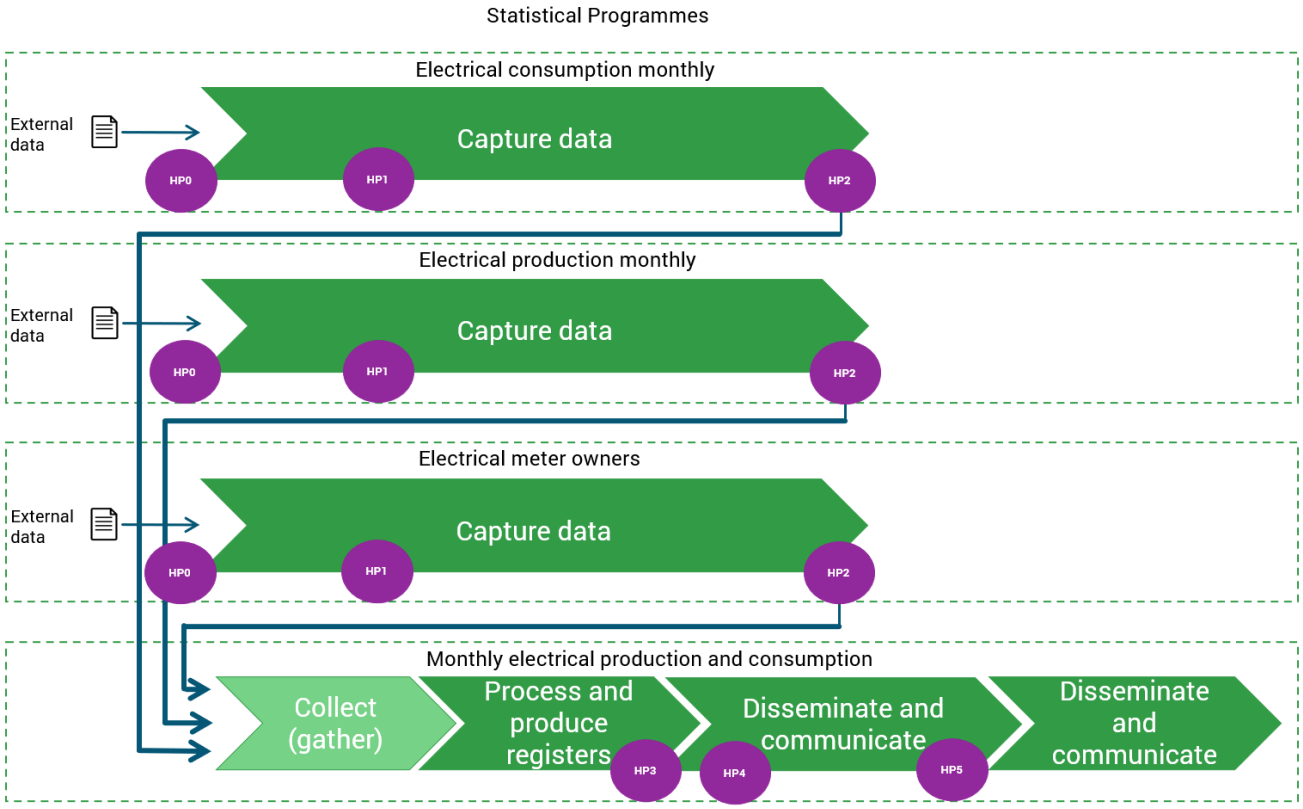

Figure 1- Statistcal programmes interacting to each other.

2.2 Using the frequency - cycles

The frequency of a Statistical Programme can depend on the frequency of the collection of data, the reference time of the data or the timing of the dissemination of the statistics. The Statistical Programme Cycle will allow us to address datapoints of a dataset before the dataset is produced, since we will know the relative path:

- {Statistical Programme}\{Statistical Programme Cycle}\{Steady State}\{DatasetName}

By leaning on the frequency, we can then divide between the design and the execution of the design. At design time we can address datasets not yet created, and let the design execute, during its cycle, when the dataset is available.

More how to use the Statistical Programme Cycle in a generic way, please see 2.6.

2.3 Steady states in production

The target for Statistics Sweden is to gain control over the produced data through metadata management and have clear handover points, *steady states*, between process steps and between statistical programs. The Steady States outlined are the following:

- HP 0: Raw data
- HP 1: Extracted raw data
- HP 2: Data collection register
- HP 3: Final observation register
- HP 4: Statistics
- HP 5: Published statistics and data

Data in the Steady States will be described with metadata and documented including a quality assessment. As a dataset is created in a Steady State, it is catalogued in the dataset catalogue. The dataset then holds references to meta-data such as: data structures (definitions of variables a.s.o.); Statistical Programmes; cycle, information security; if it contains personal information; when expected to be deleted or archived; and how and from where the datapoints can be retrieved (see also 2.4). In the figure below are the steady states or handover points (HP) indicated in the processes, according to Statistics Sweden business process model.

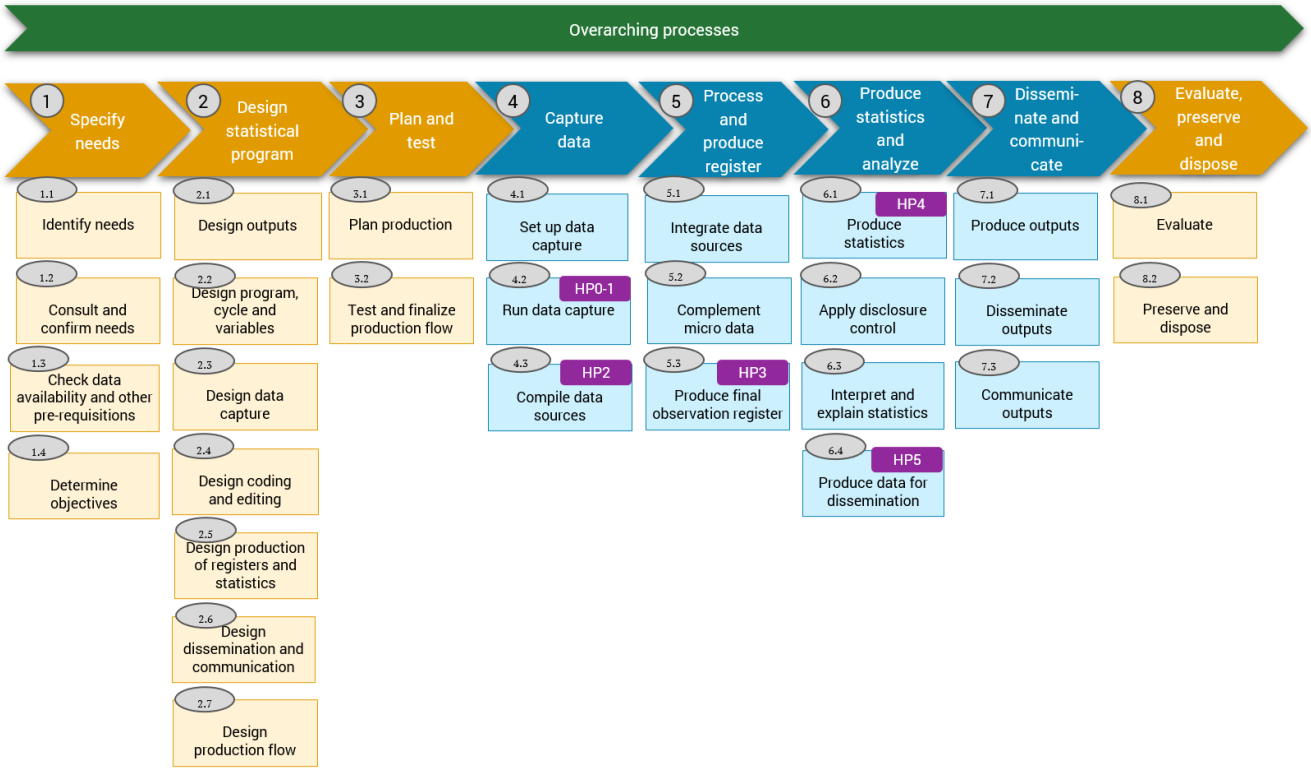

Figure 2 – Statistics Sweden business process model with Steady States indicated.

2.4 Dataset-catalogue

As soon as a dataset is produced for a Steady State (HP), then it will be registered in the central dataset-catalogue. This allows us to keep track of all data within statistical production, when setting up new Statistical Programmes but also for subscribing for data between different systems and or Statistical Programmes.

The dataset itself does not contain any information but keeps references to metadata. The dataset in the catalogue also includes a resource path where the data points are located. The resource path typically points to an API that returns the data points for the dataset. Examples of metadata attached to the dataset:

- Dataset identifier
- Dataset name
- Dataset version
- Datastructure identifier (reference to the meta-data system, where the structure is defined incl. variables asf.)
- Producer system, i.e. which system has produced the dataset
- Type of resource, i.e., how the data can be retrieved
- Resource path
- Statistcal programme identifier
- Process step instance
- Cycle identifier
- Steady state
- Information security classification

- If the dataset contains personal information
- Deletion date, i.e when the dataset will be deleted
- Creation date

2.5 Meta-data driven design: Creating a dataflow graph that controls the execution of the production flow

Statistics Sweden has developed support for creating Statistical Program Designs and associated Process Designs as dataflow graphs. The image below illustrates the principle of designing a production flow for a Statistical Program in the form of a dataflow graph. The dataflow graph starts with validating incoming data ("Steady state 0") from three sources and ends with creating a dataset according to the definition of "Steady state 2". Each transformation in the image creates a dataset that serves as input for the subsequent transformation. Note, only datasets created in a Steady State will be registered in the Dataset-catalogue.

The arrows in the image indicate the direction of the flow of datasets and should be read as "is input to". The transformations can be configured to start automatically, manually, or at a specific time of the day. A statistical program can have multiple dataflow graphs.

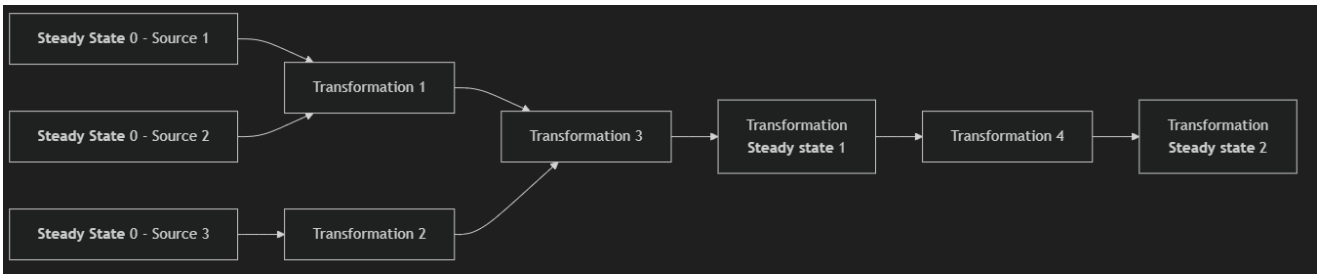

Figure 3 - the principle of designing a production flow for a Statistical Program in the form of a dataflow graph.

Before production starts, a "statistical program design" is chosen for the production cycle. Each transformation in the graph creates **one** dataset with a specified name and description.

Behind each transformation in the graph, there is a process method, rule, and a business service (IT service) that executes and creates the dataset. Rules are often expressed in code/scripts using Python, VTL, SAS, or R. Each transformation in the graph creates one dataset, and the transformations that create a steady state in the design are linked to a data structure that specifies the structure and content of the dataset.

During execution, each transformation starts in the correct order according to the designed dataflow graph. The business services behind these transformations listen for events to determine when they should begin reading the specified input datasets and start creating the specified output datasets.

It is possible to configure the transformation execution to start automatically once the previous transformation has completed and a new dataset has been created. Additionally, a transformation can be set to start manually or be scheduled for a specific time. Manual start is typically used when the user wants to analyse the results before continuing with production.

A dataflow graph is put into production when it is assigned to a statistical program cycle. All produced datasets then get a context: {Statistical Program}\{Statistical Program Cycle}\{Steady state}\{DatasetName}.

At design time, all "steady states" in the dataflow chart have a reference to a datastructure, see figure below.

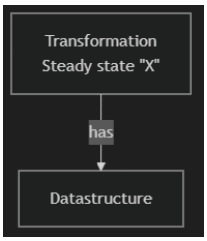

Figure 4 - A transformation in the dataflow graph will create a dataset with a specified datastructure.

When a transformation is executed according to the dataflow chart, the dataset and the data points are created, see image below:

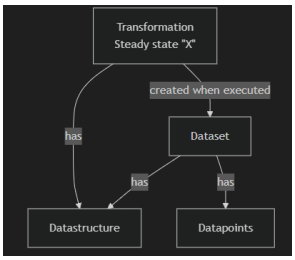

Figure 5 - Conceptual references to a dataset.

Below is a conceptual image of how datasets are structured in the dataset catalogue.

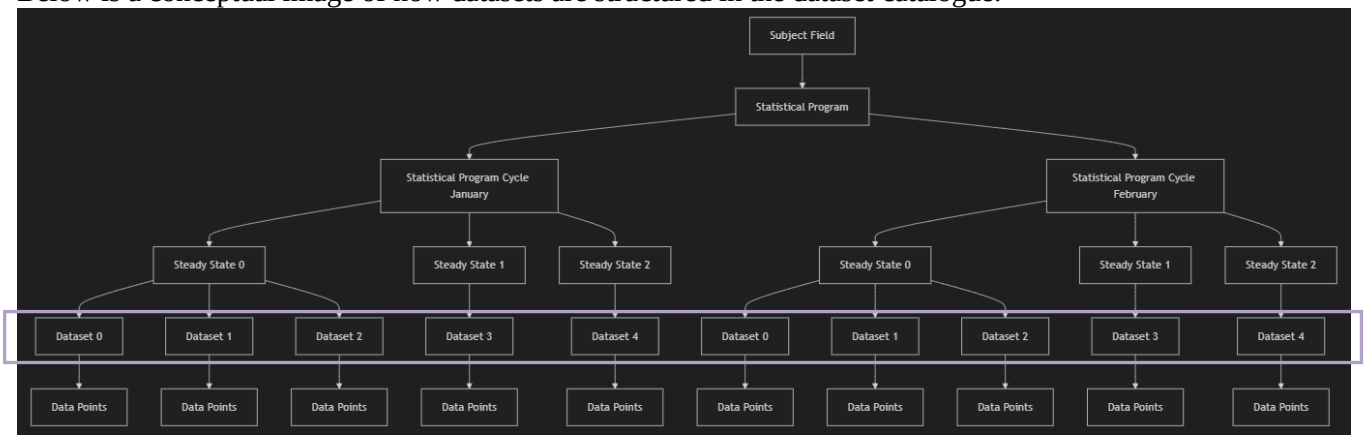

Figure 6 - Conceptual structure of datasets in the dataset catalogue.

The data points can be stored using various techniques such as Parquet files in buckets, tables in relation database, CSV files in folders, etc. The dataset catalogue contains a technical reference to the data points.

2.6 Access to datasets through relative paths

At design time, it is possible to include a reference to any other dataset in the catalogue, provided access has been granted by the information owner. The access to read a dataset is constructed relative to “T” with the following syntax:

{Statistical Program}/{**A[a-1]:M[m-1]**}/{Steady State x}/{Dataset_Name}

‘T’ is specified on the Statistical Program Cycle. For example if ”T” is set to 2024 October:

{A[a-1]:M[m-1]} Would then be resolved to ”2023M09”. The ”A” equals year and ”M” equals month.

For example in CPI it could look like this:

CPI/2023M09/SS3/A_Dataset_Name

The ability to use relative paths to access datasets in the catalogue allows for the reuse of dataflow graphs without hardcoded time references. In the dataflow graph, the "T" reference input can be seen as a "proxy" for a dataset in another context.

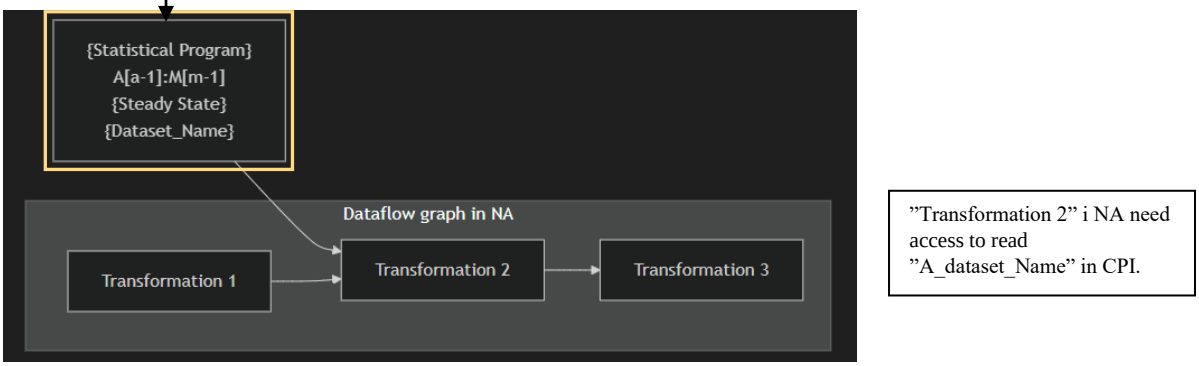

Figure 7 - Access to datasets through relative paths.

In the above example the information owner at CPI must set up an agreement with National Accounts (NA) that grant NA the right to read "A_Dataset_Name" in CPI.

3 Automated statistical production

As the pre-requisites for automated statistical production are in place, the different Statistical Programmes have to be transformed in order to suit an automation. Making the transformation means that the Statistical Programmes have to be re-designed or at least that the designs have to be established as meta-data driven designs, which means that the transition to a fully automatic statistical production while take some time. The part in the production that is most easy to automate is the collection or harvesting part, although we in many cases use different systems for the collection part versus the process and analyse part.

3.1 Automation using known context between different states

Since we can refer to datasets not yet created, we can utilise the design made for another system within the same Statistical Programme by requesting our own software to get the points as soon as the dataset has been written to the catalogue. This will allow us to automate the processing of collected data by subscribing to a certain dataset within the same Statistical Programme Cycle as the processing is done, e.g.:

- Statistical Program = My Statistical Programme
- Statistical Programme cycle = Current
- Steady State = HP1 Extracted Raw data
- Dataset_Name = Name of the dataset produced

If the Statistical Programme is CPI, which would like to set up a design in order to process the collected (web-scraped) data for prices, then the subscription of datasets would be: Subscribe to: CPI\Current\HP1\"Scraped_Data", see also 2.6. The system for collecting the data will be independent from the system performing the processing the data.

3.2 Subscription of data between different Statistical Programmes

The dataset catalogue contains a subscription service, allowing to share data not only between different systems within the same Statistical Programme, but also between different Statistical Programmes. Since we need to have some kind of protection of the data (not all Statistical Programmes can have access to all data), data can be shared from a Steady State, but the Statistical Programme which owns the data must acknowledge the

subscription request from the Statistical Programme that requires the data. For this purpose, we use a role based authority system based on Statistical Programme and role in said Statistical Programme.

Inter-Statistical-Programme sharing of dataset is manifested by a provision agreement between the Statistical Programmes, set up the role “Information owner” for both the information consumer and information provider. The consumer will set up the agreement and the producer will acknowledge it. The provision agreement will cover a named dataset of a Statistical Programme in a certain Steady State.

As the agreement is in place the subscription can be set up in the design of the information consuming Statistical Programme. The subscription will the use the notifications of the dataset-catalogue and thus be notified each time a new (version) of the subscribed dataset is produced and act accordingly in order to retrieve the datapoints for automated production.

Statistics Sweden has chosen to separate the actual storage of the datapoints from the retrieval of them. As a starter the datapoints are to be retrieved through an API, with a JSON-schema that corresponds to the datastructure of the wanted dataset. We foresee that we can offer the datapoints in several different formats, which actually has to be stated in the dataset in the dataset catalogue. The advantage by the separation, as we see it, is that we can gradually move our IT-platforms to support new storage formats without interrupting the retrieval of data.

4 Conclusions

By gradually adopting a metadata-driven design, which can be used for several Statistical Programme Cycles, we are able to automate many of the processes that are currently conducted manually. The starting point has been the National Accounts with the data flow graph designs, which have been extended to work across the entire agency through an agency-wide dataset catalogue. By using the context and structure provided by GSIM, in combination with steady states throughout the statistical production process, we can achieve a higher degree of automation.

This approach ensures that data and processes are documented in a standardized manner, facilitating the reuse of data to generate new statistics. It also creates opportunities for data scientists to search for, find, and understand the data more effectively. By increasing the provenance of data, we enhance its reliability and usability. This comprehensive documentation and structured approach not only streamline current operations but also pave the way for innovative uses of data in the future.