STATISTICS SWEDEN
Klas Blomqvist U/MET-S
K-E Kristiansson U/MET-S

2006-03-03

Developing a system for description of microdata at Statistics Sweden

Abstract

At present the Statistics Sweden metadata system consists of a number of tools and templates. The Quality Declaration template and the SCBDOK template for documentation of the production processes are the cornerstones of the metadata system. The software tool Metadok is a system for creating formalised metadata for the purposes of describing final observation registers in SCBDOK. However, owing to increasing demand for metadata, the support Metadok can offer is now considered inadequate for coordinating metadata and standardisation.
The aim of the ongoing project is to build a new model and a corresponding software tool that will provide an overview of the contents of the microdata registers at Statistics Sweden. Another important aspect is to facilitate documentation by making it easier to reuse and modify existing documentation.
The system is going to contain metadata about object classes, populations, variables and value domains including classifications. It will be a tool for standardisation of populations, variables, value domains and for documenting microdata.
A search tool function making it possible to search for metadata about macro and microdata by different criteria or combinations of criteria like object class, variables, reference times, statistical products as well as subject matter areas will be a key function in the system. This and other types of functionality will support staff in several situations such as:

- Evaluation if a need for information can be fulfilled with existing statistics or existing microdata without having to collect additional data, thereby reducing respondent burden and costs.
- Survey design (frame construction, auxiliary information in estimation or editing).
- Documentation (recycle existing metadata, step-by-step documentation of new systems/products).

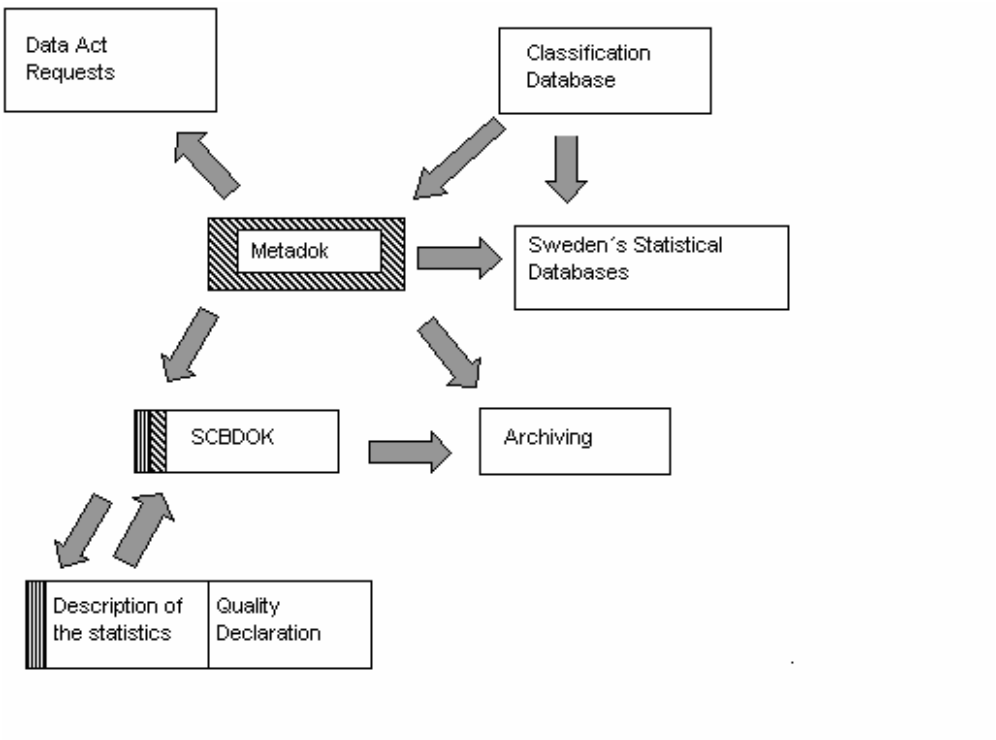
For external users and researchers the system will also provide a better overview of the data at Statistics Sweden.
The system is built on a conceptual model that is based on existing metadata systems and templates used at Statistics Sweden such as SCBDOK and Metadok and the demands for metadata within the organisation. It is influenced by ISO-11179 and ongoing work on a variable model in the Neuchâtel group.
The first version of the system, which is for testing and filling the database with content, is to be released in January 2006.

The paper discusses the project goals, the conceptual model and the functionality of the system.

The present situation, an overview

The metadata system at Statistics Sweden today consists of several tools and templates. For each of these there are guides that describe the functionality and interaction with other parts of the system. The figure below gives an overview of how these interact.

Figure 1. The present metadata system at Statistics Sweden



A brief description of the different parts in the system is presented below.

Description of the statistics

The purpose of this template is to provide a short description of the quality of the statistics and other basic facts. It contains one section with general information and a quality declaration section. All official statistics should have a *Description of the statistics* according to law.

SCBDOK

Since 1994 all final observation registers and production systems that Statistics Sweden is responsible for should be documented in SCBDOK. The purpose is to provide a detailed description of the creation of a statistical register from data collection to dissemination. SCBDOK is a Word template. Some of the information in SCBDOK is included in *Description of the statistics*.

Metadok

Metadok is a program that is used for describing a physical database for a microdata register. The metadata in Metadok are formalised and can therefore be used by other software. The purpose is to make multi usage of the documentation possible, such as dissemination on the Internet, Data Act Requests, archiving and aggregations in PC-Axis. Metadok can also be added as a part of SCBDOK, for describing the content of the final observation register. The Metadok documentation is published on The Statistics Sweden Webpage as a part of Sweden's Statistical Databases. In order to reduce the documentation workload it is possible to import already existing metadata from several sources for example from the Classification database.

The classification database

The Classification database contains national as well as international classifications, such as regional classifications and activity classifications.

The Project

Background

The current system for documenting microdata *Metadok* has some limitations when it comes to content as well as functionality. The most important are listed below.

- It is difficult to get an overview of the metadata due to that the variable content of a statistical product is tied to one register at a time.
- The information in the registers is not sufficient for evaluating the possibilities for matching data from different surveys or registers.
- Lack of information regarding comparability over time.
- The insufficient information about the registers and the difficulties to get an overview of the data makes coordination and standardisation very difficult and time consuming.
- The support for documenting in Metadok is inadequate.

Goals

One important aspect is to develop a system for documentation of microdata that has an improved functionality and higher quality of the content than in Metadok. The system is going to work together with other parts of the metadata system. The system will:

- Provide an overview of metadata for the microdata.
- Provide information for evaluating the possibilities for matching data from different surveys or registers.
- Operate as a tool for increased standardisation and harmonisation and therefore give better possibilities for data coordination.
- Provide a better support for documenting so that the overall quality of the documentations can be increased.

Delimitation

The project does not deal with harmonisation and standardisation of content. Support and future development of the system is not a part of the project either.

Working practice in the project

The system is going to be an important tool for a large part of the staff at Statistics Sweden. Therefore user groups consisting of interested persons were tied to the project from the beginning. The user groups were categorised in subject-matter statisticians, methodologists and system developers. These groups took an active part in developing the content and functionality of the system in order to be a useful tool for them in their everyday work. Other requirements have also been collected by the project group and taken into account when developing the system.

Project organisation

The project group is led by staff from the Methodology unit (Research and Development Department). In the project group there is also staff from the IT and Register and Microdata for Researchers units (also from the Research and Development Department) and the Information and Publishing Department. The project has a Reference group consisting of representatives of all departments at Statistics Sweden and a Steering committee where the current situation in the project is reported on a regular basis.

Examples of frequent work situations where VHS is a tool

The work with the collection of requirements and in the user groups has resulted in several areas or situations where the system can give a substantial support for the production.

General

- New enquiries.
- Consider whether an enquiry can be dealt with using already produced statistics or collected data.
- Searching for metadata on microdata that consists of relevant object, variables and population.
- Can different registers be used in order to make an integration register.
- Comparability over time, information on time series.
- New surveys.

Survey design

- Frame construction.
- Auxiliary information in the estimation.
- Auxiliary information in editing, coding and imputation.
- Reduction of over coverage in surveys considering certain subpopulations.

System development

- Modelling databases.

- Using code tables from production systems in connection to editing and coding.
- Search and provide information of where data is stored physically.

Documentation

- Reuse of old documentations
- Starting point in already existing metadata
    - Classifications and standards
    - Documentations made by others
    - Searching the metadata
- Document once – reuse

Extended demands for functionality and content in VHS.

The above-mentioned demands means extended requirements for functionality and content in VHS.

The system shall be coordinated with other parts of the metadata system.

This means that:

- There is a connection to other types of documentation such as SCBDOK, Description of the statistics, the system development tool SiP and a future questionnaire database. VHS also has to be connected to the product database, the archiving system, the system for Data act requests and the Statistical database SSD.

The system shall provide information for evaluating the possibilities for matching data from different surveys or registers.

This means that the system will contain:

- Content metadata, i.e. on variable, value domain, object class, and population
- Historical information about variables and value domains
- Comparability over time – inform about breaks in time series and changes in definitions.
- Accurate and up to date information on technical metadata, i.e. column name, server, database and table

The system shall provide an overview of the microdata repository.

This means the system will contain:

- A search function for survey design purposes, for example flexible search functions for selecting variables.
- The search function should be non hierarchical with multiple search entry possibilities.

The system shall be a support tool for increased harmonisation and thereby better possibilities for coordinated use of data.

This means the system will contain:

- Standards for object classes, variables and value domains.
- Consistent and distinct definitions.

The system should give appropriate support for documentation so that quality in all documentations can be increased.

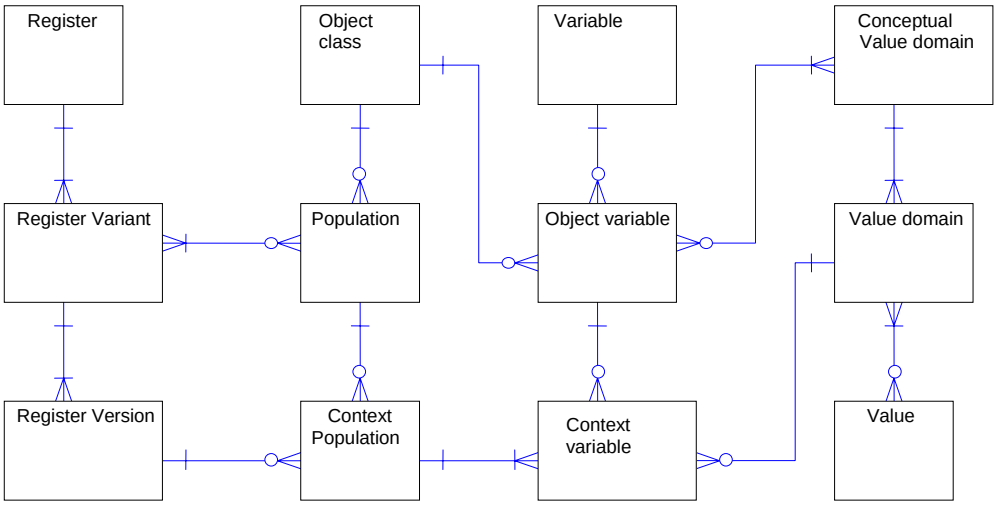
This means that in the system you:

- Will be able to document variables over time, without having to document a whole register.
- Will have one metadata container for each statistical product or enquiry, from which different types of documentation can be created.
- Will be able to create a metadata holder that does not belong to a specific statistical product, register or enquiry.

The conceptual model

The development process of the model has to a large extent been driven by the content demands and requirements stated above as the purpose of the project. Already defined metadata concepts and terms in the organisation have been used when possible. ISO 11179, Information Technology -- Metadata Registries (MDR) has been used as an input in the making of the model as well as ongoing work in the Neuchâtel group.

Figure 2. The figure below shows a summary of the core of the conceptual model



Concepts

***Object class*:** An Object class is an abstraction of an object. An object is an in itself independently existing entity. When used in a statistical context it becomes a statistical object. There are two types of objects that can be isolated (described) in the model.

- *Register object:* The Object class that is connected to the physical register.
- *Target object:* The Object class that the target population and the survey population consists of.

Example: person, local unit, organisation.

***Population**:* Description not related to a specific survey round of the quantity of objects that the survey intends to collect data about. The description can be related to either the Population from which the register objects are originated, i.e. the Register population and/or the Population from which the target objects are originated, i.e. the Survey population.

- *Survey population:* The part of the frame population that is included in the target population.
- *Register population:* Description of the Population of Register objects that the Register encompasses totally or contains a sample from.

Example: persons in Sweden on December 31, local units in Sweden at January 1, organisations in Stockholm during the year.

Context population: Description of the quantity of objects that the survey intends to collect data about in the specific survey round. Context population concerns Context register population and Context survey population. The realisation of Register population and Survey population gives the actual populations for the current survey round.
Example: persons in Sweden at December 31 2005, local units in Sweden at January 1 2005, organisations in Stockholm during 2005.

***Register**:* An overall denomination for the register that is the basis for the statistics.
Example: The population register, the business register.

***Register variant**:* Common non-survey round dependent description of the Register variant that is used in the survey. One Register can have several Register variants. Different Register variants are distinguished by different definitions of their Register population or differences in variable content.
Example: The population register complete variant, the business register sample frame variant.

***Register version**:* A Register version is a realisation of a Register variant at a certain point of time/survey round.
Example: The population register complete variant 2005, the business register sample frame variant 2005.

Variable: A Variable states a characteristic that can be connected to one or several Object classes.
Example: monthly income, annual turnover, industrial activity.

Object variable: A Variable that has been connected to an Object class.
Example: person monthly income, local unit annual turnover, organisation industrial activity.

Context variable: A Context variable is connected to an Object variable and has a connection to a Context population, a Value domain, a Register version and has an information source.
Example: monthly income for persons living in Sweden December 31 2005, local unit annual turnover 2005, organisations in Stockholm after industrial activity 2005.

Conceptual value domain*:* Presentation of level of detail and definition of a set of categories for a value domain and its categories conceptual meaning. *Example: SNI 2002 (Swedish Standard Industrial Classification 2002).*

Value domain: The level of detail for a Variable and its representation with measure unit or alphanumerical specification (code) and definition of the different categories in the given level of detail.
Example: SNI 2002 (Swedish Standard Industrial Classification 2002) 5-digit level, 0- (Swedish kronor).

Technical information

The system provides links to where (i.e. server information, column name etc) the data is physically stored. This means that the user does not have to change names in their production systems as long as they link to the correct metadata in the system. Data is not a part of the system.

Sources

On all Register levels and Context variable there is information on sources for the given Register level or Variable.

Structure

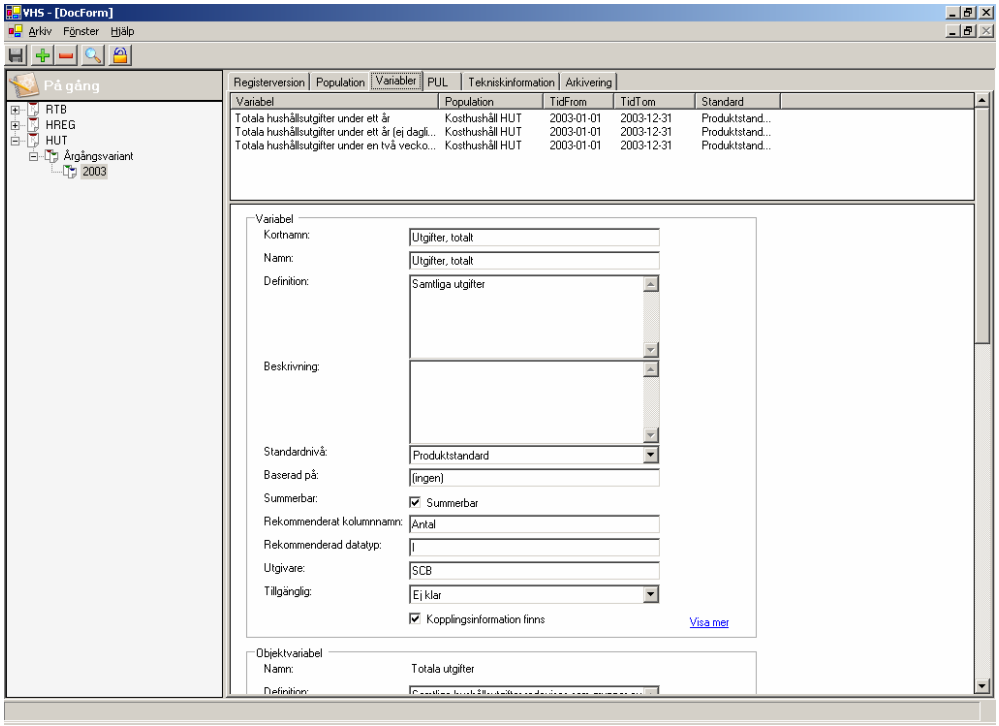
The application as based on the model gives the following content structure for a given Register:

- **Register123**
  - Register variant A
    - Register version A1
      Content:
      - Population A1
      - Variables
      - Value domains
    - Technical information
    - Sources
    - Register version A2
      Content:
      - Population A2
      - Variables
      - Value domains
    - Technical information
    - Sources
  - Register variant B
    - Register version B1
      Content:
      - Population B
      - Variables
      - Value domains
    - Technical information
    - Sources

What's going on at present

The collection of requirements for the system considering content and relations to other systems is closed. Based on this the database is now close to stable and the application is under construction. Due to this a lot of issues concerning functionality arise on a continuous basis. A beta version of the application was released in January. The user groups are now working actively giving input to the development of the application. In July a new version will be released ready for starting to fill the system with essential basic metadata about registers that are sources for many other registers and surveys. After that the system will continue to be filled completing it with the rest of the registers.

Figure 3. The figure below shows a "print screen" of the application version 1.



Version 1

In January 2006 a first version was released. It has very limited functionality and is primarily used for testing and to get input from user groups. Information can be added in this version, but it is not intended to be used in production. The database contains Classifications, some standardised object classes, and standardised variables and a few test examples.

Version 2

Version 2 will be released in the beginning of the summer. This version will have increased functionality and support more working operations. It will

contain registers, register variants, register versions and frequently used value domains. At this stage migration of metadata from Metadok can be started. How and to what extent metadata is to be migrated from the Metadok system is not yet established.
When the migration of metadata from Metadok is finished for a register the metadata has to be adjusted and completed in accordance to the structure of the system. Variables and value domains can then be harmonised in the system.

Version 3

Version 3 is a version where users can start using the application in the production. How much content the system will have from the start is determined by the level of ambition set by the departments in the migration process.

Harmonisation and standardisation

The project has not dealt with actual content standardisation and harmonisation. To be able to fully take advantage of the functionality of the system there is a great need for content related work, mostly with standardisation issues. The project is responsible for the development of the system, not the content standardisation and harmonisation of the content. The system will as mentioned earlier be a tool for this and is also to a large part dependent on the content. If the system is to be used effectively there has to be information in the system to reuse.