WP.37
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (vi): International collaboration

A strategy to improve the register system to store, share and access data and its connections to a generic statistical information model (GSIM)

Invited paper

Prepared by Anders Holmberg, Klas Blomqvist, Jakob Engdahl
Hans Irebäck, Lars-Göran Lundell and Jörgen Svensson, Statistics Sweden

I. Introduction

1. Improving the register system is one of Statistics Sweden's strategies to meet future challenges of official statistics[1]. When developing a fully register-based census, opportunities also emerge to produce new statistics or more of the regular statistics by means of registers only or suitably combined with directly collected sample survey data.

2. To facilitate this, an efficient production system is necessary, where information about the data that flow through the production process is documented, monitored and used for decisions. The Generic Statistical Business Process Model (GSBPM) provides a framework for the development. In line with Statistics Sweden's adopted data warehousing and register-coordination strategy, this paper proposes some preliminary ideas of how a statistical information model can be designed. We demonstrate the approach by taking a closer look at the editing steps involved when data from administrative sources are used and entered into the register system and the statistics production environment as a whole.

3. Initially, this paper describes the types and the main stream of administrative data through Statistics Sweden's statistical production process. Thereafter we continue by defining some key components of editing these data when they are used for statistics. And finally, a timid synthesis of the two parts is given by focussing on the information needed to edit administrative data in a statistical production process.

II. Administrative data and registers in the statistical production process

A. A general overview

4. Although it has until recently caught comparably little methodological attention, producing statistics from originally administrative data has a long history. In Sweden this was done long before the advances of scientific sampling methodology in the first half of the 20:th century. The pressure on NSI:s to cut costs, decrease response burden and to make production more efficient has meant increasing international interest in using administrative data for statistical purposes. Within the European statistical

[1] E.g. see Eurostat and CBS (2009) PG-TF2 Statistical Challenges paper from the Conference 'Work in Progress' The results of the Cracow Action Plan 15 & 16 January 2009, World Forum The Hague

system a collaborative project is ongoing within the MEETS program, a work package in the Blue-ETS project of the 7th framework program aims at developing quality indicators, and an increasing number of European countries are developing production systems using administrative data and registers in their Census.

5. Statistics Sweden has decided on a data warehousing and register-coordination strategy. A vital part of the strategy is to improve the register system by linking base registers together and create an environment where the registers are well integrated with the statistical production process. This comprise administrative data and registers not only being used as basis of population frames and auxiliary information in survey sampling based statistics, but also being used as main sources for statistics, as sources of quality assessment and as core components of the systems of economic and social statistics as well as new statistics where target objects are suitably formed by linking base registers. It also opens possibilities to get statistics on small populations.

6. Figure 1 shows Statistics Sweden's three base registers and their links. The base registers define the populations of businesses, individuals and real properties. They are each the centers of their spheres of interest. They are continuously updated and contain history, mainly stock variables and time stamped data. They are independent of the other base registers but connected by standardized links.

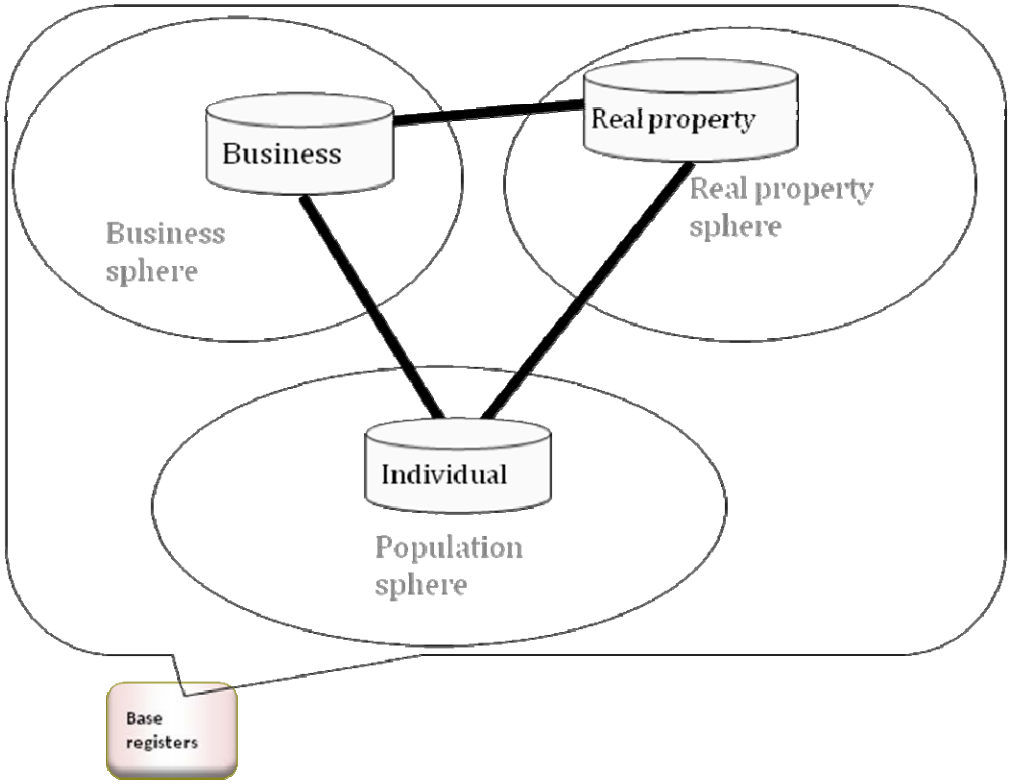


Figure 1 Illustration of Statistics Sweden's base registers and their spheres of interest

7. The spheres of interest contain definitions and rules of how to derive other objects from the base register's basic object. They also contain variables connected to the object types of the sphere. These can originate from other register sources or from surveys. Maintaining the base populations and their core variables is a central function of a base register process. Statistics on object types from business, real property or individual populations should stem from base registers and thereby increase comparability and coherence.

8. Before having a base register, data have to be retrieved and processed. Statistics Sweden receives most of its administrative data from other authorities. The Swedish tax agency and the Swedish mapping, cadastral and land registration authority are the most important for the base registers but for the spheres there are also other authorities that provide data. A large majority of the statistics done in Sweden is based on administrative sources. To standardize the retrieval Statistics Sweden use a system called

Indatatratten or Tratten[2] (the Funnel in English). Today the data that pass the Funnel are being disseminated and preprocessed at many different locations and products.

B. The future data warehouse and information management

9. Figure 2 illustrates the objective of the decided data warehouse strategy and its relation to Statistics Sweden's variant of the GSBPM.

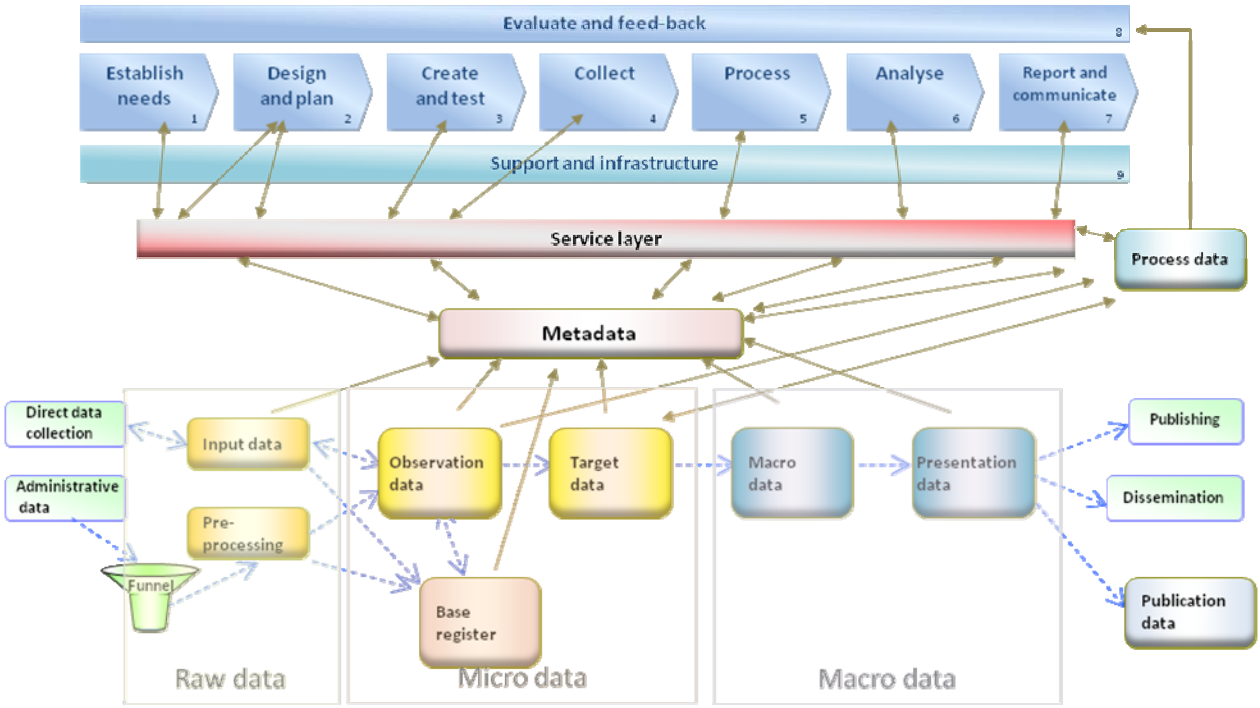


Figure 2 Description of Statistics Sweden's future data warehouse

10. The lower left part of Figure 2 demonstrates where in the future data warehouse environment the administrative data and the base registers fit. As data are refined through the production process, from raw data via micro data to macro data and published statistics, the register data are combined with directly collected data in a data storage environment (i.e. a type of data warehouse). To facilitate this, the production system has to be such that the necessary information of the data that flow through the GSBPM is documented and monitored and also used for decision-making. The idea is that this flow is made possible by a Metadata layer, a Service layer and a warehouse with Process data that provide survey managers and production systems analysts with vital information to detect errors and improve the quality of the statistical product, or the statistical production process system as such. An ideal would be a condition where survey managers have possibilities to design and use the organizations survey resources optimally.

11. One precondition to the description of the above production environment is to operationalize metadata. To make the statistics production metadata-driven has been an ideal and goal for many years. One particular example worth to mention is described in Zeila (2009). He shows how an early system was developed in Latvia in 1996–1999. The work is partly based on papers by Sundgren, and an another interesting paper in this context is Sundgren (1999).

12. Recently efforts are put in internationally to develop a Generic Statistical Information Model (GSIM). Its purpose is to aid statistical organisations to agree on common terminology and definitions to assist their discussion on developing metadata systems. GSIM is also essential as a reference model. The goal is that it can be operationalized whenever defining the information required driving statistical production processes.

[2] There are still today some amounts of administrative data that are retrieved outside the Funnel.

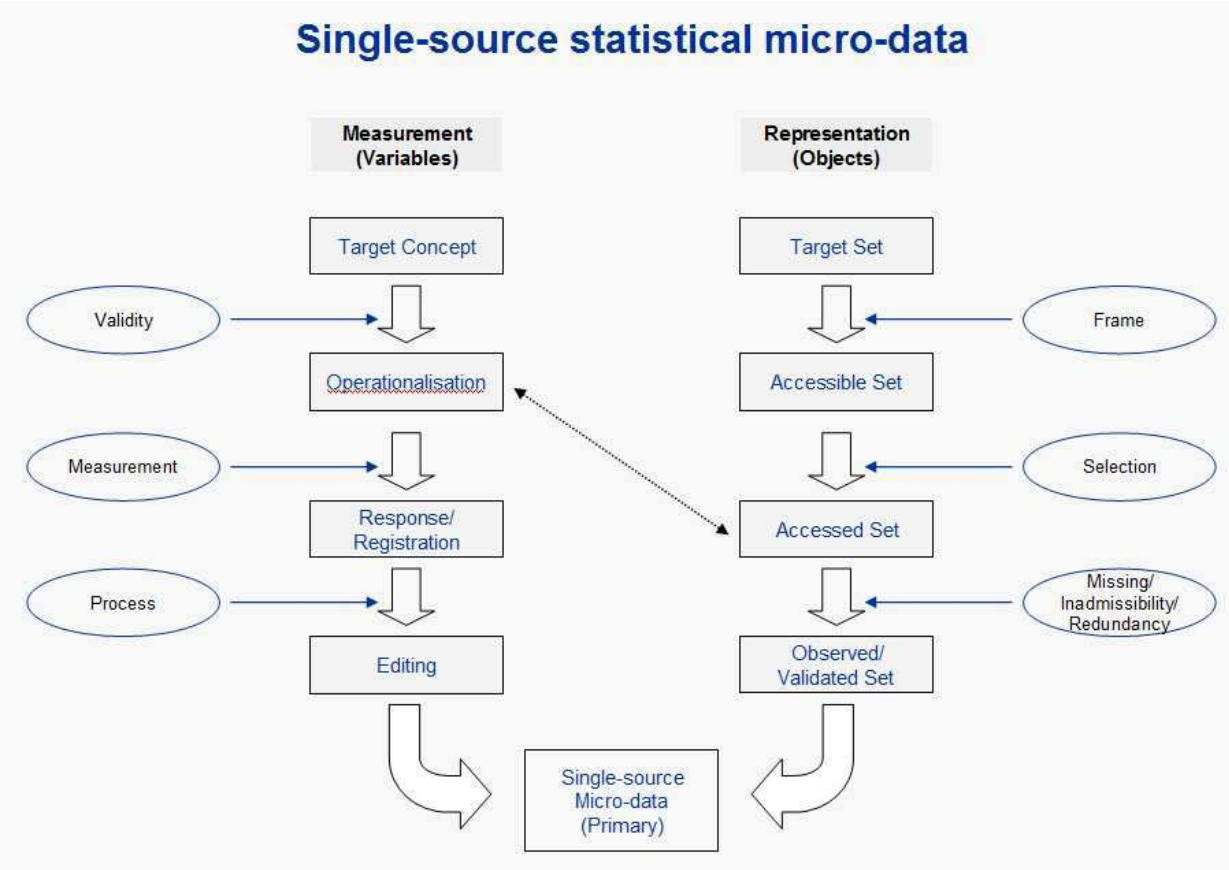
13. By looking closer at the editing process of administrative data and registers we will present some preliminary ideas and needs when a statistical information model is designed. The points presented are ongoing work and the result is a small contribution to the discussion of GSIM development, which is done as an international collaboration effort.

III. Editing administrative data

A. Concepts of quality of administrative data and editing

14. To avoid increasing costs and delays, the quality control of statistics should start at earliest possible stage. The editing process is one part of the quality control. One overall objective of editing is to find error sources and then eliminate them. In statistical surveys where the purpose of the data collection is set and controlled by the statistics producer (i.e. in surveys with direct data collection), a lot of resources can be saved if error sources are found and eliminated. This is probably the case for pure register-based surveys as well; however it is much harder to do. One reason is, in Statistics Sweden's case, inability to contact the respondent. Moreover, only in very rare occasions is it possible to contact and influence the responsible authority (the data owners). Another reason is that multiple usage of the data does not always allow clear conclusions of which errors that are more serious than others and most cost efficient to take care of.

15. Let us take a closer look at the various types of editing that are applicable to administrative data and registers. Our starting point will be quality models, the error model described by Zhang (2010) and the division into input data quality, production process quality and output quality proposed by Laitila, Wallgren and Wallgren (2011). Figure 3 below from Zhang (2010) sketches a two-phase cycle of microdata. The last cycle results in an integrated data set from multiple sources, which is particularly common when register data are used in statistics.



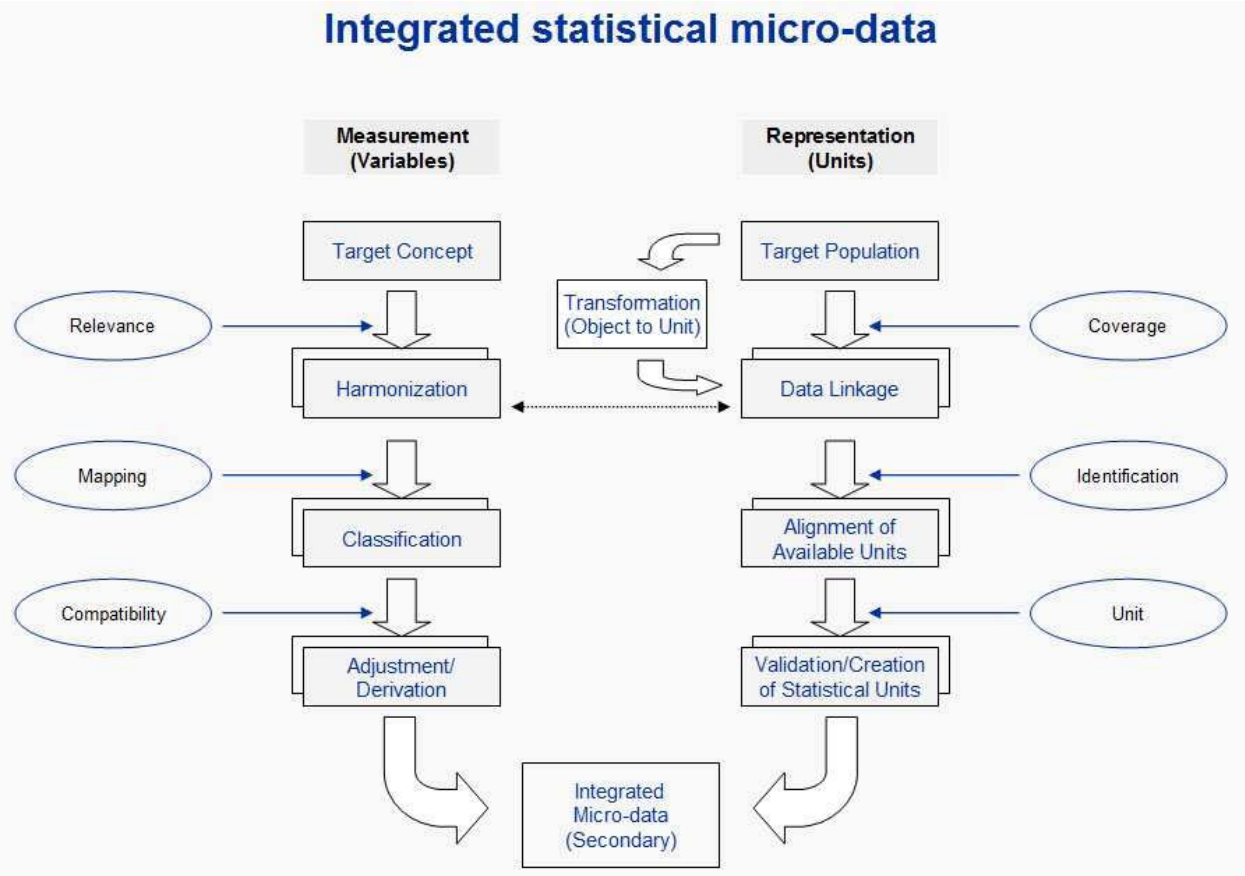


Figure 3 Two-phase cycle of error microdata, single (primary) source and integrated (secondary) source

16. Figure 3 is an extension of Bakker (2010) and borrows inspiration from the total survey error framework by Groves et.al. (2004, Figure 2.5). Conceptually, it illustrates sources of errors as well as roughly the type of error in a specific register survey, i.e. when there is a determined target population, given target parameters, domains etc. Compared with Figure 2 the end-box of the cycles contains data from the observation data or target data (the boxes in the microdata area of Figure 2).

17. When errors appear along the measurement and representation chains of Figure 3, editing tasks should be directed at detecting, minimizing effects of and (in recurrent production) eliminating the error sources. To be able to do this and also be able to distinguish between important error sources which you can treat effectively and others that are less important or important but very hard to rectify, we need effective metadata that contain process information and quality information. This is a notion of concern when developing a GSIM model. Hence, in addition to metadata about the study population and study variables a GSIM model must contain this process and quality information on different levels.

18. To give a thorough picture of the editing processes of administrative and register data at an NSI such as Statistics Sweden, we have to add aspects not visible in Figure 3. Since the perspective there ends at integrated microdata, it is single survey oriented and it does not necessarily cover multiple uses of the same register data. Editing activities on microdata with the purpose to check coherence with other aspects of a statistical system is not covered. Pure output editing perhaps aided by geographical information is also a part of the editing toolbox that falls aside. A GSIM or at least the metadata tools developed and used by NSI:s must of course support their whole production system and not be a survey by survey support only. Some of this production system's information can be in the process data environment of Figure 2, but it can also be pieces of information to a register system dash board, not necessarily applicable to other organisations.

19. It is easier to identify the information needed for such *system* parts of statistics production and editing if we divide the quality of register data by input quality, production process quality and output quality. In Laitila et al. there are examples of quality measurements that can be useful to describe properties of multiple usage of records and variables in a register.

20. Indicators based on microdata are also important pieces of information that can improve coherence of a statistical system as well as being used in a Deming-cycle to improve a system of registers. A statistical production environment using register data would benefit from such data but it remains to be discussed if such information has to be part of a GSIM.

B. Editing processes on administrative data and in a register system

21. Conceptually, the editing processes of the administrative data in Statistics Sweden's register system are shown in Figure 4. It is simplified to show possible editing steps from the lower left corner shown in Figure 2, through one of the spheres of the register system and the observation data to the processes number 5, 6 and 7 in Statistics Sweden's GSBPM.

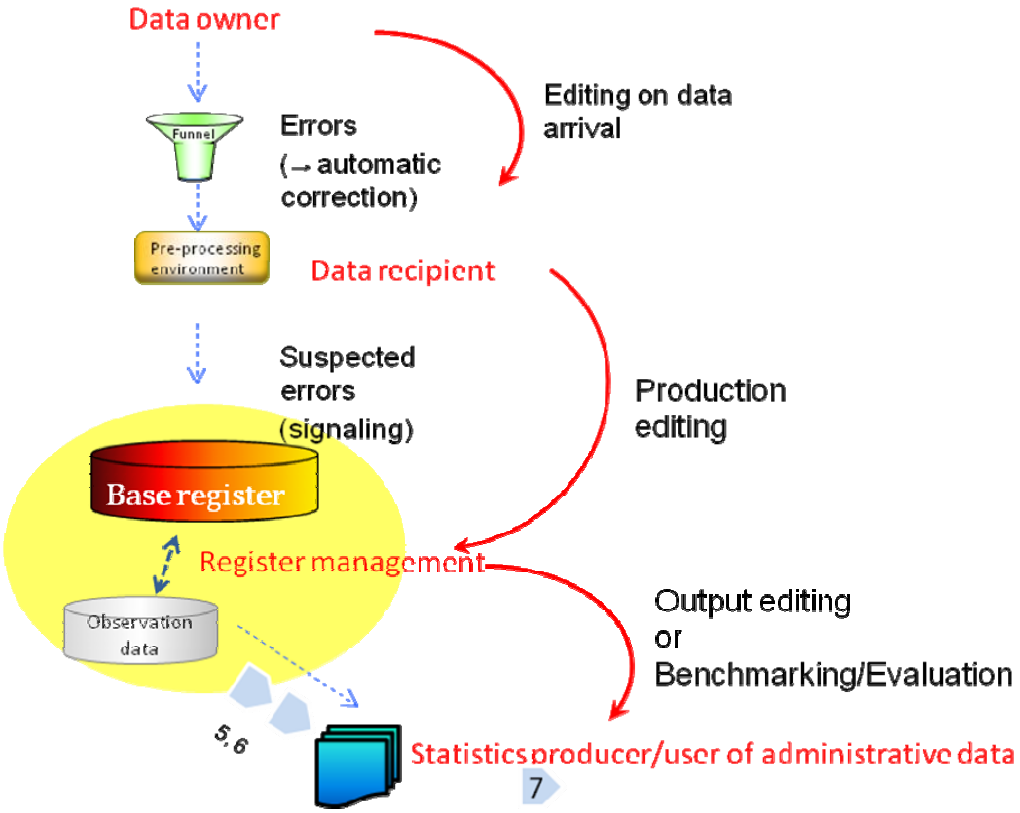


Figure 4 A simplified overview of editing activities in Statistics Sweden's register system

22. There are four main roles involved, the data owner, a data recipient, register management and the end user of the administrative data within the organisation. We call the bulk of the editing work production editing. It involves the editing tasks with goals mentioned in the previous section on both register populations and variables. Some editing will be part of register maintenance and keeping register quality, some will aim at specific surveys and its (perhaps) integrated microdata and some will be on microdata directed to improve statistical systems. (Wallgren and Wallgren (2007) give examples of the latter for the system of economic statistics.)

23. The viewpoint of doing editing work at the earliest possible stage has to be adapted to the multiple uses of register data and other practical circumstances of the administrative data source. The possibilities to contact the sources are normally limited and this makes it hard to make corrections. Caution of using imputation techniques too early should also be applied, if there are many uses and users of the data.

24. Statistics Sweden has made a decision that all administrative data should be received by the same system, The Funnel. Today, very little editing is done there, but there is a potential to already at that phase automatically correct obvious errors and signal suspected errors. The advantage (e.g. for VAT data)

is that a first check is done before data are disseminated further to the production processes of different users who today most likely repeat much of the same editing process independently of one another. Saving these signals and making them easily accessible to all users could avoid this.

C. Editing between the spheres and use of other sources

25. Editing across the system of registers is another area which should be given attention. One example is the Swedish Census and future household and housing statistics that will depend on the quality of the underlying registers. Defining the editing process and thereby relevant methodologies of quality control will be in focus the upcoming years. This includes of course data registration editing, some micro editing and output editing, and in the build-up phase collaboration with the authorities that collect the administrative data. An example where editing between the register spheres is necessary is register-based statistics on living space standards. The relation between sizes of household units and living areas comes from different spheres and both have to be of good quality to provide good statistics.

26. Cross-checking with other sources is another possibility. If future household and housing statistics are to be mainly register-based, there is a need to evaluate the quality using sources outside the registers. A designed system of complementing questions in regular sample surveys could be a way of doing this. Naturally the information collected there can be used in editing processes.

IV Summary

A. Examples of information needed in the editing processes

27. If a goal of a GSIM is to aid statistical organisations to agree on common terminology and definitions and to assist their discussion on developing metadata systems, then an important starting point is the quality of the statistics and the work done to assure this. Editing is one example of such work and we have begun to look closer at its processes with regards to administrative data and registers. The use of that type of data in official statistics has increased and is likely to continue to increase worldwide in the near future.

28. The GSIM development is in its initial stages. A final structure of a model will most likely have many levels and most of the editing information will probably be on lower levels. However, an important part is the process information and the quality information that are made visible by editing. A couple of introductory examples of this are the following:

- Edit rules (both on microdata and macrodata). It could be acceptance regions, instructions and other decision rules such as maintenance and validation rules of statistical registers.
    - Dependencies between editing steps that are done in a sequence, motives of edits and information about multiple uses of data.
    - The linking rules applied in a register system.
    - Links in a survey system as a whole, which make objects traceable from base registers to the surveys in which they are included.
    - Inspection rules of a statistical system including register status and quality, e.g. benchmarking statistics and rules of how objects are activated and deleted in populations.
- Status codes on observations with respect to editing, both on items such as signal flags and on population sets such as an indicator of the state in processing.
- Edit indicators (e.g. summary measures from ongoing processes).
- Error codes from edits.
- Time stamps of changes (on different objects and levels).
- A description of the type of edit information (any of the above).

29. Statistics Sweden’s work with the strategy to improve the register system and creating the future data warehouse environment (Figure 2) has the use of metadata as the general principle of data processing and it has to handle the above components to be fully realized. The more parts that get

implemented the more we will get an efficient use and availability of statistical data by using the common data warehouse. Users (statistics users, statistics designers, statistics managers) will be provided with adequate data and tedious and time-consuming tasks replaced by value-added activities.

30. To implement this, collaboration in various fields is necessary. Information management combined with methodology and quality management on content as well as technology are all prerequisites. The probability of success and the development speed will increase if there are common points of references (and goals). Models such as the GSBPM and a future GSIM could play an important part in this if the resources put into international collaboration can focus on common solutions. This is a big opportunity for official statistics production that is easy to see, but it is at the same time a challenge.

31. Although NSI:s roughly do the same business, their organizations differ in culture and available resources. To make models as the GSBPM and GSIM useful, overview knowledge of the whole statistical production process and its dimensions are required. This can, probably better, be achieved with international collaboration. A question is then if the models should be parts of an international architecture for statistics production that acts as a role model to national systems or if architectures developed nationally gradually become more similar by being inspired by international work? The answer remains to be seen. The really difficult challenges of the international collaboration are to communicate the results of the work and then implement it nationally. The proposed solutions must both be and appear as useful to the production systems, and primary keys of success are competence among staff, timing and high-level support.

V. References

Bakker, B. (2010) *Micro-integration: State of the Art*. Paper for the Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses, The Hague, The Netherlands.

Groves, R.M., Fowler Jr., F.J., Couper, M., Lepkowski,. J.M., Singer, E. and Tourangeau, R. (2007) *Survey Methodology*, New York, Wiley

Laitila, T., Wallgren, A. and Wallgren B. (2011) *Quality Assessment of Administrative Data.* Research and Development – Methodology reports from Statistics Sweden 2011:2

Sundgren B. (1999). An information systems architecture for national and international statistical organizations. Meeting on the Management of Statistical Information Technology (Geneva, Switzerland, 15-17 February 1999)

Wallgren, A. and Wallgren, B. (2007) *Register-based Statistics: Administrative Data for Statistical Purposes*. New York, Wiley

Zhang (2010). Notes from the Blue-ETS WP4 meeting in Heerlen September 15, 2010, and *Developing Statistical Theories for Register-Based Statistics*, in Qvintensen nr 4, 2010, pp 21-22.

Zeila, K. (2009) *Meta Data Driven Integrated Statistical Data Management System*. Paper presented at the conference Modernisation of Statistics Production, (Stockholm, Sweden 2-4 November 2009). http://www.scb.se/Grupp/Produkter_Tjanster/Kurser/ModernisationWorkshop/final_papers/C_2_SOA_Zeila.pdf