*Seminar on Registers in Statistics - methodology and quality*
*21 - 23 May, 2007 Helsinki*

# *The database for social statistics in Denmark*

*Finn Spieker*
*Statistics Denmark*
*fsp@dst.dk*

## *Chapter 1. Change of register strategy*

The register system for statistics on persons has to a very high extent been influenced by the stepwise development during the last thirty-five years. The number of data has increased heavily and the use of general identifications, mainly the PIN-code, has lead to plenty of opportunities for producing statistics based on varying combinations of variables. The data have been organised in registers as physical units of the statistical system and still more registers were created completely or partly containing variables found in other registers. For different reasons it had to be so, but the system as a whole had grown to a size that hardly anybody could cope with it, and the expected continuing extensions would contribute to further complications. So a new strategy for the register system was decided and it is implemented for the time being.

The results of the reorganisation will be:

- harmonised definitions of concepts used in the different fields of statistics
- clarified responsibility for creation, maintenance and documentation of each variable
- simplified and more transparent documentation
- improved guarantee for optimal choice of variables to be used in the individual case
- a more efficient operation of the system
- a higher degree of flexibility by means of knowledge sharing achieved by closer contact between collection and deduction of data and the final use of the information
- less space required for storage of data

These improvements will contribute to ensure the quality of the statistics and a more optimal use of the resources.

Further it was decided to change platform from mainframe to PC/LAN which means that the system by now is organised in SAS and Oracle databases. Oracle is expected to be used for the whole system in the end.

## *Chapter 2. The structure of the system*

The basic elements of the system are a number of primary registers. Part of a primary register makes a topic related data modules. The current statistics covering a specific traditional domain is produced on the basis the primary register concerned complemented if necessary with variables from data modules related to other domains of statistics to deduce new variables or as supplementary information. Integrated datasets are seen as logical links between these modules and they are the basis for production of transverse statistics, which typically is carried out by means of the so-called research databases, and other similar integrated datasets.

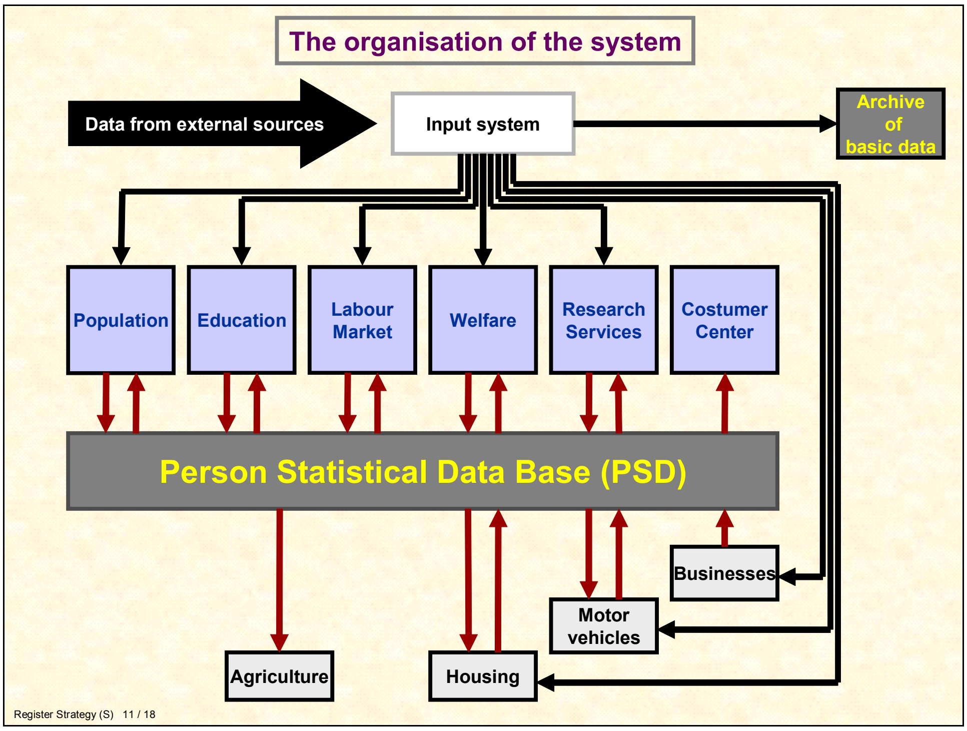


**Figure 1. The  structure of the register system**

The structure of the system is illustrated in figure 1. Data provided from external sources are processed together with data from internal data modules in preparation for control, deduction of new variables and addition of supplementary data. The results from these exercises are stored in one or more primary registers, each one of them refering to a specifik domain of statistics. Together these registers make the so-called Person Statistical Data Base or in short PSD. A data module is part of a primary register and it is available to be used in the integrated datasets and in other primary registers if required.

The reason for the distinction between primary registers and data modules is that only a limited part of the variables from a specific domain of statistics is in demand from other domains or from integrated datasets. So there ought to be a limitation of the possibilities to access primary data. It may seem needless to operate with the two kinds of variables but the data security and especially the outside evaluation of it makes this limitation necessary.

## *Chapter 2.1 The purpose and the content of the primary register*

A primary register is basis for a current primary statistics covering a specific domain. New variables associated with the statistics concerned may be deduced, and it provides data in demand for other statistics or integration projects.

The content of a primary register is the following types of data:

1. Data from external sources
    - used for the primary statistics concerned
    - auxiliary data for deduction of new variables
    - data to be investigated for possible suitability
2. New variables
    - deduced
    - standard concepts
3. Identification data
    - the unit of the primary register
    - links to data modules in other primary registers

Often you will find more than one data source to a primary register, and the data from the register may be spread to more than one data module. However a certain variable should not be stored in more than one data module.

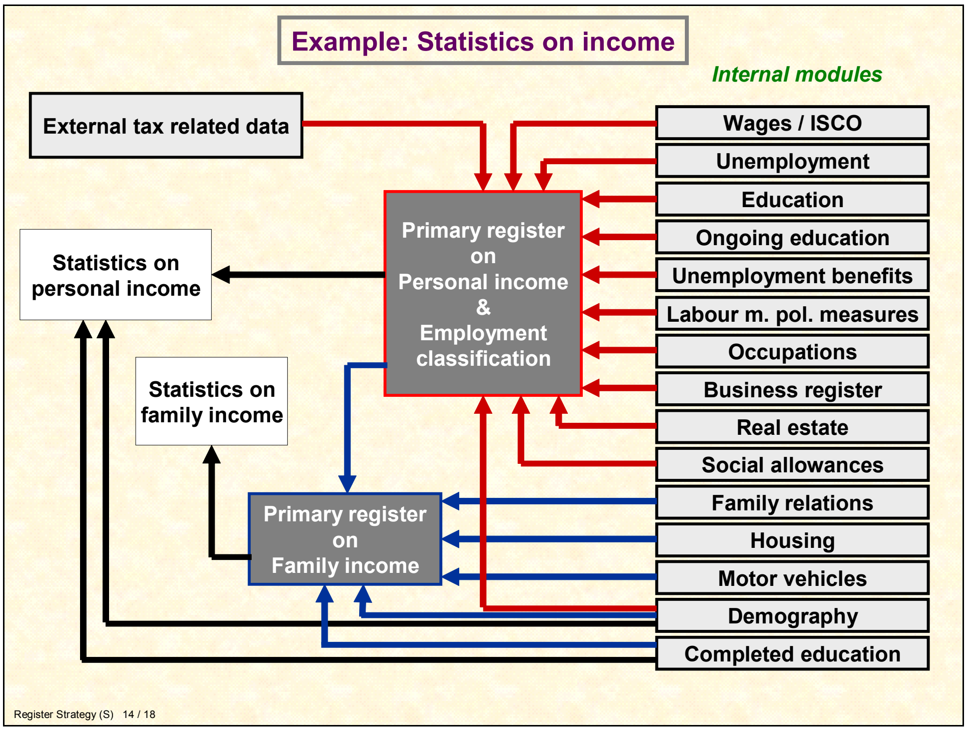

Example: Statistics on income
Internal modules
External tax related data
Wages / ISCO
Unemployment
Education
Ongoing education
Unemployment benefits
Labour m. pol. measures
Occupations
Business register
Real estate
Social allowances
Family relations
Housing
Motor vehicles
Demography
Completed education
Primary register on Personal income & Employment classification
Statistics on personal income
Statistics on family income
Primary register on Family income
Register Strategy (S) 14 / 18

**Figure 2. The subsystem on personal income and occupational classification**

Figure 2 shows as an example the sources used for the updating of the primary registers in the subsystem of personal income and occupational classification and the categories of data contained in these registers. Two primary registers have been established in this subsystem concerning occupational classification, personal income and family income.

Compared with the situation before the reorganisation the number of registers in this domain have been reduced due to a merging of three registers concerning personal income to one register and the number of variables will be reduced because of a more restrictive attitude to the selection of external data to be stored. The consequences of a model based on logical links between physical units (the data modules) will contribute to a further reduction in the number of data stored in the registers.

## Chapter 2.2 The purpose and the content of the data modules

A data module is part of a primary register. The data content is collected or deduced in relation to a specific domain of statistics. It provides data for other domains.

The role of a module will be:

- to provide data through direct access
    - supplementary data to other statistics
- to ensure quality through
    - use of harmonised definitions of concepts
    - correct choice of variables for the individual project
    - minimised delays by means of centralised updating
- to ensure transparency by means of
    - limited data redundancy
    - centralised documentation

The content of the individual data module is delimited to variables related to specific topics, which are used in different fields of statistics. For instance a module concerning health will only contain variables directly suitable for description of health conditions. Other variables must be picked up from other modules. The precise delimitation of a data module and with that the number of modules or even more detailed principles for that has not yet been decided. It depends on what might be appropriate taken the different requirements for integration of variables in various combinations into consideration. The modules must in some way be flexible in the way that in many cases the number of variables available will exceed the number required for a specifik project. That may be regarded as having negative effects, but the need for maintenance of a certain possibility for alternative selection of data and the need for a simple overall administration of the permissions of access dictate the flexibility.

Within the social statistics there is a need for different types of units. As far as the income statistics is concerned a data module on persons and a module on families have been established. The first mentioned is the main module and following the principles described above it aught to be the only module covering this field leaving statistics on families and households to be produced by means of the primary register combined with data from the module concerning household and families. However that would lead to repeated calculations of income for families, which is very irrational. So family variables are regarded as deduced variable gathered in a special family module. Other units can be events/activities for instance hospitalisations and employments.

It could be seen as an obvious possibility just to let the primary registers serve as data modules as well. However discretion considerations dictate restrictions on the access to data for the individual staff member to what is necessary to carry out the decided projects. Further there would be a risk that wrong data are chosen. So distinction has to be made between primary registers and data modules. They do not need to be and are usually not physically divided units. Primary users can have permission to access all data while the access of module users is limited to only a part of the data. By means of authorisation and control of access it is ensured, that the users only get the data they are entitled to process.

Figure 3 below illustrates as an example the relation between the data module and the corresponding primary registers concerning statistics on income and employment classification (EC). In the previous register system you found three traditional registers, an income register mainly based on data provided by the tax authorities, another income register including nontaxable incomes and a register containing employment classifications. The numbers of variables are respectively 341, 84 and 55 of which 417 are different. The new primary register on personal income will contain 322 variables and the EC 54 variables. The reduction is due to an avoidance of redundancy and omission of variables that are not used. The final income/EC data module includes 201 variables.

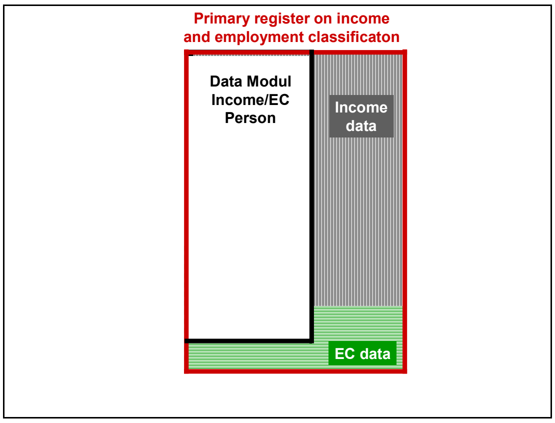
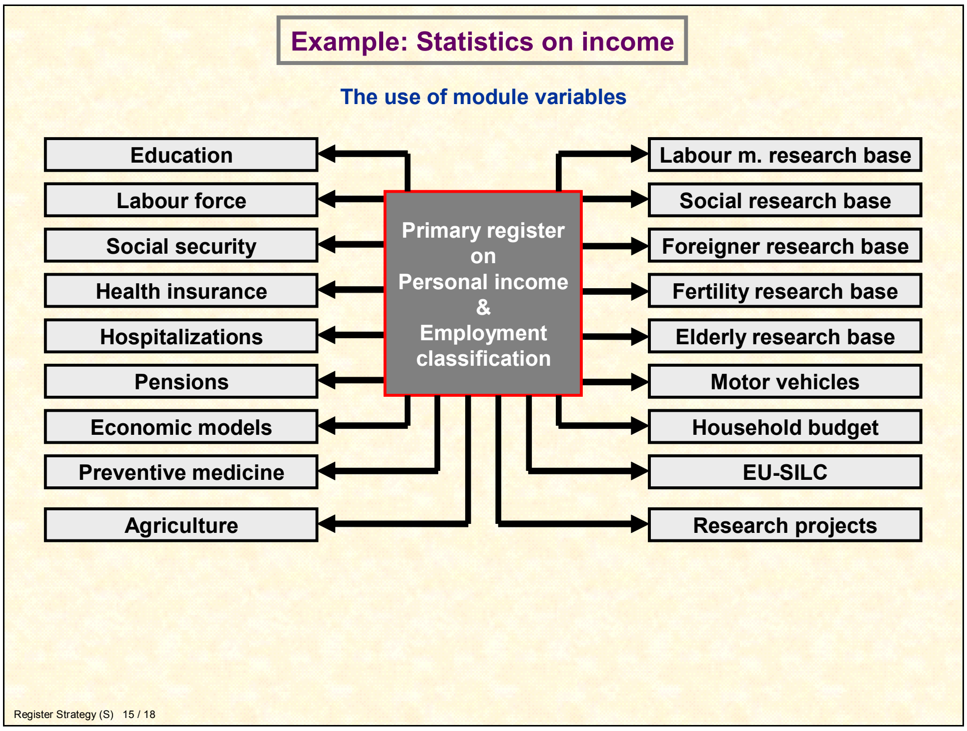


**Figure 3. The relation between a data module and the primary registers**

For reasons of flexibility and administrative simplicity permissions to access data modules are not differentiated. So the delimitation of variables in a module must be done with full attention to all the requirements that the module is supposed to comply with. The example concerning income/EC mentioned above aims to all applications of income and employment classification data in Statistics Denmark.



Example: Statistics on income

The use of module variables

Primary register on Personal income & Employment classification

Education
Labour force
Social security
Health insurance
Hospitalizations
Pensions
Economic models
Preventive medicine
Agriculture

Labour m. research base
Social research base
Foreigner research base
Fertility research base
Elderly research base
Motor vehicles
Household budget
EU-SILC
Research projects

Register Strategy (S) 15 / 18

**Figure 4. The current use of the Income/EC module**

Figure 4 illustrates the current use of the data module on income and employment classification in different fields of statistics. From this it appears that there is a very widespread use of income and employment classification variables. For some other data modules the picture will be the same.

The delimitation of the data content is of course dependent of confidentiality. No staff member should have access to unnecessarily many identifiable data. That principle shall continue to be in force. It might create some uncertainty outside Statistics Denmark if there is not a selective access to data at an acceptable level.

The frequency of updating a data module will as a principal rule be the same as the corresponding primary register, but it does not necessarily have to be so. It depends on a judgment of the expected use of the module. In this context it is very important to organise the modules in a way that leaves no doubt of the time reference of the variables contained in the module.

Certain domains outside the statistics on persons are included as well. In particular dwellings and businesses are concerned. Data modules will be established for these fields too. It is an organisational matter to decide how these modules should be connected to the system on persons.

# *Chapter 3. The future integration register*

The integration registers have been the basis for producing statistics across the primary statistics including research projects. Some registers are established as a general readiness for projects of that kind while other registers more directly aim at a definite problem.

In the traditional integration register all data are gathered and stored separately from the sources and they are used for the purposes that were the reasons for the establishment of the register and as supplementary data for other projects. It is a physical unit, which is manageable for the staff member responsible for the register. The access is rather simple for person authorised and it is easy to set up rules for dealing with the data. The responsibility for the maintenance of the register includes all data, data collected or deduced by the responsible himself and data provided from other statistical registers.

The integration register in the future will be a logical unit defined as a set of drawing rights to a number of data modules given to the staff member responsible for the project. Thus the delimitation of the content of a certain register is laid down by a specification of data modules accessible and the availability of the data / groups of data in these modules.

A certain variable must not be available from more than one source. So the responsibility for maintenance and documentation is settled. The access to data shall continuously be simple for persons authorised. Rules for dealing with data will be laid down with a clear reference to certain data modules.

The linking of data that occurs in the integration registers may lead to deduction of new variables. It could be one of the purposes of the linkage. Further there are situations where data are collected from external sources for direct use in an integration register. Finally the wide comparison of data may reveal errors in the data modules.

Deduction of new variables at the integration register level must result in the establishment of a new data module or extension of an existing primary register to form the basis for deducing the variable and provide it for further use through a corresponding data module. The last solution is preferable if it is possible to do it in a way, which is consistent with the organisational division of statistics and the placement of responsibility. Collected external data ought to be treated in the same way.

Discovery of errors must lead to correction of the data module concerned and in this way are the integration registers involved automatically updated. Users of the modules must be inform in order to explain the consequences of the correction for the statistical results up till now and for the use in the future.

Datasets that are deduced from an integration register must never be the basis for the updating of a data module. Neither shall such a dataset be a part of another integration register as a module. This limitation of possible data streams aims to avoid the uncertainty that could arise if data are not passed directly from the original source.

An integration register should be characterised by its technical simplicity from the user point of view. Steps must be taken to ensure that appropriate means for handling the data are introduced. Uniform definitions of the concepts and the degree of updating among the different registers are another characteristic and so is the quality of the documentation as well. Avoidance of parallel processing of data with the same purpose is an important advantage too.

# Chapter 4. Organisation and responsibility

The registers in the system of statistics on persons as it function to day are with few exceptions organisational placed in the divisions of the department *Social Statistics*. Among the exceptions are research and service related registers placed in the department *Customer Service*. In the department Business Statistics you will find a few other exceptions. The responsibility for the maintenance and use of the registers including the content and security is clear. It is assigned to the division to which the register belongs. A globalised documentation system developed during the recent years have to a certain degree moved part of the responsibility for the data documentation closer to the data capture.

It has been an essential disadvantage to the disposition of responsibility that it to a considerably extent was the physical location of data and with that the organisational composition that determines who is responsible for a certain dataset. For instance you found find many staff members responsible for income variables because that type of data are included in many registers. The often used method to transmit data from the most convenient source in stead of the original one means that the staff member who has collected or deduced the variables concerned could not be sure to know anything about the further use in other fields of statistics.

In the future system of primary registers, including data modules, a person responsible will be appointed for each of the primary registers and the associated modules. The responsibility shall cover maintenance of the register including deduction of variables according to standard definitions to be used in all relevant domains of statistics, documentation of one's own variable and authorisation of access to data modules.

Regarding access to data modules it is the person responsible for the register who assigns authorisations but a written request from the applicant endorsed by the head of his division is required. The person responsible enters assigned authorisations into a journal and the content of that will have to be confirmed once a year. In addition to that any attempt to access a data in a module shall be logged with information about user identification and time.

The authorisation to access a data module could include all variables, standard groupings of variables or selected variables. Regarding the data security and the general opinion about it dictate a precise selection of variables, while a demand for easy administration tend to allow access to a data module as a whole. An appropriate compromise could be to allow full access to small modules and a differentiated access to groupings of variables in big modules. Access to specific variables must be a clear exception and only used where data are particularly sensitive. For the data module income/EC mentioned above access could be allowed for many users to a group of selected broadly used variables and to the whole module for the limited number of user who require very detailed information about income.

# *Chapter 5. The implementation of the reorganisation*

The starting point of the reorganisation of the register system was a number of traditional registers each one of them containing all variables relevant for a certain field of statistics and some integration registers again containing all variables of interest for analysis related to a specific topic or a specific part of the population. All these registers are established as independent physical units. There is a rather comprehensive data redundancy, data transmissions go off through ingenious channels, and between some registers there is a certain parallelism in the updating processes. Most of the data processing was carried out on the mainframe.

The future technological conditions for data processing have changed because of the replacement of the mainframe in favour of PC/LAN, but the basic principles for a reorganised register system are independent of these conditions. During a period of transition you had to operate partly on the mainframe and some modifications, especially concerning organisation and storage of data were necessary in order to keep the register system workable.

In line with the development of data modules and the determination of their content the definition of concepts have to be considered in order to achieve a higher degree of uniformity between the different fields of statistics. An examination of income variables has revealed some differences which even that they are documented implies a certain risk of misinterpretation or create confusion among the users of statistics.

Income statistics was one of the first domains of statistics to be reorganised and it was among the first registers to be transferred to PC/LAN as well. There are quite a lot of variables contained in the income register so it is not typical of registers in Statistics Denmark but the variables are very broadly used and many users will be affected by the reorganisation. In this way it contributes to create attention to the construction of the register system.

At first the reorganisation aims at future register versions but variables related to previous years and still used must be organised and processed in the same way. The questions are how much effort should be spend on harmonisation of concepts related to previous years and how do we handle differences over time.

All historically data are removed from the mainframe by now but there is still quite a lot of work to do organising data according to the register strategy and making exchange of data between domains possible in a easy and smoothly way.