Working Paper No. 4
ENGLISH ONLY

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE-EUROSTAT Work Session on Registers and Administrative Records for Social and Demographic Statistics
(Geneva, 9-11 December 2002)

Supporting paper

A STATISTICAL SYSTEM OF REGISTERS BASED ON LOGICAL DATA CONNECTION

Submitted by Finn Spieker, Statistics Denmark

I. THE NEED FOR REORGANISATION OF THE REGISTER SYSTEM

1. The current Danish register system for statistics on persons is to a very great extent influenced by the stepwise development during the last thirty years. The number of data has increased heavily and the use of general identifications, mainly the PIN-code, has led to numerous possibilities for producing statistics based on different combinations of variables. Researchers and others are very much aware of this situation and they are keen to utilize the possibilities of analysing specific topics in the different fields of living conditions.

2. Today, the data are organised in registers as physical units of the statistical system and still more registers have been created totally or partially containing variables found in other registers. This was done for different reasons, but the system as a whole is now so large that hardly anybody can cope with it, and the continuing extensions that are expected will contribute to further complications. That is the reason why reorganisation of the register system has been discussed in Statistics Denmark and is now underway.

3. The results of the reorganisation are expected to 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 individual cases;
- 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.

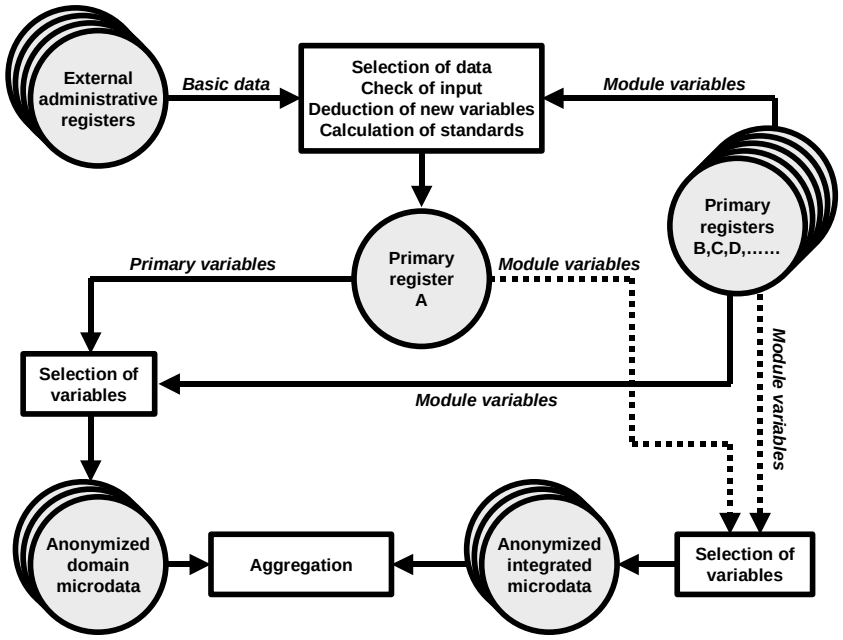
4. These improvements will contribute to ensuring the quality of statistics and a more optimal use of the resources.

II. THE STRUCTURE OF THE REGISTERS

5. The basic elements of the system will be a number of primary registers serving as physical units. Each one of these registers includes a topic-related data module. Besides these registers, integration registers are defined as sets of logical links between the modules. They form the basis for the production of transverse statistics, which typically is carried out by means of the so-called research databases and other similar integration registers. The current statistics covering a specific traditional domain are produced on the basis of the primary register concerned, complemented if necessary with variables from data modules related to other fields of statistics in order to deduce new variables or as background information.

6. 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 complementary data. The results of these exercises are stored in one or more primary registers, current statistics are compiled and the module part of the register is updated. The data modules are available to be used in the logical integration registers and in combination with other primary registers if required.

Figure 1. The logical structure of the registers



7. The reason for the distinction between primary registers and data modules is that only a limited number of the variables from a specific field of statistics is requested by other fields or by integration registers. So there ought to be a limitation of the possibilities to access primary data. It may seem needless to operate with the module as a kind of subregister but the data security and especially its evaluation on the exterior necessitates restrictions on access. Figure 2 illustrates the data modules and the logical superstructure to a number of integration registers of which each one consists of a set of well-defined links to the data modules. The data modules will make up the physical elements of the integration registers. The content of them decide what is available for integrated statistical projects.

Figure 2. The integration register model

	Integr. reg. A	**Integr. reg. B**	**Integr. reg. C**	**Integr. reg. D**
Data module 1	**X**			
Data module 2		**X**		
Data module 3			**X**	
Data module 4				**X**
Data module 5	**X**	**X**	**X**	**X**
Data module 6	**X**			**X**
Data module 7		**X**		**X**
Data module 8	**X**		**X**	
Data module 9		**X**		**X**

III. THE PURPOSE AND THE CONTENT OF THE PRIMARY REGISTERS

8. A primary register is a physical unit that, together with data from the module part of other primary registers, forms the basis for current primary statistics covering a specific domain. Further new variables associated with the statistics concerned may be deduced, and finally the primary register provides data for the updating of data modules in demand for other primary statistics or integration projects.

9. The content of a primary register includes the following types of data:

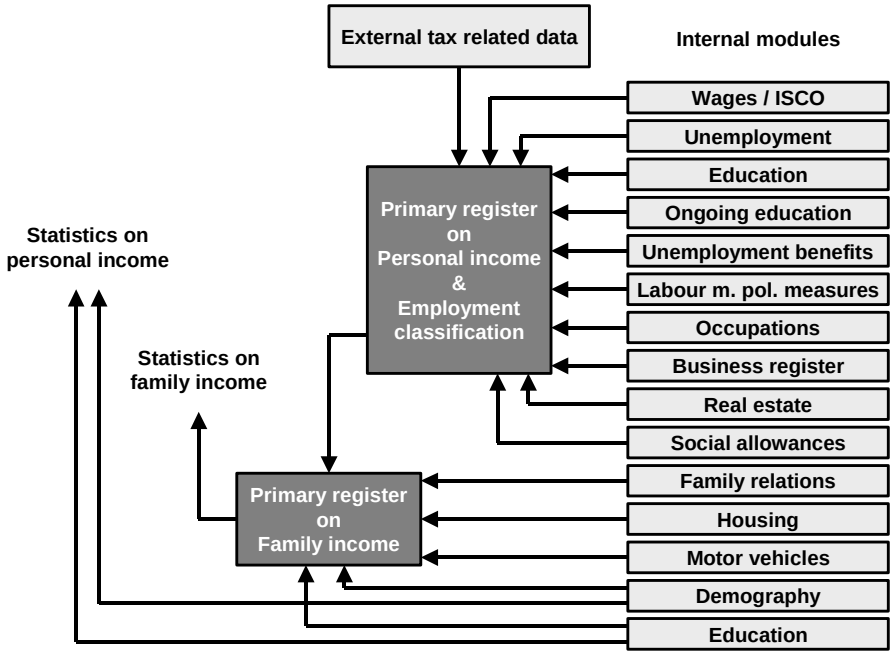
Data from external sources
- used for the primary statistics concerned;
- required for one or more integration registers;
- auxiliary data for deduction of new variables;
- data to be investigated for possible suitability;
- data for control purposes.

New variables
- Deduced;
- grouping standards of variable values.

Identification data
- the unit of the register;
- links to other registers.

10. It may be that more than one primary register makes use of the same external input. In this case it must be ensured that a given variable only is available from one data module.

Figure 3. The function of the primary registers on personal income / occupational classification and family income



11. Figure 3 shows as an example the sources used for the updating of the primary registers in the subsystem of income and employment classification and the categories of data used to maintain these registers. It is proposed to operate with two primary registers in this system, the first concerning personal income and employment classification and the second family income. From the figure you will see that data from two data modules are used in the production of statistics on personal income without being included in the primary register. The education data included are used for deduction of new variables.

12. This represents a reduction compared with the existing number of registers in this field, due to the merging of three registers concerning personal income into one register. 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.

IV. THE PURPOSE AND THE CONTENT OF THE DATA MODULES

13. A data module is a logical part of one or more integration registers and it provides complementary data for other primary statistics. The data content is collected or deduced in relation to a specific field of statistics.

14. The role of a module will be:

- To provide data
  through direct access
  - complementary data to primary registers
  - module in integration registers

- Ensure quality through
  use of harmonised definitions of concepts;
  correct choice of variables for the individual project;
  minimised delays by means of centralised updating

- Ensure transparency by means of
  limited data redundancy;
  centralised documentation.

15. 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 the description of health conditions. Other variables must be picked up from other modules. The precise delimitation of a data module, the number of modules or even more details have not yet been decided. It depends on what might be appropriate, taking into consideration the different requirements for integration of variables in various combinations. Somehow, the modules must be flexible in such a way that in many cases the number of variables available will exceed the number required. This might be regarded as having negative effects, but the need to maintain a possible alternative selection of data and the need for a simple overall administration of access authorization dictates the flexibility.

16. The demands on the content of a data module are as follows:

Uniform units must be used;
The variables shall make a subset of the variables in primary register to which they are related;
The variables are used for at least one of the following purposes:

- to be part of an integration register;
- to compliment a primary register;
- the basis for creation of deduced variables;
- planned to be used in an innovation project;
- to identify the units of the module;
- to identify links between units in the module;
- to identify links between units in the module and in other modules.

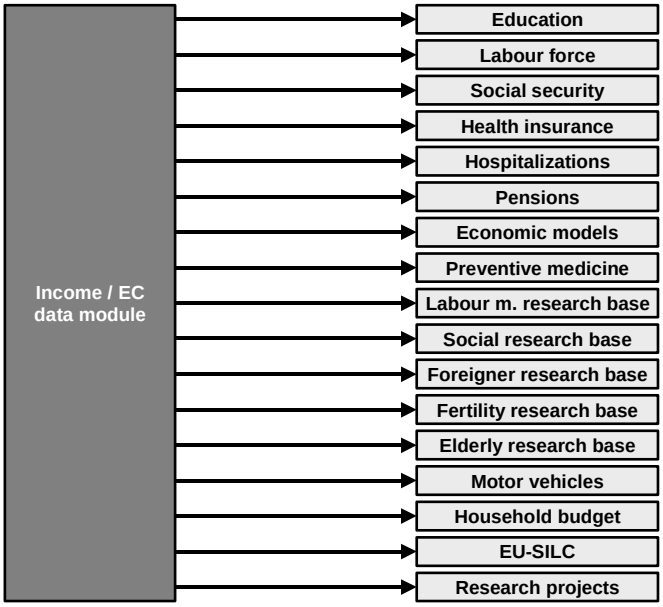
17. Within the statistics on persons, there is a need for different types of units. As far as the income statistics are concerned, a data module on persons and a module on families is proposed. The former is the main module and, following the principles described above, it ought to be the only module covering this

field, leaving statistics on families and households to be produced by means of the main module combined with data from the module concerning household and families. However, that would lead to repeated calculations of income for families, which ought to be avoided. So family variables are regarded as deduced variables gathered in a special family module. Other units can be events/activities, for instance hospitalisations and employments.

18. An obvious possibility would be to just let the primary registers serve as data modules as well. However, discretion limits the restrictions on access to data for the individual staff member to what is necessary to carry out certain projects. Furthermore, there would be a risk of incorrect data being chosen. So a distinction has to be made between primary registers and data modules. They do not need to be 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 can only obtain the data they are entitled to process.

19. For reasons of flexibility and administrative simplicity, authorisation to access data modules will only be slightly differentiated. Therefore, the delimitation of variables in a module must be undertaken with full attention to all the requirements with which the module must comply. The example concerning income/EC mentioned above is aimed at all applications of income and employment classification data in Statistics Denmark.

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



20. The content of a data module must be considered and settled individually. It is a matter of the number of variables and the domain of statistics to be covered. The former must be based on the requirements and must follow the principles described above, while the delimitation of the field is more dependent on organisational matters, including the responsibility for the maintenance and documentation of the module combined with professional qualifications.

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

22. The delimitation of the data content is, of course, dependent on confidentiality. No staff member should have unnecessary access to too many identifiable data. This principle will continue to be applied. It might create some uncertainty outside Statistics Denmark if selective access to data is not maintained at an acceptable level.

23. The frequency of updating a data module depends on the estimated expected use of the module. In this context it is very important to organise the modules in a way that leaves no doubt as to the time reference of the variables contained in the module.

24. The period covered by the data modules must follow the rules for storage of confidential data. For every version of a module, it must be decided how long time it will be stored in Statistics Denmark and whether it will be passed on to the national record office at the end of the period.

V. THE PURPOSE AND THE CONTENT OF THE INTEGRATION REGISTERS

25. The integration registers will be the basis for producing statistics across the primary statistics, including research and explanatory projects. Some registers are established in general readiness for projects of this kind, while other registers aim more directly at a definite problem.

26. 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 complementary data for other projects. The integration register is a physical unit, which is managed by the staff member responsible for the register. Access is simple for the person authorised and it is easy to set up rules for the handling of the data. The register includes all data collected or deduced by the responsible person and data provided from other statistical registers.

27. In future, the integration register 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 the data modules accessible and the availability of the data / groups of data in these modules.

28. Any variable must be available from only one source. So the responsibility for maintenance and documentation is clear, and access to data will be simple for the authorised persons. Rules for dealing with data will be determined with a clear reference to certain data modules.

29. The linking of data that occurs in the integration registers may lead to the deduction of new variables. This could be one of the purposes of the linkage. 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.

30. 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 providing 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 assignment of responsibility. Collected external data should be treated in the same way.

31. Discovery of errors leads to correction of the data module concerned and, in this way, the integration registers involved are automatically updated. Users of the modules must be informed in order to explain the consequences of the correction for statistical results to date and for use in the future.

32. Datasets that are deduced from an integration register should never be the basis for the updating of a data module. Neither should 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.

33. An integration register should be characterised by its technical simplicity from the user's 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 same degree of updating among the different registers are another characteristic, as is the quality of the documentation. Avoidance of duplicate processing of data is an important advantage.

VI. INCLUDED REGISTERS

34. The model involving the establishment and maintenance of primary registers should be applied to all fields of statistics on persons, based on identified individual data collected from external sources or deduced from a combination of data. The primary registers replace the subject-oriented statistical registers. The content of the physical unit is limited to what could be called one's own data and, by means of links to different data modules, complementary data are gathered, making it possible to complete the current statistics for a specific field.

35. Certain fields outside the statistics on persons are included as well. Dwellings and businesses are particularly concerned. Data modules must be established for these fields too. It is an organisational matter to decide how these modules should be connected to the general system.

36. Four types of integration registers can be distinguished according to their purpose:

- Production of statistics and analyses initiated by Statistics Denmark;
- Production of statistics initiated by Statistics Denmark combined with readiness for research projects (Semi research databases);
- Readiness for research projects (Multi research databases).
- Research registers established for a specific project (Single user databases)

37. The distinction between primary registers and integration registers, established only to serve as the basis for the production of statistics on the initiative of Statistics Denmark, is not evident. In both cases, the integration of data is involved but the functions of processing data from external sources, deduction of new variables and updating of data modules ought to be attached specifically to the primary registers.

38. Micro datasets deduced from integrations registers for calculation of tables or for research projects are still created as physical units.

39. Survey data are treated in the same way as register data involving the establishment of primary registers with links to data modules. Due to the limited samples of these registers, there will hardly be a need for creation of data modules covering survey-based variables.

VII. ORGANISATION AND RESPONSIBILITY

40. The registers of the system of statistics on persons as they function today are, with few exceptions, located in divisions of the Department of Social Statistics. Among the exceptions are research related registers, located in the Division of Research and Method in the User Service Department, and various other exceptions in the Department of Business Statistics. 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 global documentation system developed during recent years has, to a certain degree, moved part of the responsibility for data documentation closer to data capture.

41. An essential disadvantage of the present disposition of responsibility is that, to a considerable extent, it is the physical location and organisational composition of data that determines who is responsible for a certain dataset. For instance, many staff members are 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 instead of the original one means that the staff member who has collected or deduced the variables concerned has no clear knowledge about the further use of those variables in other fields of statistics.

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

43. With a register system based on the common use of a number of topic-specific data modules, it is difficult to define an exact responsibility for the integration registers. It is more relevant to talk about product liability related to the utilization of the data composition, which is a specific combination of variables for each integration register. The results of the utilization may be tables calculated to be published or to be at the disposal of service customers, or it could be the formation of an anonymised micro dataset to be used for research or similar projects. Liability could be assigned to a project group if this form of organisation is used, but one staff member should be the principal responsible person.

44. The primary registers are very similar to the integration registers. They are concerned with the production of statistics based on one's own data combined with data from other fields. So product liability could be applied to these registers as well.

45. 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 data in a module shall be logged with information about user identification and time.

46. The authorisation to access a data module could include all variables, standard groupings of variables or selected variables. Considerations concerning data security and the general opinion about it dictate a precise selection of variables, while a demand for easy administration tends 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.

VIII. THE IMPLEMENTATION OF THE REORGANISATION

47. The starting point of the reorganisation of the register system is 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 is carried out at the mainframe.

48. The future technological conditions for data processing will change because of an upcoming phase out 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 will have to operate partly on the mainframe and some modifications, especially concerning organisation and storage of data, may be necessary in order to keep the register system workable until the end of the transition.

49. The distinction between the logical links and the physical delimitation of a register may cause some troubles during the transition period, but you can as a principal rule stick to the principle that passing on variables received from a data module to other registers should not be allowed. Complementary variables for a primary register and variables for an integration register ought to be provided directly from the data modules concerned.

50. 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.

51. Income statistics is one of the first fields of statistics to be reorganised and it will be 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 the 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 reconstruction of the register system.

52. At first the reorganisation aims at future register versions starting when it is effective 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.

IX. THE INCOME/EC EXAMPLE

53. Having examined the Income/EC some principal conclusions can be drawn as regards the content of variables that can be basis for the considerations concerning the content of the system:

- 73 out of 417 person variables are not used at all
- 172 person variables are used outside the field of income statistics
- some variables enter into sum concepts named alike but with varying definitions
- almost no use of 90 family variables outside the field of income statistics

54. The 73 variables not used are part of the basic data provided by the tax administration. It is data, which at the time of delivery are doubtful as regards their further use, or data provided twice from different tax administrative registers. Variables of this kind will not be excluded from the basic data sets but they will not be passed on in the statistical system.

55. 172 variables are used outside the income statistics but the Income/EC module is supposed to contain 201 variables. 9 of the 172 variables appeared not to be used. Further were 5 variables included which according to the principles as stated above ought to be provided from other data modules. So it is proposed to enlarge the Income/EC data module with 43 variables beyond the use observed. It is mainly a matter of making standard sum variables available.

56. A limited use of family variables was observed. It may be ascribed to a possible creation of this type of variables by the users themselves based on person variables. So it is proposed in addition to the module mentioned above to establish a family version of the Income/AC data module containing 65 variables despite the apparently limited use of such variables. It will mainly be about sum variables based on person variables.

57. The Income/EC system is characterised by using a very comprehensive amount of data from external sources and deduction of key variables where external data are combined with data from a number of internal sources. Further there is a very widespread use of the variable in the statically system. There are similarities between Income/EC and other parts of the register system and there are some differences as well, but Income/AC may be a model for the arrangement of other parts of the system as part of the reorganisation of the register based statistical system on persons.

- - - - -