United Nations
Statistics Division

ESA/STAT/AC.88/06
24 April 2003

English only

Expert Group Meeting on
Setting the Scope of Social Statistics
United Nations Statistics Division
in collaboration with the Siena Group on Social Statistics
New York, 6-9 May 2003

The Dutch System of Social Statistics: Micro-Integration of Different Sources*

by

Pieter C.J. Everaers and Paul Van Der Laan**

* This document is being issued without formal editing.

** Director, Division for Social and Spatial Statistics and Senior researcher, international relations, Division for Social and Spatial Statistics respectively. The views expressed in this report are those of the authors and do not imply the expression of any opinion on the part of the United Nations Secretariat.

The Dutch system of social statistics: micro-integration of different sources

Pieter C.J. Everaers[1] and Paul Van Der Laan [2]
Statistics Netherlands

1. Introduction

Growing demand for coherent and detailed statistical information, developments in information and communication technology, growing non-response rates in surveys and political pressures to cut down cost and to minimise the reporting burden caused by official statistics are the main reasons that traditional statistical methodology will not be able to continue to meet the demands facing official social statistics in the future. The design and organisation of statistical production processes as well as the organisational structure of the statistical offices will have to be changed radically.

Designing, conducting and processing social surveys is a very time-consuming and expensive process and places quite a burden on society. Added to this, companies and the public at large are less and less willing to participate in surveys. A logical alternative would be a large-scale statistical recycling of information available in society initially intended for other purposes. Statistics Netherlands (SN) has adopted a basic principle to collect its own data only if this information is not available from other sources (Statistics Netherlands 2001). This is possible because of the increasing availability of a growing number of administrative sources and registers, all containing various information on persons and households that not only for administrative purposes also – sometimes with some small corrections and additions - can be used for official statistics. With this principle statistical offices will meet the widely felt need to reduce the response and administrative burdens and to increase cost-efficiency. In recent years many statistical offices, especially in the more developed countries, have gained experience in the use of administrative databases.

Essential in the combined use of different sources is the effort put in the harmonisation of concepts and classifications. Harmonised statistics by input (same data collection methods, questionnaires, etc.) as well as by output (i.e. based on the same concepts, definitions and classifications) will be the main source for European social statistics and (for example in the field of income, Canberra group) western social statistics as a whole. Harmonisation is mainly reached via theoretical discussions in (inter)national discussion groups, as is the implementation procedures via regulations and gentlemen's agreement procedures. All these efforts are mainly driven by harmonising

1 Director, Division for Social and Spatial Statistics
2 Senior researcher, international relations, Division for Social and Spatial Statistics

concepts in the input sources of the statistical process (registers, questionnaires). However, next to the theoretical considerations technical (IT) developments also play an important role in facilitating as well as forcing harmonisation further.

2. Statistics Netherlands has made several steps to improve the coherence in its statistics.

In the late eighties the development of Blaise as the main tool in computer-assisted (personal and telephone) interviewing forced the harmonisation of standard blocks in questionnaires. Micro integration forced researchers to make concepts with regard to persons and households in business surveys comparable with those from household surveys. In the early nineties the first steps were made in the more extended use of administrative sources: the municipal population administration allowed Statistics Netherlands to link demographic register information with information stemming from specific surveys. As a consequence of the integration of the whole set of separate isolated surveys into one integrated modular survey on living conditions an important step was made for further integration.

By linking these administrative databases with each other and with data from specifically targeted surveys as well as the modular general survey on living conditions, Statistics Netherlands has created in the late nineties initial versions of a *Social Statistics Database* (Arts and Hoogteijling 2002). This database has revealed the contours of a new way of producing social statistics: the transformation of social statistics from a wide variety of largely isolated and usually expensive single statistics into an integrated, cost-efficient statistical system. Statistical integration of different data sources at the micro-level is as an essential tool to produce these databases (Van der Laan 2000; Bakker 2002). The paper will demonstrate that this approach to social statistics will improve the coherence and detail in our statistical output and will increase the efficiency of the statistical process and in that way provide the best value for money.

The development of the Social Statistical Data Base (SSB) initiated other developments in achieving consistency. The system of repeated weighting, as implemented in the program Bascula, allowed the combination of data from large (population) registers and surveys. Such a system asks for high computing capacity. Combining data from these sources forces researchers to make concepts in surveys comparable with those used in registers (and as far possible vice versa). The recent reorganisation of Statistics Netherlands focuses on comparability in processing the data. Input data bases, baselines in different stages, are considered the main in between products of the statistical process. The Social Statistical Data Base (SSB) is seen as the resulting database using all the combined data from surveys and registers. In this data base in principle every event can be measured via several original sources, the figure for a specific event in the data base combines the good elements of the via several methods measured or observed event. Data matching and linking techniques further force harmonisation.

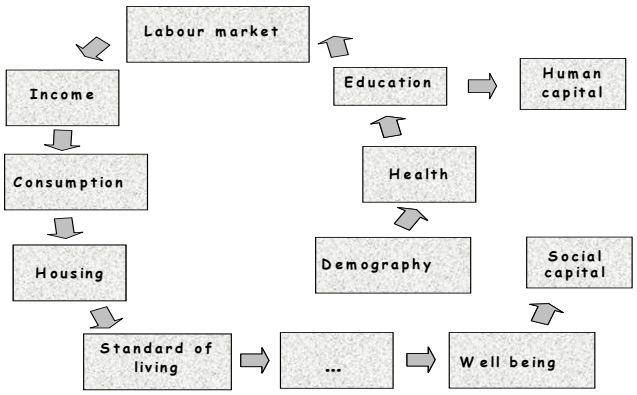
The introduction of computer-assisted interviewing, the advanced (re)weighting procedures, the matching programs and data storing systems like those of data warehousing have all forced harmonisation of concepts and classifications.

3 Integrated micro-data files in social statistics

The Social Statistics Database (SSD) will contain all the relevant information on persons, families, households, jobs, benefits and living quarters, from which consistent statistical outcomes can be produced with more regional detail and with more information on specific population groups in society. As the conceptual model behind the SSD the 'life cycle' model of a person is used (Figure 1). This model allows a good modelling for the main government interventions that unequal opportunities in the life cycle may cause. Interventions that normally in the western societies are initiated by the national governments and that are considered important criteria for the performance of the governmetal system in the domains of the living conditions.

Based on this model the SSD will cover all aspects of a person's life cycle and can be used to analyse socio-demographic, socio-economic and socio-cultural situations of population groups, transitions of persons between situations and spells of situations.

Figure 1. The 'life cycle' model



The SSD is primarily based on register information and data from business and household surveys which are not available in registers. The registers frequently contain complete information on all relevant units. In the Netherlands, this is surely the case for demographic data, income tax data, the participation in the labour market, the dependence on social security benefits, the participation in education and housing facilities. The files of the Population Register form the backbone of the database, as all the other files are linked to this register (Arts et al. 2000).

The combined register data are used as a sample frame for surveys, in order to collect the survey data in a most efficient way. For that purpose, the method of pre-stratification is used. If e.g. information is needed on poverty, low income households are over-represented in the sample. This is possible as the sample frame contains information on household income.

One of the problems with administrative records is that they are delivered to Statistics Netherlands with a delay of sometimes more than two years. Although household surveys will only collect information that is not available in registers, register information that has a serious delay and needs to be published timely, is also collected through surveys.

The completeness of the information in the registers brought in the idea of *micro-integration* of data. Linking records from different sources provides a check on the completeness of the registers and the occurrence of double records. If different sources contain information on the same variable, consistency checks are made and correction procedures are followed. This kind of data editing leads to consistent statistical information for those variables that are processed. It makes macro-integration later far easier, as a large part of the inconsistencies is solved at the beginning of the process. After linking, the statistical variables have to be derived from the characteristics recorded in the linked records. In cases of missing data for smaller fractions of the total population, imputation can be used as a form of integration to arrive at completeness.

High non-response makes survey estimates questionable, because it introduces a potential bias that is difficult to measure. One of the possibilities of the SSD is to adjust more effectively for selective non-response. Linking administrative registers with survey information makes it possible to search for the characteristics that correlate highly with the probability of response and with the target variables in the survey. Furthermore, it is possible to select those characteristics to weight the survey. It is evident that the use of more register information to weight the data has serious advantages over traditional methods to reduce non-response bias. This method results in the starting weights for further reweighting procedures.

After collecting, linking and editing the data, it is necessary to estimate the frequencies and cross-tabulations which are to be published. For an output database like the Dutch StatLine to function properly it is important to enter consistent results from different sources. The method of consistent and repeated weighting is used for that purpose (Kooiman et al. 2000). The cross-tabulations are estimated in such a way that they are consistent with the marginals of earlier produced tables. Therefore, the starting weights from the non-response bias reduction method are calibrated.

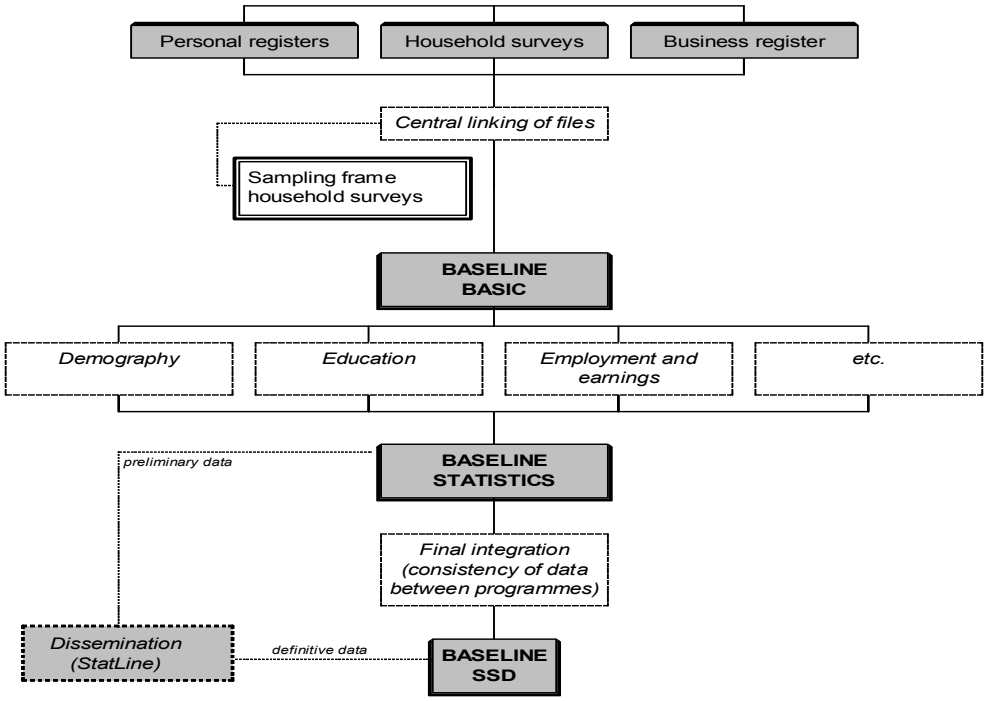
One of the advantages of the SSD approach is that a large part of the integration process that was done so far on a macro-level will be performed on a micro-level. However, macro-integration will still be necessary, e.g. for the consistency with National Accounts or with the outcomes of business surveys. For example, the total expenditure of government on social benefits must equal the total of benefits received by individual

persons. In some cases the macro-totals can be used as a restriction in the production of the individual weights.

4. System of baselines

Besides data linking, data sharing is one of the key elements of a SSD-based production process. To facilitate the process of data sharing and integration a system of so-called 'Baselines' is developed (Figure 2). This system shows that statistics and information technology are more or less equal. In the SSD production process three different baselines are distinguished. The first one 'Baseline basic' is the result of all linked micro-data from administrative sources and surveys. These data sets are not yet analysed and contain only the most basic edits. So, baseline basic is still an administrative register. The basic baseline covers monthly, quarterly and annual data based on daily, weekly, monthly, quarterly and yearly data flows.

Figure 2. The system of 'Baselines'



The next baseline is 'Baseline statistics'. This baseline contains all the edits on the original data of baseline basic. The edits are made by the different programme units, such as demography, education, labour force, employment and earnings, social security, income and consumption and housing. Baseline statistics also includes the transformation from baseline basic data to statistical variables. Baseline statistics is a statistical register used to publish preliminary, timely output.

The final baseline is 'Baseline SSD' which includes the data after final integration of all data sources. The definitive statistical results are based on baseline SSD; baseline statistics only produces provisional results. Baseline SSD takes more time to construct than baseline statistics. The data in baseline SSD have a high accuracy, the data however are usually not very timely. From baseline statistics more timely data can be produced, but these data are preliminary and may be revised in baseline SSD. The output of baseline statistics and baseline SSD is entered into the output database StatLine. The input of StatLine is based on 'data cubes', a large multi-dimensional table containing all relevant publishable data on a specific theme. With the help of a menu the user can select the ranges of the table he or she is interested in.

5. Conclusions

The design and organisation of statistical production processes have to change radically. These changes are triggered by the growing demand for coherent and detailed statistical information, by developments in information and communication technology, by growing non-response rates in household surveys and by political pressures to cut down staff cost and to minimise the reporting burden caused by statistics. The paper has argued that the statistical integration of different data sources (both administrative registers and household surveys) at the micro level is an essential tool to meet the new circumstances. As statistical information systems should provide accurate, relevant and authoritative information at the lowest possible costs, the transformation of social statistics from a wide variety of largely isolated and usually expensive statistics into an integrated, cost-efficient statistical system is the logical consequence of these prerequisites. After all, it is in the common interest that the goals of government policy are carried out on the basis of the best available information with the lowest cost to produce this information.

REFERENCES

Arts, C.H., B.F.M. Bakker and F.J. van Lith. 2000. 'Linking Administrative Registers and Household Surveys'. *Netherlands Official Statistics*, Vol. 15 (Summer 2000): Special Issue, *Integrating Administrative Registers and Household Surveys*, ed. P.G. Al and B.F.M. Bakker, pp. 16-22.

Arts, C.H. and E.M.J. Hoogteijling. 2002. 'The Social Statistics Database 1998 and 1999'. *Monthly Bulletin of Socio-economic Statistics*, Vol. 19 (December 2002): pp. 13-21. *[in Dutch]*.

Bakker, B.F.M. 2002. 'Statistics Netherlands' Approach to Social Statistics: The Social Statistical Dataset'. *The Statistics Newsletter for the extended OECD Statistical Network*, Issue No. 11 (October 2002): pp. 4-6.

Kooiman, P., A.H. Kroese and R.H. Renssen. 2000. *Official Statistics: An Estimation Strategy for the IT-era*. Statistics Netherlands, Division for Research and

Development, Research Paper No. 0018. Voorburg: Statistics Netherlands, Division for Research and Development. May 2000.

Laan, P. van der. 2000. 'Integrating Administrative Registers and Household Surveys'. *Netherlands Official Statistics*, Vol. 15 (Summer 2000): Special Issue, *Integrating Administrative Registers and Household Surveys*, ed. P.G. Al and B.F.M. Bakker, pp. 7-15.

Statistics Netherlands. 2001. *Statistics That Count: Strategic Plan for the Medium Range, 2002-2005*. Voorburg and Heerlen: Statistics Netherlands. May 2001.