



**Economic and Social
Council**

Distr.
GENERAL

CES/AC.71/1999/9
19 November 1998

ENGLISH ONLY

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Meeting on the Management of Statistical Information Technology
(Geneva, Switzerland, 15-17 February 1999)

Topic (i): The impact of Internet on the statistical production and dissemination process

**BETWEEN ELECTRONIC INPUT AND ELECTRONIC OUTPUT -
SOME THOUGHTS ON STATISTICAL PROCESSING BEYOND 2000**

Submitted by Statistics Netherlands¹

I. STATISTICS AND TECHNOLOGY: INPUT

1. The traditional approach to producing statistics is to collect data by means of surveys using paper questionnaire forms. For household and personal surveys, interviewers are used to obtain the answers to the questions, whereas establishments surveys are usually of the mail out / mail back type.

2. Technological developments (including the Internet) triggered a change from paper-streams into bit-streams. The paper questionnaire is gradually replaced by EDI (Electronic Data Interchange). For example, NSI's use EDI to collect data from the electronic data systems of establishments (including enterprises and government/public sector agencies). This totally new way of data collection requires Business Process Redesign (BPR) at NSI's.

3. Already in the eighties, BPR activities were initiated in NSI's as a consequence of the introduction of computer-assisted interviewing techniques

¹ Prepared by Wouter J. Keller and Jelke G. Bethlehem.

such as CAPI (Computer-Assisted Personal Interviewing with laptops) and CATI (Computer-Assisted Telephone Interviewing) [see e.g. Bethlehem (1995)]. Paper questionnaire forms were replaced by electronic forms, using laptops/PC's with software such as Blaise. This mainly affected household surveys. Several tasks in the survey process (interviewing, data entry and data editing) were integrated. We call this task integration.

4. The change to computer-assisted interviewing made it also possible to integrate different surveys in one combined electronic questionnaire instrument. An example at Statistics Netherlands is the POLS project, which combines five household surveys into one. This is survey integration.

Figure 1. Data collection at SN



5. Survey integration is also being implemented in establishment surveys. EDI is increasingly used to collect data from establishments. Software, e.g. EDISENT [see De Bolster, 1997], automatically retrieves data collection from financial bookkeeping systems of enterprises. The focus is on the data source (e.g. financial systems) instead of on the surveys. The collected data is used for several statistics. Combining tasks at the enterprise (data entry and editing, including the automatic translation of financial and administrative concepts to statistical concepts as required by the electronic questionnaire) constitutes a form of task integration. This is electronic data collection still involving respondents, and therefore it is called primary EDI.

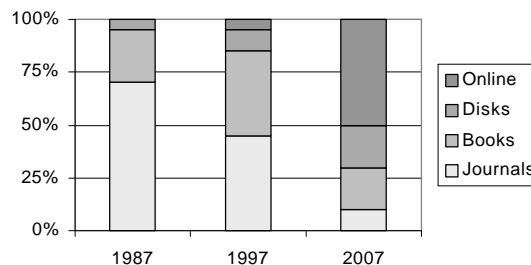
6. A next step in the evolution is data collection from secondary sources. This means using electronic administrative data sources (mainly public registers) like those of the IRS and social security administration. This is called secondary EDI. Usually, information of primary EDI sources (respondents) and secondary EDI sources (registers) must be combined to obtain the required statistics.

7. The change from paper questionnaire data collection to primary and secondary EDI is illustrated in Figure 1. This figure contains a prediction for 2007.

II. STATISTICS AND TECHNOLOGY: OUTPUT

8. In the past, there was a separate set of (paper) publications for each statistical survey. Most of these publications consisted of tables with lots of numbers. They were printed and published at regular intervals, e.g. monthly. The organization of activities was based on internal differentiation with respect to the statistics to be published. This organizational structure can be characterized as "stove-pipes". There was not much statistical integration of publications over different surveys (except for National Accounts - NA), nor was there complete statistical co-ordination of common concepts and classifications. Customers looking for consistent information on a topic, say housing, were forced to consider many different (paper) publications (e.g. on production, labour, use of houses), often with inconsistent concepts (number of employees with and without part-timers) and different classifications (e.g. for products and branches).

Figure 2. Dissemination of statistical publications at SN



9. It is clear that the main future medium for dissemination of statistical publications will be the Internet and other electronic media such as e-mail subscriptions, CD-ROM, etc.. Of course, paper will always be an important medium, particularly for thematic publications including a lot of visual elements like charts and photographs. However, most statistical figures will come from statistical datawarehouses, freely accessible on the Internet as a public service by the NSI's. Integrated statistical databases will result in a limited number of consistent, integrated views (so-called datacubes). An example is the IT literature on OLAP and datawarehouses, and IJsselstein (1996). Figure 2 displays the trend in the change from paper to electronic publications. The figure for the year 2007 is a prediction.

10. The datawarehouse is a collection of datacubes. The dimensions (axes) of each datacube represent a multidimensional framework of classifications (time, region, branches) and cells of the datacube contain measurements such as income, consumption, number of unemployed, etc.. The user can slice-and-dice the datacubes in order to obtain different views on the data.

11. The main problem of this structure of the datawarehouse is the lack of consistency between statistical departments and their publications. Different statistical departments use different concepts (e.g. different definitions of the number of employees) and different classifications, making it nearly

impossible to combine data from different departments into one datacube. Efforts in the area of statistical integration and co-ordination must result in more consistent cross-sectional and timeseries data (besides NA) and on better (centralized) metadata.

III. COMBINING DATA FROM PRIMARY AND SECONDARY SOURCES: THE SYNTHETIC CENSUS

12. With the rapid developments in and application of information and communication technology, more and more electronic data will become available as a secondary source for statistics. Important sources will be administrative (public) records in registers like those of the IRS and the Social Security Administration, but also records from other private sources. It is to be expected that in the future these registers will become the main data source for NSI's. Combining this secondary information with its own survey data will become an opportunity of the utmost importance. But it will also constitute a major methodological challenge.

Figure 3. Surveys and registers

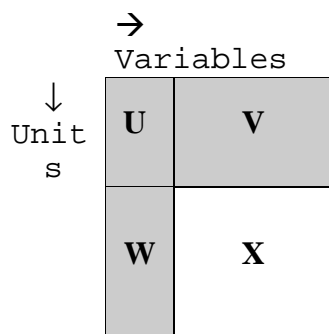
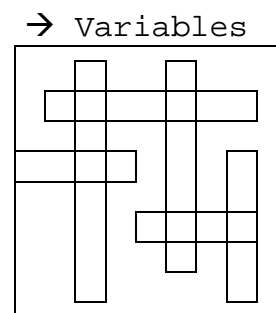


Figure 4. The Synthetic Census



13. To explain the differences between administrative data sources and survey data sources, a graphical representation is helpful. Figure 4 contains a table where the columns represent the variables (or concepts) of interest, and the rows represent the units (persons, establishments, etc.) or objects on which measurements are made. Surveys are characterized by many variables and a limited number of units (samples). The survey data are represented by U and V in the table. Both sets of variables are defined by the NSI. Secondary sources such as registers are characterized by many units and a limited number of variables. This part of the information is denoted by U and W in Figure 4. The NSI does not control measurement.

14. The combination of primary data sources (surveys) with secondary data sources (registers) implies the estimation of the missing part denoted by X in Figure 4. To be more precise, it means the estimation of certain aggregates over units (vertically) of the area X. In other words, the survey data are used to model the relationship between the variables in U and V. Then this model is applied to the variables in W to predict the values of the variables in X for the units not in survey. This approach is sometimes called mass imputation. Modelling of the relationship is the vital part of this approach.

Methodologies like record linkage, imputation and synthetic estimation will become dominant once the potential of this approach is better understood.

15. The example above illustrates the combination of one register with one survey. In practice, focus should be on the more general case of combining many surveys and many registers, all containing measurements on the same object. This situation is graphically represented in Figure 5. For each type of object (or object-type), the corresponding population can be enumerated. Starting with an empty table, all available data from surveys and registers can subsequently be filled in. Since the resulting database table (see Figure 5) ultimately contains the same number of units as the population, the resulting fully imputed table will be called the Synthetic Census (SC) or micro-database.

16. Two SC's are clearly very important: one for the object-type persons (with households as an aggregate) and one for enterprise/establishment object-type (with companies as an aggregate). Besides these two obvious object-types, other object-types (e.g. buildings, cars, jobs, etc.) can also be considered. Compare the limited number of registers in the Nordic countries, concentrating on persons, establishment units and buildings. At SN, two experimental SC's are under construction, one for households (SSB) and one for establishments (Microlab).

17. Imputation of the SC might result in nonsense data at the micro (unit) level. This is the case when the number of units in surveys (corresponding to U and V in Figure 4) is small compared to the population, and the relationship between the variables in U and V is weak even with multiple surveys/registers available. At the same time, disclosure protection will inhibit publication of the individual (possibly imputed) values of the SC. Therefore, a second database table, called the Reference Database, is constructed. It contains all aggregates (over units) in the SC that can be published. In practice, one might methodologically combine these two steps, imputation and aggregation, into one step.

18. In the future, surveys should be designed in such a way that the result of combining survey data with data from secondary sources (registers) is optimal. An example of such an approach is the Dutch Labour Force Survey (EBB). SN combines the survey data with the public register of unemployed persons in the Netherlands. In this case, the EBB is oversampled with respect to the register (giving a higher probability to people registered as unemployed). The survey results are used to assess the quality of the register (which is known to contain many persons who have already found a job but did not have their names taken off the register). Next, the sample is post-stratified using the (corrected) unemployment register and other population registers.

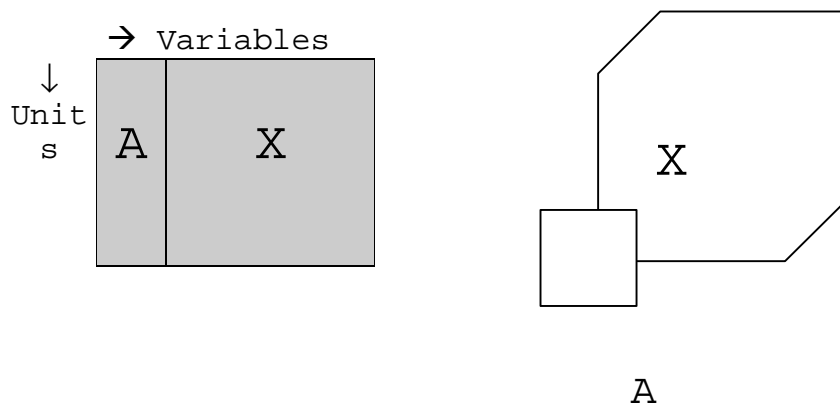
19. Statistical processing of data collected from primary and secondary sources will be dramatically different from the way we process our paper forms today. Administrative editing is needed at the input stage, but this will be file-based instead of record-based. The main concern of the file-based

procedures are conceptual errors (is the metadata changed in the respondent's system?) and technological errors (data communication and other errors). All contacts to respondents should take place at this stage. The resulting administratively clean input database should be viewed as a set of virtual respondents for the subject-matter departments which focus on the construction of the micro- and reference database. Statistical editing (often done before or after the construction of the Synthetic Census) is integrated in the estimation procedures, ultimately resulting in the Reference Database.

IV. COMBINING PUBLICATIONS: THE OUTPUT DATABASE

20. On the output side of the process, NSI's are facing similar demands for integration as on the input side. As an example, take the datawarehouse which SN has published (for free) at the Internet: Statline (at www.cbs.nl). A Statline query for, let's say, income or hospitals will result in dozens of electronic publications (datacubes) all dealing with certain aspects of the selected topics. However, each of these publications will look at the topic from a different perspective, and not many publications are co-ordinated (with respect to concepts and/or classifications), let alone integrated. As a result, users will be confused and will look for other sources in the future.

Figure 5. The survey database table and the corresponding datacube



21. This lack of statistical co-ordination and integration arises because each publication (datacube) is based on its own separate database table, arising in turn from its own separate survey. Figure 5 displays the relationship between the survey database table and the datacube in graphical form. The rows of the table represent units and the columns represent variables. There are two types of variables: quantitative variables representing measurements (income and consumption), and qualitative variables representing a classification of the objects in categories (year, region, economic activity). Note that sometimes a concept can be measured in both ways. For example, income can be measured on a continuous scale as an quantitative variable, and in intervals as a qualitative variable. In Figure 6, X denotes the set of quantitative variables (measurements) and A the set of all qualitative variables (classifications).

22. In the multidimensional datacube, the quantitative variables in X show up as (aggregate) values in the cells of the cube, indexed by the categories of the quantitative variables A along the axes. Of course, aggregation is over different units with the same categories. Different variables in X can be combined into one datacube by adding a new dimension (axis) for all the variables X. For more information about the concept of the datacube, see Willeboordse and Altena (1998).

23. There are hundreds of datacubes at SN, because there are hundreds of surveys and therefore hundreds of separate survey database tables. In order to offer our customers an integrated (and coordinated) view of all statistical publications, we should try to reduce (without loss of information) the number of datacubes and therefore all survey database tables to a very limited number. Now the question arises: which tables (and cubes) should we focus on?

24. In view of the discussion in section 3 (combining input sources), we should focus on the core object-types (persons and establishments). Then we should try to combine the database tables from different surveys with the same object- types into one Synthetic Census, in order to guarantee consistent and integrated views (cubes) for the customers. Any cube constructed from the same SC will be consistent and integrated, assuming that the database is complete both horizontally (with respect to variables) and vertically (with respect to units). In other words the database table should be a Synthetic Census or an aggregate of such!

25. As we have seen in section 3, aggregation of a SC into a reference database is necessary because of statistical quality and/or because of disclosure protection. Therefore, all published cubes/views should be based on at least two reference databases: one for persons and one for establishments. Again, it might not be necessary to physically construct the imputed micro-databases (SCs): what matters is that our estimation/imputation procedure results in a consistent reference database on the lowest possible level of aggregation, taking into account guarantees with respect to minimal requested statistical quality and privacy.

V. CONCLUSIONS: THE FINAL ORGANIZATION

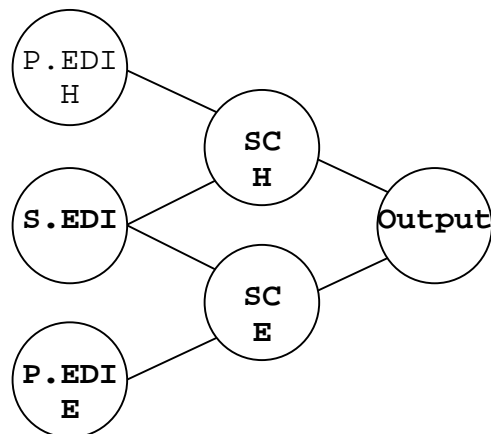
26. Due to technological change outside and inside NSI's, data collection (input), data processing (throughput) and data dissemination (output) will have to change dramatically. The most important aspect is a focus on external data sources such as administrative electronic registers for input, and data users (requiring a consistent and integrated view on statistical data) as output. This means the focus is no longer on separate statistical surveys. Both the input process as well as the output process can be improved by the introduction of two databases in the throughput process: the Synthetic Census and the Reference database.

27. We see in the future three important bit-streams entering the NSI's: primary EDI from personal/household surveys (mainly CAPI/CATI/CASI, with CASI standing for Computer-Assisted Self-Interviewing, e.g. on the Web), primary

EDI from establishments/enterprises, and secondary EDI from public registers (administrative records like those of the IRS and Social Security Administration). All this input data is collected in a number of so-called input databases, after administrative editing performed by the data collection (input) division. It is possible that the organization of the NSI reflects these three streams, with separate organizational (input) units for the processing of each of them, in view of the amount of data and the number of external sources.

28. After input processing, the (administratively clean) data goes to the two main integration divisions: one for persons and one for establishments. It is here that all data from the various sources are combined (using record linkage) into Synthetic Censuses (micro databases) according to object-type. Additionally, these divisions use imputation and estimation procedures together with statistical (macro) editing in order to build the reference databases, with publishable (aggregate) data. Finally, the output division wraps up the process with its final integration (possibly after NA) into a datawarehouse with a limited number of cubes. The total number of organizational units is six, as depicted in Figure 7.

Figure 7. Future statistical data processing (H=Households, E=Establishments)



REFERENCES

J.G. Bethlehem (1995): Improving the Quality of Statistical Information Processing. Proceedings of the 6-th Seminar of the EOQ committee on Statistical Methods, Budapest, 1995, pp. 87-114.

G.W. De Bolster and K.J. Metz (1997): The TELER-EDISENT Project, Netherlands Official Statistics, Autumn 1997, Statistics Netherlands, Voorburg.

H. IJsselstein (Ed.) (1996): Proceedings of the Conference on Output Databases. Statistics Netherlands, Voorburg.

A. Willeboordse and J.W. Altena (1997): Matrixkunde of 'The Art of Cubism'. Research Paper 9756, Department of Statistical Methods, Statistics Netherlands, Voorburg.