Working Paper No. 14
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)

Invited paper – SESSION 2

THE IMPACT OF NEW TECHNOLOGIES ON CONCEPTS IN SOCIAL STATISTICS

Submitted by Pieter Everaers, Statistics Netherlands

I. INTRODUCTION

1 Users of statistical information ask for timely, flexible and coherent information. As a consequence of recent methodological developments in social statistics and technical developments in data gathering and data processing, these criteria for high-quality statistics can be better met. New data collection techniques, for example via Computer-Assisted Personal Interviewing, integrated statistics combining register and survey data in statistical systems (labour, health), data storing in data warehouses and data dissemination techniques via public statistical databases have stimulated progress in providing harmonised (coherent) information. Progress is greatly influenced by the new opportunities offered by modern Information Technology. ICT developments have, without doubt, contributed greatly to the current state of manufacturing and dissemination of social statistics.

2. In this paper a very recent development in Information Technology - the advance in the technical possibilities to link data in statistical output databases - is described as an important step forward in order to achieve harmonisation of statistical concepts and definitions.

II. HARMONISATION OF CONCEPTS AS A PREREQUISITE FOR COHERENT STATISTICS

3. An important feature of modern statistics is (the need for) data consistency: each social, economic or financial event described by only one number. This one number is supposed to give the most reliable picture of an event in reality. This strategy is known as “one number policy” and is considered to be a good reference point for high quality statistics. Competing and inconsistent results on the same issue introduce doubt on the users’ side regarding the quality of the data and, as a result, the authority and credibility in statistics diminishes.

4. In the field of statistics, the harmonisation of concepts and classifications plays an important role in this one number policy. The use of one definition for each concept, the classification of the variables and agreement on the most appropriate measurement method are the basis for (internationally) comparable data and the backbone for authoritative statistics. The Siena group has contributed to the harmonisation of concepts as well as stimulating the use of harmonised concepts in social statistics. 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.

5. Harmonisation is mainly reached via theoretical discussions in international discussion groups, as are the implementation procedures via regulations and gentlemen’s agreement procedures. All these efforts are mainly driven by harmonising concepts in the input sources of the statistical process (registers, questionnaires).

6. However, alongside the theoretical considerations, technical (IT) developments also play an important role in facilitating and forcing harmonisation further. In this context the developments at Statistics Netherlands can be used as an example.

7. Since the middle of the eighties of the last century, 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 resulted in the harmonisation of standard blocks in questionnaires;
- micro integration forced researchers to formulate concepts with regard to persons and households in business surveys comparable with those from household surveys;

- 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 requires a high computing capacity. Combining data from these sources forces researchers to formulate 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 databases, baselines in different stadia, are considered the main liaison products of the statistical process. The Social Statistical Data Base (SSB) is seen as the resulting dataset, using all the combined data from surveys and registers. In this data base in principle every event can be measured via several original sources. Data matching and linking techniques further encourage harmonisation.

8. The introduction of computer-assisted interviewing, the advanced (re)weighting procedures, and the matching programs and data storing systems like those of data warehousing have all resulted in harmonisation of concepts and classifications. For the user of statistics, all these initiatives in themselves were relatively invisible but they resulted in the reduction of inconsistencies and the improvement of the quality of statistics. However, these actions were not directly linked to user dissatisfaction.

III. HARMONISATION AS A CONSEQUENCE OF VISIBLE INCONSISTENCIES IN THE OUTPUT

9. Traditionally, statistical offices publish their results in statistical yearbooks, theme publications, etc. A statistical office typically publishes a large number of statistics. Its traditional and still most popular product is the statistical yearbook, the statistical image of a country or region. Conventionally, users of statistical data try to find their data in the tables of a yearbook. Although not impossible, tables from a yearbook are difficult to compare and combine. In the nineties the electronic version of the statistical yearbook was developed. In the Netherlands this system is named StatLine. Many countries and international organisations use comparable systems and even comparable acronyms for their statistical output bases, approachable via the Internet.

10. StatLine can be seen as a very large electronic version of the statistical yearbook of the Netherlands, but with much more data (1.5 billion data-elements) and many more features for browsing of the contents. The data are organised in so-called cubes. Because StatLine cubes are digital it should be easier to combine figures from different cubes. However, the current implementation of StatLine makes this difficult. Each cube has it own topics, variables and categories. Therefore, it is currently very difficult to (physically) share a dimension/variable with another cube. This is considered as an element of StatLine that requries redesigning.

11. As a consequence of this structure, users still cannot easily compare data from different parts of the statistical results as stored in StatLine. For example, results stemming from a business survey may differ from those of a household survey, not because the statistical data are different – in fact they might be the exactly same - but because the cubes have been developed in different parts of the statistical office based on concepts and classifications that differ from each other, even in their minor elements.

12. New IT developments allow the gluing of cubes that share one or more dimensions. In the future StatLine, cubes sharing a dimension can be "glued" together. A simple example: a cube containing unemployment figures and a cube containing figures on bankruptcy share the same region and time dimension. The user will be able to combine figures from those two cubes into one output table. This is only possible if, and only if, the dimensions shared are related and these relationships are modelled in the underlying database. It is clear that this mode allows even very naive users to combine all kinds of tables, just because the variables are the same. Of course, as a consequence, every possible inconsistency will be more easily found.

IV. IMPROVEMENT OF COHERENCE

13. The cubes in the StatLine and comparable programmes have proved to be an excellent structure for storing, presenting and selecting statistical data. Statisticians designing cubes tend to make small and compact cubes. The side effect is that numbers belonging to a statistical theme are scattered in multiple small cubes. Designing good statistical cubes is an art, as is acknowledged in the 'Art of Cubism'. By developing good cubes and using (advanced) techniques as satellite cubes, many data questions posed by end-users can be answered. But cubes are built around statistical themes, representing general areas of interest. Even a real cubist will create cubes that will not meet the need of a specific user with specific needs. This user may want to combine this superb cube with a cube from another theme. The sharable or common dimensions/variables make this possible, and therefore the unstandardised (unharmonised) concepts will be very apparent. This in its turn will trigger statisticians to use standard classifications. The further improvement of the coherence in specific fields and as a result of the overall database is only a question of time. A statistical office does not enjoy answering users' questions regarding certain concepts that cannot be different but that appear to be so according to the "high rated" statistical output data base of the statistical office. It is clear that this development triggers the evolvement of generic dimensions e.g. a generic time (year) dimension and a geographical dimension as well as exerting enormous pressure to further reduce inconsistencies by harmonising and standardising concepts. These dimensions will be used in almost any cube and so they make it possible to combine those cubes. The next generation of StatLine will revolutionise harmonisation and will enormously facilitate – when used in an international context – improvement of the discussion on the comparability of data.

V. CONCLUSION

14. Methodological and IT developments have facilitated the harmonisation of concepts and classifications in social statistics. To date the accessibility of the output of statistical processes has not been an important factor in harmonisation. The digital statistical output databases were hampered because they were not so technically different from their paper predecessors. The new generation of digital output bases, accessible for public use, allows the gluing of the cubes of statistical tables. The pressure of the general (layman) public to generate data without inconsistencies will cause a revolution in harmonisation and standardisation. Over the next three years Statistics Netherlands will undergo this revolution. This will create a win-win situation: the end user of statistics will have access to coherent statistics, the statistical office will underline its role as the bureau of standards!

- - - - -