



**Economic and Social
Council**

Distr.
GENERAL

CES/AC.71/2001/27
28 November 2000

ENGLISH ONLY

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN
COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE/Eurostat Meeting on the Management of Statistical Information Technology
(Geneva, Switzerland, 14-16 February 2001)

Topic (iv): Integration of statistical (survey) data with registers (administrative) data

INTEGRATION OF STATISTICAL DATA WITH ADMINISTRATIVE DATA

Submitted by Statistics Denmark ¹

INVITED PAPER

I. INTRODUCTION

1. There is a tendency in all or most ECE countries towards increasingly making use of administrative sources in order to reduce the burden on citizens and enterprises, and to save resources in the NSI. The increasing use of administrative sources confronts us with a number of opportunities and problems that are worth discussing. The opportunities have been discussed in a number of international meetings, and the perspective of a general register-based statistical system is outlined in e.g. (Statistics Denmark 1995). In this paper we shall focus particularly on some of the difficult problems, and we shall attempt to relate the discussion to practical examples.

¹ Prepared by Lars Thygesen.

II. DIFFERENT KINDS OF ADMINISTRATIVE SOURCES

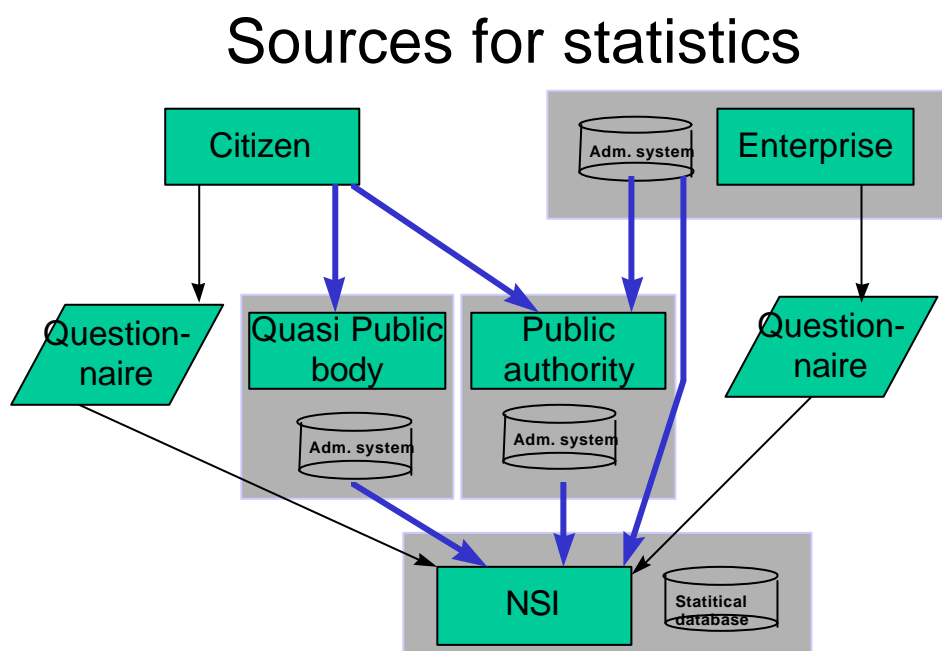
2. In this paper we shall use the words “administrative sources” and “registers” as synonyms. But when we talk of register-based statistics, what do we mean? Normally, we understand registers as data collected for some administrative purposes by somebody whose interest is not the production of statistics. Production of statistics is thus a secondary use of the data. Often we tend to think only of registers as data collected by public authorities, like the tax authorities, but in this paper we shall include data collected or gathered inside some organisation, e.g. an enterprise, for its own administrative purposes, e.g. book-keeping or administration of wages.

3. In order to be labeled register, the data has to be kept in some structured and standardized system, some kind of database dealing with certain entities or objects. When talking about sources for statistics, the most important entities are persons, enterprises and real estate units (land plots, buildings, etc.).

4. The organizers of this meeting have suggested a typology of registrations that could be utilized in statistics, according to their origin. Data stem from:

- ◆ Enterprise internal registers: created by administrations of enterprises (here the aspect of internal book-keeping systems is relevant);
- ◆ Service bureau registers: created by administrations of enterprises who keep records for other enterprises (accountant practices);
- ◆ Public registers: created by administrations/registrations within central or local and regional governments;
- ◆ Quasi public registers: created by administrations/registrations within institutions that carry out certain specialised and dedicated functions (e.g. social security registrations).

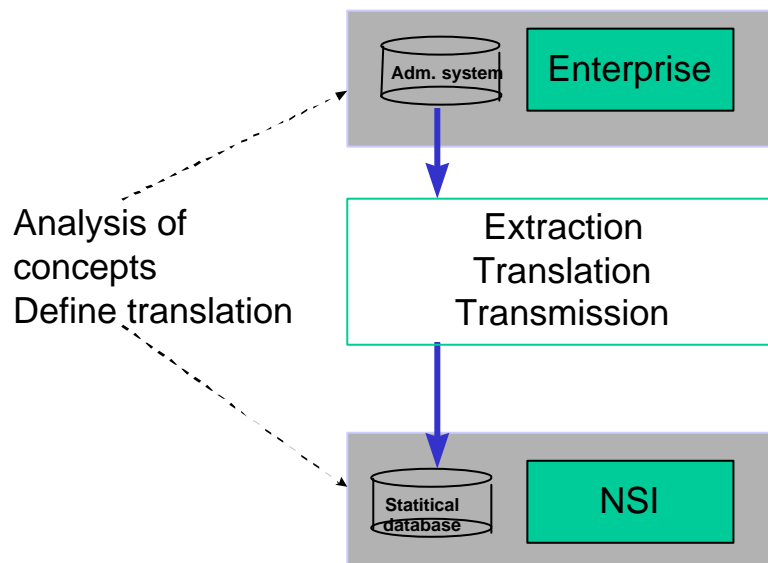
5. The different sources for statistics are illustrated in this diagram:



6. The arrows marked in bold indicate the routes of administrative data to the NSI. The diagram does not explicitly show the Service bureau registers, as they constitute a special case of Enterprise internal registers.

7. It may be argued that there is no great difference between the situation where an enterprise reports directly to the NSI (using a questionnaire of some kind) and the situation where the data is retrieved from the administrative system of the enterprise. It is reasonable to expect that in both cases the data has to be deducted from the same administrative system, but whereas the process has to become automatised in the latter case, we imagine some kind of intelligent human intervention - a translation - in the first. Consequently, we will have to reproduce this intelligent translation process when organizing register-based business statistics.

8. Whereas the use of public and quasi-public administrative registers for socio-demographic statistics in particular is already well-developed in a number of countries, there is much less experience with the use of Enterprise internal registers and Service bureau registers. In Denmark both of these methods have been employed for many years for reporting statistics on wages and salaries. We all have great expectations of the prospects of using electronic data interchange (EDI) methods for extracting and reporting statistical business data directly from enterprise accounting systems:



9. Promising applications of this kind have been developed through the EU R&D project TELER, resulting in the software Edisent. It has been proved that the process of analysing the concepts of the administrative system vis-à-vis the statistical concepts requested is a rather heavy task, demanding intensive communication with each enterprise. Once this has been done, the potential gains for the enterprises as well as for the NSI are obvious. Some examples mentioned below illustrate the range of

statistical tasks we are talking about.

II.1 Enterprise internal registers

10. These involve direct use of the company's own accounts as the basis for business economic statistics. In Denmark we have not yet fully developed the Edisent kind of approach mentioned above, but a pilot project is testing the possibilities of EDI reports from business accounts. In the meantime, enterprises are invited to hand in their internal accounts, and Statistics Denmark will then interpret and translate these accounts into statistical terms.

II.2 Service bureau registers

11. Here the most prominent Danish example is the wage statistics. A few service offices handle the wage administration of a large proportion of private companies, using standardized systems. On authorisation by the enterprise, the bureau will send the standard wage statistics message defined by Statistics Denmark for each employee. Statistics Denmark liaises with the offices in order to allow them to adapt their systems to the statistical demands.

II.3 Public registers

12. In Denmark, most of the socio-demographic statistics are based on public registers, governmental as well as municipal, taking advantage of common identification numbers like the Person Number. In business statistics a Central Business Register (CVR) constitutes part of a similar infrastructure. A somewhat different example is the livestock statistics. Up to now, these statistics have been collected directly from the farmers who indicate e.g. their number of cows, but a reform is underway, because every animal is registered in a public register, the Central Livestock register. There is still some uncertainty concerning the quality of these new statistics and whether the new figures will be consistent with the former ones.

II.4 Quasi public registers

13. Statistics estimating short-term employment trends based on total employer contributions to the Labour Market Supplementary Pension Scheme (ATP).

III. A COUNTRY PLAN

14. How could a country's NSI go about approaching these potential sources? A country list of relevant administrations and registrations can be drafted. Secondly, the sources should be examined regarding their conceptual usefulness (including the aspect of data pollution). This means that the concepts in the registers and administrations should be compared with the statistical concepts. Thirdly, they should be looked at from the aspect of legal and institutional accessibility. Fourthly, it should be looked at from a data administration and IT point of view, whether use is possible and in which ways (primary use and/or secondary use). But while undertaking this work, it should remain clear that the analysis is made using the target variables - what we want to measure - as starting points, and that we don't focus on the availability of data only.

IV. SOME PROBLEMS

15. While we recognize the potential benefits of embarking on the journey towards register-based statistics, we have to be well aware of the problems, which are considerable. They range from data quality, including comparability, through feasibility and legal aspects to resource considerations.

IV.1 Validity and relevance

16. The most important problem is naturally whether the register data can be used for estimating the concepts that are sought in statistics. Are the data relevant? The problem is that we do not control the measuring mechanism in the same way as we are do in a traditional statistical survey. Administrative sources have their own concepts which are defined according to their administrative purposes, and these concepts are not necessarily in harmony with our statistical purposes. Therefore, a translation or estimation must somehow be carried out from the administrative concepts towards the statistical ones. To what extent will the validity, reliability and international comparability be affected?

17. In favourable cases, the statistician may be fortunate enough to exert an influence on the content of registers. In some countries, e.g. Denmark, the legislation empowers the NSI to exert influence on registers which may be used for statistical purposes. However, our expectations in this respect should not be too high.

18. Another solution which has proven fruitful in many cases emanates from data linkage. It may be possible to let different data sources supplement one another through linkage, so that the desired information is provided by comparing a number of data from different registers which document the subject more or less perfectly. The use of several data sources is discussed below.

19. On the point of relevance, it must also be borne in mind that a statistical system which is exclusively or to a very large extent based on the registers of public authorities may end up giving a picture of the World as seen through the eyes of those authorities. Concepts or person groups who do not exist in the frame of reference of the tax authority or social affairs administration may also not show up in the statistic. An example might be homeless persons.

20. To exclude this risk it is important that there should be certain basic registers, such as the Danish Central Business Register (CVR), whose task is to record all units and collect information from many independent sources, without reference to any specific administrative purpose. We also know that many records from different sources can be linked so as to provide a statistic which is not tied to any single administrative viewpoint. At the same time, it is important to bear in mind that the registers cannot be expected to give us a complete picture of society and that it is therefore necessary to collect statistical data specifically in order to reveal aspects which are not covered by the registers. These additional data can be collected via interview or questionnaire surveys, and they should ideally include identifiers which make it possible to compare and supplement them with the register data

IV.2 Reliability and precision

21. The next requirement to be met by basic data for statistics concerns reliability. This applies irrespective of whether the data comes from questionnaires or registers. There must be a high degree of certainty that the recorded data faithfully reflect reality. If it is stated that a person is a painter and he is in

fact a government statistician, there is a risk of drawing incorrect conclusions from the final statistics.

22. It can, of course, be argued that, as long as the errors are not systematic, they will not result in skewed marginal distributions since errors in opposite directions have a tendency to cancel one another out. But when, as is often necessary, we look at statistical incidence between several variables, such errors become a problem since they can distort the assessment of those correlations. If the errors tend systematically in a particular direction, the problems are immediately magnified but their systematic nature may make it easier to correct them.

23. Finally, it is important that data be recorded with a degree of precision suited to the needs of the statistic, i.e. the scale used should be sufficiently detailed. For example, the degree of specification in occupational information must be fairly detailed if the information is to be used for epidemiological studies where the risk of disease is to be viewed in terms of job-related effects.

24. The need for precision remains even when the information is not to be used in statistical reports at the recorded level. In many cases groups are formed on the basis of variables derived from discrete items of data. The definition and application of such variables in statistical contexts depends crucially on the precision of the basic material.

IV.3 Comparability over time

25. Major problems may be posed for a statistic when legislative or regulatory changes result in alterations to the data content of administrative registers. On the one hand, it may be difficult or impossible to assess the long-term trend in a particular indicator if different definitions are used in the base material. On the other hand, problems may arise in deciding what changes in data values are to be viewed as reflecting actual events and what changes merely represent new concepts or definitions.

26. Altered data in administrative registers may be due to changes in the legislation applicable to the field, and the statistical consequences depend on what type of statistic is involved. If it is in fact a statistic for the monitoring of legislation, the function of which is to show how the administration of a law affects ordinary people, for example, the statistic merely has to go along with and adopt the concepts of the new legislation. Statistics on social benefits must thus, at any given time, reflect the rules applicable to the field and it may then be difficult to assess aspects such as behavioral changes.

27. If we are concerned with a more general statistic, on the other hand, it is not acceptable that it should be impossible to compare the statistical concepts before and after the legislative change. General statistics seek to elucidate certain concepts which are not defined in legislation, e.g. unemployment, a concept whose definition is to be found in the ILO convention. If the statistic relates to the payment of social benefits (e.g. daily cash benefits), and the rules for these are changed, it may be difficult or impossible to compensate for the change in the statistic. An attempt must at least be made to estimate the significance of the changes, so that time series can to some extent be linked together.

28. A Danish example of this is statistics estimating short-term employment trends based on total employer contributions to the Labour Market Supplementary Pension Scheme (ATP). As the main purpose of this statistic is to describe short-term trends, it is vital to eliminate or reduce the effect of rule changes on the measurement of employment. In some cases, the ATP legislation has been changed, e.g. extended to

cover new groups of employees, such as government civil servants or young wage earners formerly exempted from this scheme. In such cases, it has been necessary to correct the administrative figures, using the Labour Force Survey as auxiliary information.

29. Another example is the short-term statistics on turnover. These statistics are based on the turnover reports from business units to the VAT system. Until 1999 all the business units outside agriculture reported to the system quarterly, but from 1 January 1999 the system was changed so that large units have to report on a monthly basis, medium-size units report quarterly and the small units only have to report twice a year. The existing statistics are quarterly statistics. To be able to continue this it has been necessary to make some estimations of turnover for the small units whose reports are missing when the statistics have to be made and published. In this way, an uncertainty has been introduced into statistics formerly based on total turnover.

30. Finally, there are also instances of administrative registers changing without this being due to any changes in the basic rules. Typically, these are changes introduced for technical reasons or to achieve rationalisation gains. A problem may arise, for example, if it is decided that a particular item of data is no longer necessary.

31. System changes may also have a more local character, such as when a municipal boundary or the house-numbering in a particular street are changed with the result that a certain number of persons acquire a new address without having moved. Here it is useful if the changes can at least be identified as corrections.

32. Clearly the changes described are serious for the statistics in question if they are undertaken without sufficient consideration for their requirements. It is therefore important that statisticians be involved in the preparation of the changes so that at least their consequences can be assessed before they are introduced.

IV.4 Can statisticians influence data?

33. Since 1970, when the rise of register-based statistics in Denmark began, the impossibility to exercise the same control over the content of basic data as was customary has been recognised as a major problem in the production of questionnaire-based statistics. We cannot be sure that the registers cover the units of relevance with the same degree of precision or that data are defined in accordance with the needs of users. Further problems arise when the content of the registers changes over time and, paradoxically, a problem may arise if the registers become more reliable for, even then, we get data discontinuity.

34. It is of course desirable for statisticians to exert a certain influence on data content, but it must be remembered at the same time that registers are kept for quite specific administrative purposes and that the task of the "register keepers" is to serve those purposes in the best way possible. They cannot therefore pay too much attention to needs and wishes that "only" serve statistical purposes.

35. In Denmark, however, Section 1 of the Act on Danmarks Statistik states that anyone planning to set up a register must discuss the plans with Statistics Denmark so that the register can be set up in such a way that it also caters adequately to statistical requirements. There is also very good and close co-operation between Danmarks Statistik and register keepers in many fields.

36. Statisticians should not expect their extensive demands for additional data, different definitions and the like, to be met. Register keepers must pay strict attention to efficiency in their own operations and use of resources. Statisticians must therefore be very modest and only put forward major requests when an adjustment to a register can be expected to yield very substantial benefits to society. Very few instances of extra data being collected by way of registers exclusively for the purpose of statistics are known.

IV.5 Timeliness

37. Timeliness is an important quality dimension of statistics, and it is therefore important to find out how using administrative sources affects timeliness. However, we have not been able to identify any clear tendency. In Denmark we have a number of register-based statistics with extremely rapid production, e.g. general demographic statistics which can be published in detailed form within a couple of months. The Edisent model of business statistics also seems to be able to speed things up, since the processes became automated. On the other hand, some Danish statistics, e.g. on income and labour force, are based on a number of sources in the tax administration which take a long time to obtain sufficiently complete data, meaning that publication cannot take place sooner than one year after the period to which the statistics relate.

IV.6 International comparability

38. Ever since the dawn of modern register-based statistics around 1970, there has been much international discussion about the risk of impairing the international comparability of statistics. During the same period the importance of international comparability became increasingly recognized along with the growth of international cooperation. This led to a growing standardization of concepts and definitions, agreements on classifications and statistical systems. In most cases the focus of international standardization of statistics has been on the conceptual side while, with the exception of a few cases like the Population Census, there have been no agreements concerning which measurement instrument should be used. However, the development of the European Statistical System in the EU has gradually changed this situation, so that there exist today some total measurement vehicles, like the Labour Force Survey, where member states are bound by legal regulations or agreement to use a standardized measurement method.

39. Denmark has for many years fought to defend the honour of register-based statistics, claiming that in many cases their quality is not inferior, or in fact superior, to traditional statistics. This has not always been easy because statisticians have become used to their own methods and have disregarded the doubts about what really happens when a business manager receives a statistical questionnaire, completed by statisticians who may not fully understand the phenomena they are trying to measure. Do we always measure something meaningful?

40. But now the question is somewhat different: Can we produce comparable figures when using two completely different measuring methods? First of all, it is important to remember that we should not use the concepts of the administrative sources - unless they should accidentally coincide with the statistical concepts. We should not produce statistics simply because the data can be found in some accessible source. On the contrary: we should start with the concepts we want to measure - often concepts agreed upon internationally. And only then should we look at how best to measure them.

41. As has already been described above, there are techniques which can be used to secure a good and central measurement: translation of basic information, estimation based on combination of data from many different sources. It is essential that these procedures be based on the full range of data available at the most detailed level. In this case there is a good chance of reaching a fair result, which should then in turn not be too different from that measured using traditional methods, assuming that these work reasonably as well.

42. However, the fact that register-based statistics have certain quality dimensions which are not present in traditional surveys and censuses may create a problem for comparability. If we consider the Labour Force Survey, which is carried out in all EU member states: these surveys have varying non-response rates ranging from 30 per cent to 5 per cent in different countries - which may by the way also impair comparability between the LFS results from one country to another. We could easily contemplate substituting this survey in some countries with a register-based count, but this would give rise to lack of comparability, because the register-based statistics have virtually no non-response.

43. Far too little research into differences between statistics based on surveys and register-based statistics has been carried out to date, although some reports exist (Statistics Finland). A recent Swedish study (Håkan Lindström 1999) deals with quality standards in statistics based on administrative sources.

44. When focusing on international comparability, the solution has been proposed to use some common measurement vehicle carried out in exactly the same way in several countries (like the LFS) as a linking mechanism, calibrating the estimates from other register-based statistics to a common norm

IV.7 Feasibility: Legislation and confidentiality

45. While it might be very useful to combine administrative sources to create new information, this could be forbidden by law for confidentiality reasons. We know that this is actually the case in many countries. What are the limits and how can they be addressed? What kind of legislation or rulings regarding confidentiality promotes or hinders the use of administrative sources?

IV.8 Are register-based statistics cheaper?

46. It is evident that it is more expensive to collect new data from scratch than to use data which are already available in a computer system. However, the development of register-based statistics is far from trivial. It demands a lot of knowledge from the NSI, and a lot of work too, a fact which we sometimes tend to forget. There is an investment to be made in the development phase.

47. For instance, to use internal enterprise information requires a far deeper understanding of business systems than that required when simply sending a questionnaire. Unless the enterprises make use of standard systems, we have to develop and maintain a close relationship to each respondent. This is probably extremely useful since it should allow us to better understand the relevance or feasibility of the statistical concepts (questions) we are requesting. Maybe it will show us that what we traditionally request is not really meaningful. But the process is also extremely resource demanding. We should have an in-depth understanding of the business of individual companies, especially the bigger ones, in order to properly define the necessary translation mechanisms; the analysis of concepts is resource intensive and demanding, etc. This also places a larger burden on the NSI who has to make the data fit the harmonised concepts that are used.

48. Danish experience clearly shows that once the investment has been made, the NSI can reap considerable savings in the daily operations, and that these savings far outweigh the initial investment. This was demonstrated by the register-based Population and Housing census of 1981, where the cost of Statistics Denmark amounted to no more than one third of the costs of a traditional census. The experience from business statistics is similarly that we can save on resources. At the same time it should not be forgotten that respondents will generally be able to save resources since the data is re-used and does not have to be created a second time. In the public and political debate it has been stressed that it is important to relieve particularly the business sector of this response burden. In Denmark the weight of the burden of form-filling for private enterprises was estimated at 245 man years in 2000.

V. POSSIBLE INTERNATIONAL STANDARDS

49. The transition towards statistics based on administrative sources has developed more or less independently in different countries, although best practices have, of course, sometimes provided inspiration across borders. In some other fields of statistical work, international standards have proven to be helpful, and it is worth considering whether a similar approach might apply to certain aspects of register-based statistics in order to overcome some of the problems mentioned above in Section IV. However, many of the problems are of a nature which cannot be solved by standards, e.g. political problems. But one NSI may very well learn from and be inspired by other countries' experience.

50. It must also be remembered that the first prerequisite for attaining comparability is the international agreement on common definitions of the concepts to be measured. This is also the most difficult part, occupying many statisticians for years in international working parties.

51. Providing legal conditions for use: probably the most serious obstacle to extending the utilization of administrative sources in many countries is the lack of legal support for such activities. Could a standard for such legal rules be contemplated? Could rules functioning well in some countries be a model for others? This is a highly controversial and political question.

52. Statistical protection techniques: one of the reasons why there is strong political opposition to statistical use of administrative sources is that this will leave a lot of personal and business confidential data in the hands of the NSI, leading to fear that information will leak or individuals be recognized from statistical information. There is a strong desire, e.g. from researchers, once the detailed and potentially valuable data has been collected for statistics, to be able to make use of it for advanced research. In Denmark we maintain very strict protection rules, for example even unidentified micro data are not given to any user outside the NSI, while other countries issue public micro data files. This situation may put stress on restrictive NSIs to be more open. If we could invent and internationally accept some technique for protecting confidentiality in micro data, this would be most useful. Some EU R&D projects in this area have been promising, but so far no solution exists.

53. Book-keeping systems: it would be most helpful to comparability if binding international standards for book-keeping systems could be agreed upon, taking into account the international statistical standards. However, experience indicates that this is an extremely difficult task, even within one country. In Denmark we tried some years ago to introduce such a standard which would satisfy the needs of statistics, as well as the tax and other public authorities. The standard was also agreed upon, but even before it could be

introduced business lobbyists succeeded in persuading the Government that it would be very burdensome for them. The tax minister in several steps reduced the standard, and in this way the standard was made insufficient as the sole source for an accounts statistics. As a result, Statistics Denmark was forced to supplement this system with questionnaires to a sample of enterprises to be able to produce accounts statistics of a sufficiently high quality standard.

VI. THE FUTURE

54. In our opinion the further development of using administrative sources for statistics will concentrate on improvement of the conceptual translation from administration to statistics, especially harmonisation of the rules for processing data in the different fields of statistics in such a way that the same definition is used for a certain concept independent of the statistics concerned. So far there has been a tendency to a rather narrow horizon with focus on specific statistics when the selection of data and the rules for treatment are decided. We will see more cooperative work between the different fields of statistics and a more appropriate organization of the data processing. The result should be a higher quality of the statistical data deduced from a combination of administrative data and better comparability of concepts in statistics.

55. The work underway in Eurostat preparing the ECHP or its replacement after 2002 is very interesting. The problems of comparing data collected directly by means of questionnaires or interviews and data from administrative sources are being considered. Hopefully, the result of this work will be that there is sufficient comparability to allow both kinds of data to be used.

56. Access to statistical data on individuals and enterprises is a very urgent question. The balance between protection of integrity and the needs of statistics is very fine and gives rise to political problems in many countries. This problem will probably persist.

REFERENCES

Statistics Denmark 1995: Statistics on Persons in Denmark, A Register-Based Approach. Eurostat, Luxembourg 1995

Statistics Finland 1994: Evaluation Study of the Census (1994). Statistics Finland, Population Census, Volume 6B. Helsinki

Håkan Lindström 1999: Kvalitetssäkring i register för statistikproduktion med administrativt underlag. Statistics Sweden, Örebro