

NATIONS UNIES

COMMISSION ECONOMIQUE
POUR L'EUROPE

ОБЪЕДИНЕННЫЕ НАЦИИ

ЭКОНОМИЧЕСКАЯ КОМИССИЯ
ДЛЯ ЕВРОПЫ

UNITED NATIONS

ECONOMIC COMMISSION
FOR EUROPE

SEMINAIRE

СЕМИНАР

SEMINAR

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN
STATISTICIANS



CES/SEM.35/SV/1
25 June 1996

Original: ENGLISH

Seminar on Official Statistics - Past and Future
(Lisbon, Portugal, 25-27 September 1996)

SESSION 5: OUR LEGACY TO FUTURE GENERATIONS

**PREPARATION FOR THE REUSE OF NON-CONTEMPORARY STATISTICS -
SOME DANISH EXPERIENCES, PROVISIONS AND IDEAS**

Report submitted by Statistics Denmark */

INTRODUCTION

1. Following an introduction to the topic, section 1 of the paper provides a brief overview of various types of re-use and re-use preparedness relating to the Danish statistics. It shows that there are a number of possibilities for re-use but also that it can be difficult to achieve optimal results owing to the many changes that have taken place in statistical information down through the ages.

2. In this light - subject to the optimistic proviso that the resources can in some way or other be obtained - the following sections discuss what might be done to improve the scope for using statistics as a historical source in future. Here a distinction is made between statistics that already exist and those that are to be generated in future. In the former

*/ Prepared by Poul Jensen, Director.

GE.96-

instance there is no way in which conditions can be influenced. In the latter instance it will be possible in the context of general planning to open up future avenues of reworking which, in principle, are greater than has hitherto been the case.

3. As a statistician, one of the reasons for reverting to the activities of bygone ages may be to examine how problems that have long been familiar but are still topical were once dealt with. It may also be to shed light on the historical development of statistics as a social phenomenon or to construct long-term comparable time series. In such cases it is striking how difficult it is to access older statistics when they need to be used in the present. Linguistic usage has changed; the conceptual apparatus has changed; the methods, design and presentation of statistics have changed; and all this makes comparisons between information from different points in time problematic. The differences may vary in size, they may be visible or concealed, and they may be of greater or lesser importance in a specific context. Problems of interpretation must therefore be evaluated on the basis of a knowledge not only of the characteristics of the data but also of their intended application.

4. For historians and other social science researchers, official statistics represent an important source - especially where they can be taken to have been originally compiled with a view to shedding impartial light on the social conditions of the time. Subject to a certain delay - some feel it is often too long - they also reveal with a degree of probability something about the statistically measurable phenomena that were originally considered essential. The statistics can also be used to elucidate such phenomena, therefore, without necessarily calling for comparisons to be made over lengthy periods of time.

5. When it comes to employing less recent statistics, however, the user is nearly always faced with a difficult task. This has hindered, though not prevented, researchers from applying statistics as a primary source. Improving the scope for efficient and correct application of older statistics is an important task which, rightly considered, is of common interest to researchers and statistics makers alike.

6. However, that scope can only be feasibly and substantially improved by investing not inconsiderable effort. Some feel that it is a self-evident mandate for the central bureaux of statistics to ensure and facilitate the re-use of earlier information. Yet as a rule the resources of the central bureaux barely suffice to make internal ends meet. Nonetheless, since the re-use requirement has essentially been brought about by research demands, it is only natural that research funds should contribute to their financing.

7. The purpose of this paper is not to make proposals as to how the financing aspects should be solved, however, but rather to report on various Danish experiences and provisions regarding re-use (section 1), and

- to posit some preliminary considerations as to what **might** ideally be done from two angles: What can facilitate future use of

statistics from the past (section 2), and in connection with the ongoing production of statistics how can one prepare for efficient use in future (section 3)?

Provisions and experiences

8. Actual re-use as well as setting-up and preparedness for re-using census statistics are done in a number of different ways and from a number of different angles. This section contains a brief summary of the most important measures of this nature in Denmark.

The organization of statistics production

9. The way Danish statistics production is organized - particularly in the case of personal statistics - makes it largely suited to re-use, albeit in relation to a shorter time frame.

10. Danish personal statistics have been described in "Statistics on Persons in Denmark" (1).

11. For each statistical field, registers have been created which can be used for reworking. Under the provisions of Danish register and data protection legislation, however, Statistics Denmark stores such data for a relatively short period only, typically 5 or 20 years. After that, selected chunks of the data are transferred to the National Archives, while others are deleted. Use of these records requires the permission of the register authorities.

12. Special re-use options are available for those databanks that have sizable data volumes extending back in time. Research databases which Statistics Denmark has created in association with research councils also contain historical information. Such large research databases include the IDA (Integrated Database for Labour Market Research), which amongst other things contains combined data on companies and persons dating back to 1980. Research activities in more narrowly defined fields also give rise to the generation of data sets, which are documented and stored - inter alia with a view to subsequent follow-up of the initial results.

13. A databank with a series of harmonized financial data stretching back to 1947 acts as a data platform for Statistics Denmark's "Model of the Danish Economy" (ADAM).

Publications

14. The most obvious source of use for former statistics, of course, is the publications that have been issued with time. The amount of material involved is very large, however. An exhaustive bibliography of Danish

statistics (2), published in 1984, includes a total of about 800 items, classified by topic.

15. On closer reading, it emerges that a total of some 15 or so publications with longer harmonized time series have been published over the years.

16. Particularly in the national accounting field, there is a wealth of historical material. This is a natural consequence of the pride of place necessarily given to continuity in using national accounting information. When methods, concepts and systems are changed, however, the consideration of continuity calls for some reconstruction of succeeding series.

17. Inter alia, this is exemplified by the fact that Statistics Denmark, together with the Institute of Economics at the University of Copenhagen, in 1950 issued a publication containing studies of Denmark's gross domestic product (GDP), 1870-1950. This series was comparable with a listing published in 1962 for the period 1947-1960, which contained revisions essential in relation to previously published figures for 1930-1954. The national accounting computations now used take 1966 as their principal point of departure and build on the current SNA system, though this will be superseded by a new system in the near future.

18. Figures computed for 1870 to 1966 at 1966 system level are available in manuscript form at Statistics Denmark, but even now it is fair to assume that new versions of the older series will be a desirable item following the conclusion of the main revision of the national accounts currently under way. This is based on the new SNA/ENS system, new NACE-gearred sectoral classification and modified primary statistics.

The National Archives

19. The Danish state archive services comprise the National Archives, a number of provincial archives and some special-purpose archives, including the Danish Data Archives referred to in paragraph 25.

20. As mentioned above, the legislation stipulates that Statistics Denmark hand over the data to the National Archives on IT medium. Some proportion of questionnaires and other hard-copy data is also stored at the National Archives.

21. The purpose of archiving statistical material at the National Archives is not only statistical but above all for general historical purposes; moreover, the statistical data do not provide systematic coverage of the entire production of statistics and therefore do not systematically lend themselves to reuse.

22. Experience with reworking magnetic tapes archived at the National Archives is still scanty, but there has been brisk trade in using paper records for a number of special purposes. This is particularly true of the questionnaires for the historical censuses, which have a number of potential applications. As personal data, these censuses - the last of which was in 1970 - will only be made freely available after 80 years.

23. In future, the majority and the best of the options for statistical reworking will be linked to computer records with information in statistically processed (coded) form.

24. "True-to-source" computer registration does not occur, or at least only rarely, in statistics production. Such a thing is important for many historical applications, but if an affordable and detailed basis is to be established for reworking statistics, the best method will be to store all important primary data in as compressed numerical IT form as possible. In this respect, general historical and statistical archiving considerations scarcely have any interests in common. This problem area is elaborated in the last section.

The Danish Data Archives (DDA)

25. DDA is now a special-purpose archive under "The National Archives" but was originally founded by the Danish Social Science Research Council and for an intervening period was a unit under Odense University.

26. One of the purposes of DDA is to carry out archiving of surveys and other data sets compiled for specific research purposes. To a significant extent, therefore, the material stored by DDA is statistical in nature. Furthermore, some data originate from the official statistics. This applies, for example, to the reprocessing of geographically delimited parts of older census forms, which are archived at the National Archives, as mentioned.

27. DDA has devised a number of systems for systematic storage, documentation and content description at various levels and for searching the various data from various angles of approach.

28. The availability of the data is - by agreement with the original owners - classified in different categories. To date, the information involved has been exclusively anonymous and not subject to the restrictions of the data protection legislation, though other considerations may result in the "donors" wishing to have a greater or lesser degree of joint influence over their use.

29. DDA's stock of archived surveys now comprises some 2,500 units. The chances of the individual user retrieving and using the information of relevance to him or her are particularly good.

30. A "Danish Data Guide" is available in English. This contains a great many items of general information, including data on statistical registers and "Study Description Excerpts" as well as a "Primary Investigator Index".

The past in the service of the future

31. Taking the measures referred to in the first section as a basis, two questions can be posed:

- a. Is it possible to further improve the re-use potential of the statistics that already exist, so that the researchers of the future will more readily be able to use statistics as a primary source?

This question is dealt with below.

- b. What should be done in conjunction with the actual production of statistics to facilitate future reuse requiring more detailed reworking of the primary data?

This question is dealt with in the last section.

Historical series

32. The experience gained from historical publications in Denmark shows that there is a large and qualified corpus of work associated with the generation of long historical series. Therefore, such series can only qualify for consideration in fields where the requirement is particularly great - as it is in Denmark in important fields, statistics having undergone great changes over the past 30 years, inter alia as a result of using new methods and sources (register data).

33. In Denmark, the areas where there appears to be a particularly prominent need to harmonize time series for a longer period (e.g. covering the 20th century) are:

- labour force and general employment statistics,
- occupational structure statistics,
- central national accounting series on a new SNA/ENS basis.

There are also needs in other fields and at other levels of detail, however, based on the many divergent and varying objectives of research.

34. Compiling detailed, "ready-to-use" historical statistics in all areas is not a realistic option, however. The question is, then, whether anything else can be done to facilitate the application of statistics in such cases.

A pathfinder amid historical statistics ?

35. Gaining an all-in bird's-eye view of the overall *current* production of statistics is certainly difficult, though suitable aids are available to this end.

36. For instance, in Denmark a "Guide to the Statistics", introducing statistics in the individual areas (4), is published at intervals of several years.

37. It would be overly pretentious to attempt to compile a directory proper for all the statistics produced over the years, but spurred on by the research element, we in Denmark have contemplated whether it would be possible at least to produce a "pathfinder" that could provide guidance for historical "explorers" in their quest for historically rewarding sources.

38. Expressed in **present-day terminology** but with due respect for the statistics' **"own terms"**, the gist of the idea for each individual statistical series or individual listing of importance is to compile a comprehensive textual description of the substantive and conceptual qualities of the statistics, relations and comparability with other statistics.

39. In addition, there should be references to publications and other sources in which the statistics are to be found - and preferably also to publications etc. in which they have been used as essential input.

40. These data should be cross-linked both with a general description of the development of the statistics and with a series of retrieval systems on a modern IT medium.

41. From a research point of view, this would constitute an important instrument. It would require a large one-off investment but, once generated, overall registration could be continuously updated rather more easily with data culled from future statistics.

The near future in the service of the distant future

42. A clear distinction needs to be made between what can be done to facilitate the application of statistics from previous ages and what can be done to facilitate the future use of statistics not yet produced. In the first instance one must take the documentation and information as one

finds it. In the latter instance it is - in principle, at any rate - possible to plan not only for current but also for future applications.

43. The very awareness that the data may also assume an - as yet unknown - role in the future may possibly help the producers to make the planning, documentation, processing and presentation more universal and hence more comprehensible to the researchers of the future than normally dictated by statistical traditions. Looking at the experience from previous instances of re-use, it seems fairly clear that the information which is most readily understandable and re-usable is that accompanied by careful, all-round descriptions.

44. But if life is really to be made easier for future users, it will undoubtedly require a special effort, which in turn requires resources on which there is no visible short-term return.

45. On the other hand, the example from DDA shows that relatively modest means can go a long way. Systematic long-term preparedness in re-using public statistics can draw inspiration from DDA and its foreign sister organizations in some areas, though not in others. Whatever the case, it will undoubtedly require considerable development work.

46. No endeavour has been made to analyse these requirements in more detail here, but a few overall views should be set out as an outline for debate:

1. The detailed data in the archiving system must, in principle, be complete, which is to say that they must include all essential statistical information and all years. This view may possibly be modified for short-term statistics, however. This is a tall order to make, of course, but "gaps" in coverage, chronologically or in terms of area, may risk compromising essential future applications.
2. Conversely, in order to save space, the data must be concentrated - there must be no redundancy within the individual data sets nor between data sets where conceptual demarcations and identifications permit the transfer of data between them.
3. Only proper statistical data in coded form and at detailed level are to be archived. The many auxiliary data normally used in processing are omitted but referred to in the documentation.
4. The detailed information should be supplemented with elaborate and general-purpose documentation concerning definitions, the creation and delimitation of data as well as processing procedures and channels of conveyance for the statistics. In addition, correlations with other statistics should be thoroughly

elucidated. Retrieval systems should be constructed so as to enable specific items of information to be retrieved from various angles of approach.

5. By its very nature, the detailed information will be subject to rigorous "statistical confidentiality" in accordance with the obligations incumbent on the central bureaux. How an archive system of this nature would tie in with the data security legislation of the various countries is generally difficult to envisage, just as there is no telling in the present-day situation how conditions for accessing such types of information might be defined in the long term.
6. The documentation information, however, must be such that anyone can gain access to it.
7. Compilation of both sets of data should be done under the auspices of the statistics producers in as close as possible rapport with the primary production process but also in collaboration with researchers. How to organize long-term storage of the information is a question that will need clarifying in relation to the data security and archiving legislation in the various countries. The question of which technical media and systems are suitable will also have to be regarded as entirely open-ended.

Bibliography

1. Statistics on Persons in Denmark. A Register-based Statistical System (English version). Statistics Denmark and Eurostat, 1994.
2. List of publications: **En oversigt over Danmarks Statistiks udgivelser (in Danish only)**. Statistics Denmark, 1986.
3. Danish Data Guide, 1993 (English). DDA. Odense.
4. Guide to the Statistics (in Danish only). Statistics Denmark, 1991.