

**Economic and Social Council**Distr.: General
27 May 2013

English only

Economic Commission for Europe

Conference of European Statisticians

Sixty-first plenary session

Geneva, 10-12 June 2013

Item 4 (b) of the provisional agenda

How should national statistical offices respond - moving from risk avoidance to risk management**The rationale for micro-data services at Statistics Denmark****Note by Statistics Denmark****Summary*

This paper describes how Statistics Denmark approaches the issue of micro-data access. It grants remote access in a secure setting based on an evaluation of the research institute and the purpose of the research. The possible risks of disclosure cannot be ignored, so appropriate measures must be in place. Balancing with the risks is, therefore, a key concern of the top management of the statistical office. However, their view is that the micro-data access adds great value to the services that Statistics Denmark can provide.

The potential for producing new important knowledge based on comprehensive micro-data sets combining survey and register data is almost unlimited. A system based on well organised micro-data is the basis for statistical production in general. In addition, research can benefit from access to micro-data so as to inform important debates in society.

* This document was submitted late due to delayed inputs from other sources.

I. The Danish Statistical System

1. The system of official statistics in Denmark builds to a large extent on data from administrative sources, supplemented by surveys especially in enterprise statistics. Data from all data sources are integrated into one system (an “archive statistical system”, Thygesen 2010), in which all data on events or status are furnished with precise identification of the objects to which they relate: persons, families, enterprises, real estate; these object types can be interlinked as well. The data archive is primarily used for producing what is labelled as official statistics, i.e. all the statistics that are included in a statistical program financed by the government appropriation. These official statistics are made freely available to all users on the internet in many different forms, of which the dissemination database www.statbank.dk is the most important. The official statistics also comprise many statistics delivered to Eurostat and international organisations like the United Nations and the Organisation for Economic Co-operation and Development (OECD).

2. Data come from multiple administrative sources: Basic registers like the Central Population Register or the Central Business Register, tax registers, school systems, social protection, health systems, etc. Data are treated carefully, edited, combined and enriched in order to create good estimates of the concepts users are looking for in official statistics.

II. Secondary use of micro-data

3. However, in addition to serving the primary information needs, which can normally be well defined in advance, the system is also well suited for responding quickly to unforeseen data needs because data can be combined as needed. Data also comprise precise information on location of objects (coordinates), so that spatial analysis can be carried out.

4. The system supports longitudinal analysis where objects (persons, enterprises, etc.) can be followed through time. This way it is possible to analyse how risks of certain events vary according to the preceding history of the objects, so that hypotheses of causal interdependencies between various phenomena can be effectively tested. Large part of micro-data cover a long time span, some of them going way back to year 1970, others to the 1970s or 1980s. Of course, one has to be well aware of changes in definitions and data sources, and how these may affect the conclusions that may be drawn from the data.

5. The potential for producing new important knowledge is virtually unlimited using micro-data. The archive system based on well organised micro-data is a data goldmine that can of course be utilised for numerous purposes that cannot be financed by the government.

6. As has been mentioned, the system builds to a large extent on micro-data from administrative sources, so in principle, it would be possible for potential users to use these sources directly if they are able to get access. However, there are a number of advantages of reusing Statistics Denmark’s data, as compared to other possible sources.

7. *Efficiency.* Since Statistics Denmark has already collected the data and subjected it to necessary cleaning in order to be able to compile official statistics, a lot of work is already done and shouldn’t be doubled, unless there is some special reason for this. In addition, Statistics Denmark is used to handle requests from researchers, which is not necessarily the case for administrative bodies.

8. *The gold stamp.* Statistics Denmark is a quality label which automatically increases trust; so if research or analysis is based on these data, it adds to their credibility.

9. *Consistency.* When users build their analysis on the statistical archive, the results they produce may be consistent with the official statistics. This is no small benefit, since the results published by these secondary users will inevitably be compared to the official numbers. Even if differences are insignificant, it will give rise to long and unproductive discussions about why it is so.

III. Why should Statistics Denmark give micro-data access?

10. According to the Law, the main purpose for Statistics Denmark is to provide users with official statistics as a basis for informed decision making in society as a whole, the statistics are not developed only for the public sector. For this purpose, the institution receives an appropriation in the government budget. The Board of the statistical office decides what can and should be achieved with this budget; this is regarded as official statistics and is accessible for everyone for free.

11. But there are many more information needs, and they may be just as worthy. Therefore, the Law entitles Statistics Denmark to fulfil other needs on a cost recovery basis. It is our obligation to make good use the goldmine of data entrusted to the statistical office: We offer not only the data themselves but also the knowledge about the data and expertise in correct use of these data. The more these resources are used to create new knowledge to provide a sound basis for decision making, the better the purpose of Statistics Denmark is fulfilled.

12. In order to really make it possible for outsiders to use the data, Statistics Denmark must be able to explain in a very user friendly way what is in the archive and what it can be used for, and what it cannot be used for. Therefore, the provision of excellent metadata that is comprehensible for users is a huge challenge on its own (Thygesen & Grosen 2012), which cannot be described in this paper.

IV. Utilising the goldmine

13. In some cases, potential users may ask Statistics Denmark to carry out analysis based on their specifications. In these situations there is no need to give access to micro-data as the results will be aggregated. This kind of activity comprises at the moment all detailed spatial analysis, since addresses as the basis for geographic breakdowns are identifiers and as such cannot be released.

14. In addition to this type of analysis, Statistics Denmark may give access to micro-data in a number of situations. The micro-data must, however, always be de-identified, meaning that all information that can be seen as an identifier, including address, must be removed from the data. Even the de-identified data must remain inside the walls of Statistics Denmark because of the risk of backwards identification or indirect identification.

15. There are at the moment two main arrangements of access to the micro-data: Researchers' access and the Government Law Model.

A. Researchers' access

16. Researchers and analysts can be granted access to micro-data tailored for their research purpose. This arrangement allows for generation of knowledge that in many cases makes it possible to improve life conditions and reduce risk of undesirable events for the population.

17. Remote access in a secure setting is granted based on an evaluation of the research institute and the purpose of the analysis. The number of applications is high – 1,100 researchers are active at present – and the treatment of applications is rather liberal regarding the content of research projects. Statistics Denmark cannot and should not act as a censor on free research. The risk of an accident, e.g. revealing confidential data, cannot be ignored, so appropriate measures must be in place. Balancing the risk against the value of having access to micro-data for important research purposes is, therefore, a key concern of the top management. The management of the statistical office considers that the micro-data access adds greatly to the value of Statistics Denmark’s services.

B. Government Law Model

18. The Law Model is similar to the researchers’ access with great practical importance, aiming exclusively at the Government’s needs. It was established in 1980 in cooperation with the Ministry of Economic Affairs. The model contains micro-data on a sample of the population including a broad range of data for calculating the effects of planned changes in legislation.

19. The system is managed in collaboration with the Ministry and is hosted by Statistics Denmark. Its importance is underlined by the fact that government departments, especially the Ministry of Finance and the Ministry of Taxation, demand effective guarantees that the system will be up and running 24 hours a day, 7 days a week.

C. Bringing in additional data

20. Many researchers’ needs can be met by recombining and modifying data that already exist in the archives of official statistics. Some researchers need added data, for instance data collected in computer-assisted telephone interviewing (CATI) or computer-assisted web interviewing (CAWI) surveys taken by Statistics Denmark, or based on registers and other criteria defined by the user.

21. Some researchers or government agencies want to combine data held by the researcher or government agency, e.g. lists of persons who are members of a certain organisation or who have been subject to a certain medical treatment, or data from a survey they have carried out themselves.

22. In all such cases, the data must come to Statistics Denmark, be combined with data here and accessed through the secure environment. It is a one-way street: Data that have entered Statistics Denmark can never be released or given to a third party in a form that allows for identification.

V. Financing

23. Another concern is the financing of the arrangements. The government appropriation must be used only for the “official statistics programme” decided by the Board. Fulfilment of all other needs must be financed in full by the users. There is no payment for the data, as they have been produced by public financing to complete the statistical program, but all additional work must be paid for. This includes of course the whole management of a complex infrastructure system with servers, computer program packages, secure access mechanisms and oversight. Users must also pay for additional work generated by their tasks, e.g. for the creation of special datasets and collection of additional data from surveys.

24. As regards public sector researchers, a special arrangement with the Ministry of Research has been set up, whereby a research fund pays for the basic systems as well as for some initial assistance to each research team. The rest must be paid by the individual research projects. The research fund also pays for the creation of special “High Quality Documentation” of around 1,000 of the most utilised variables in the archive. Furthermore, this documentation is quality assured by external reviewers.

VI. Protecting confidentiality vs. facilitating access

25. It is obvious that giving widespread access to micro-data for researchers and analysts may imply a serious risk for the public confidence towards the national statistical office (NSO), even if the data can only be accessed in a de-identified form. If some data were to be backwards identified by someone in a research environment and used for improper purposes, this could cause serious damage. Therefore, the NSO has to take measures to reduce the risk.

26. On the other hand, it is important to facilitate as much as possible the good use of the goldmine of micro-data. Even though researchers might be able to find and process some of the data elsewhere, there is no real alternative.

27. The following protection measures apply:

(a) Access is always given in a secure environment at Statistics Denmark; users get access to the data situated in a server at Statistics Denmark from their own computers through a secure channel, and the access activity is surveyed in order to uncover attempts of illegitimate use;

(b) Access is only given to data without identifiers; of course backwards identification may nevertheless be possible in many cases;

(c) Only researchers from recognised research or analysis environments can get access; an agreement with the environment specifies the responsibility of the leader of the environment for any misuse from their authorised staff. The agreement highlights penalties in case of breach;

(d) Danish law applies in case of breach of security;

(e) Agreements have to be signed by each researcher / analyst, also highlighting penalties in case of breach;

(f) There are restrictions about which data can be accessed and used; for instance, Statistics Denmark will not take a special survey among a population that is delimited by values of a very sensitive variable.

28. In order to facilitate reuse, the following applies:

(a) Approval procedures are quick and rather non-bureaucratic; the research purposes are not censored;

(b) Access is rather easy once the researcher is authorised; the servers are monitored in order to secure sufficient power, and almost any software package can be used;

(c) We do not apply software for standard confidentialising, such as Argus. Users get access to the original, unscrambled data.

VII. Organisation and customer management

29. Basically Statistics Denmark has decided to merge all paid customer services in one cross-cutting department – although the decision has not yet been fully implemented. The micro-data services are, therefore, handled by a specialised department. This means that all technical issues can be solved in one place, including the organisation of safe access, handling of authorisations and questions of confidentiality. The customers can be handled in one common Customer Relations Management System.

30. Contacts with customers are managed by consultants in the Customer Department. Of course researchers and analysts need to go deep into the quality and meaning of data, so they must be able to draw on the expertise of the subject matter departments. For large projects, a proper project organisation is set up involving all necessary expertise.

VIII. Challenges at present

31. The following pressures against the system have been growing in recent years:

(a) Pressures for access to identified micro-data, especially from government departments or agencies. It may be difficult to understand why Statistics Denmark is a more secure place for these data than an agency. However, we need to take our special role in society in consideration: Statistical office provides the most comprehensive collection of data on social conditions, economic activities, health, etc. The Board of Statistics Denmark recognises a solemn responsibility to defend this unique role;

(b) Pressure from the research community: Why is it that all data must come to Statistics Denmark? If the data researchers have acquired from other sources are equally or more confidential than those from statistics, why couldn't data flow to the other direction?

(c) International researchers' access: In the Danish system, researchers have to be affiliated with a Danish, recognised research institute in order to be authorised. Many researchers from other countries have made arrangements to get an affiliation. It could be seen as discrimination and hindrance of the free flow of knowledge that not every researcher can get equal access. The reason for this limitation is that Statistics Denmark needs to be sure that penalties in case of misuse can and will be imposed, which is not the case if the institution is in another country;

(d) Need for cross-border register-based research is growing: There are efforts to create a possibility for researchers in the Nordic countries to carry out such research with similar data from several countries in one location, also avoiding multiple application requirements. There are also similar efforts within the EU, although the probability of finding really similar data in several countries is probably quite low;

(e) Confidentialisation: It is quite difficult to give access to micro-data on business units without indirectly revealing quite sensitive information on these units. Therefore, we must consider if scrambling, by using a system like Argus, should be implemented.

32. The Board of Statistics Denmark carefully balances the value of allowing for the micro-data sets to be exploited for knowledge generation against the necessity to defend the data gold mine entrusted in its custody.

IX. References

Thygesen, L. (2010): *The importance of the archive statistical idea for the development of social statistics and population and housing censuses in Denmark*. <http://www.dst.dk/extranet/staticsites/Nordic2010/pdf/bf7d6701-5b9f-4888-adc2-a45ce8debf87.pdf> Copenhagen 2010

Thygesen, L. & Grosen, M. (2012): How to fulfil user needs - metadata, administrative data and processes. *European Conference on Quality in Official Statistics (Q2012)* http://www.q2012.gr/articlefiles/sessions/21.3_Thygesen-Nielsen_How%20to%20fulfill%20user%20needs%20-%20metadata,%20administrative%20data%20and%20processes%20version%20final.pdf
