STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 8
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Washington, D.C., United States, 28-30 November 2000)

Topic (ii): Metadata modelling and terminology issues

METADATA FOR STATISTICS BASED ON ADMINISTRATIVE DATA

Submitted by Statistics Denmark [1]

Invited paper

I. INTRODUCTION

1. Increasing use of administrative sources in the production of statistics raises new types of problems closely related to the issue of metadata. Based on the experience of Statistics Denmark, this paper focuses on some of those problem areas, trying to show possible strategies.

II. WHAT DO WE MEASURE?

2. Data from administrative sources are often used to produce statistics related to the administrative process and closely related to the underlying legislation. Examples are statistics on social benefits, registration of cars or income tax related statistics. In these cases the use of administrative sources is logical and reasonably simple, but some metadata related problems still have to be tackled. In this case, the source is often a single administrative register.

3. However, to improve the efficiency of statistical offices, it is tempting to use administrative data as a basis for the general statistical system. In several countries this has been done successfully, particularly in the area of statistics on persons.[2] This usage of administrative data raises a large number of metadata related issues that are the main focus of this paper.

4. In these circumstances, it is common to combine information from a large number of sources. This is done in order to obtain or construct information matching the statistical concepts as closely as possible.

III. STATISTICS ABOUT ADMINISTRATIVE EVENTS

5. When data, produced as a result of or in connection with administrative procedures, are used to create statistics on that same procedure, this seems at first to be quite straightforward. If the whole operation takes places within the body responsible for the administrative area and the data only is consumed with in that area, this is not likely to raise many questions regarding metadata. The reason for this is mainly that the data and the procedures producing the data are well known in this context. Once the statistics are moved into the public area, i.e. when they become part of the National Statistical program, the need for metadata to accompany the data raises dramatically.

---

[1] Prepared by Soren Netterstrom.

[2] Statistics Denmark 1995: Statistics on Persons in Denmark, a Register-Based Approach. Eurostat, Luxembourg, 1995.

6. The main problem encountered is the fact that the documentation of the source IT-system in many cases does not contain these metadata. The documentation offered is most often geared to the construction of the IT-system. The variables are named (sometimes with 'strange' names), and sometimes their relation to some form or questionnaire is given. But the context as such, the underlying concepts, etc., are not of interest when building the IT-systems.

7. Thus the statistician may be relieved of the task of data collection when dealing with administrative data, but the discipline of collecting metadata still persists. What is the data really about? Why are they collected? Is there a legal base? What is the coverage (population)? As the conducting of a traditional survey requires considerable knowledge of the subject area, so does the process of creating statistics with regard to the administrative system. It is a long and often painstaking job to collect this information. As the statistical office has a response burden when collecting data, so they put a metadata response burden on the data provider within the administrative system as they try to collect metadata. Statistical offices should beware that their contact with the source body does not degenerate into dealing only with technical matters, but that from the outset focus is put on content rather than form.

8. The task of collecting metadata is a continuous effort. Administrative systems are likely to change over time. The legislation behind or the procedures of the system may change, leading to changes in the concepts of the system. These changes may be easy to spot, because they lead to changes in the format of the data received. But great care should be taken, because often only the concept of a variable changes without any visible modification, i.e. a variable keeps its name, definition and type of content. To the IT-system running the administrative system, these changes are of no interest and hence are not captured. But to produce reliable statistics, we need to capture these changes to make the statistics comparable or to produce relevant footnotes. So we need to constantly monitor the changes of the administrative systems and the legal systems behind them.

IV. GENERAL STATISTICS BASED ON ADMINISTRATIVE SOURCES

9. The moment we start to use the data from administrative sources to build general statistics, the whole situation changes dramatically.

10. As is clearly stated,[3] the point of departure for general statistics are the statistical concepts we want to measure - often concepts which are agreed upon internationally. So we do need to carefully consider whether the data collected by the administrative procedures are meaningful in this context. To what extent do the administrative concepts match the statistical concepts? What is the quality of the data? Is the population covered adequately? Without metadata collected as described above, we are not able to answer these questions.

11. In general statistics it is desirable to have concepts which remain unchanged for long periods and, therefore, we do need an assessment of the risk of changes to the key variables of the administrative system. Are they likely to remain stable or not? So rather than monitoring changes we need to predict the likelihood of changes.

12. The metadata gathered in this process should be stored, whether the data is eventually of use or not, and should be available within the whole of the statistical organisation. The data may prove to be interesting for other statistical areas and they should be able to examine them.

---

[3] Lars Thygesen, Statistics Denmark: The Use of Administrative Sources and International Comparability. Conference of European Statisticians, June 2000.

V. USING MULTIPLE SOURCES

13. A prerequisite for the use of administrative data seems to be the existence of a common identifier for the object of interest, i.e. a persons number, an enterprise number, etc. Through this number, information from different sources can be combined.

14. When such a common identifier exist, the statistical variable may be computed or estimated, at the level of the individual object, using data from different sources. Many variables may take part in such an estimate to match and determine the weight of the different sources. The typical system will handle large volumes of data (the whole population, everybody attending school, etc.), so the process of estimation is computerised. To do that, however, it must be based on a set of rules, on an algorithm that can be recorded. This algorithm should be captured by (or created within) the metadata system, possibly with notes arguing why this algorithm is chosen.

VI. WHAT ABOUT THE TIME?

15. When conducting a traditional survey, we normally fixed a point in time, the day of the survey or the like, to which all information relates. In an ideal world, we should be able to do so with administrative data as well and in some cases we are able to do so, because we receive information on all the events that change the content of data with the time of the associated event. The prime example is the Central Persons Register in Denmark, the base of the unique key (CPR-number) used in Denmark. Apart from sex and day and place of birth, it contains the actual address at any point in time. From the copy in Statistics Denmark, we can extract this information related to any given day. However, with other types of data, we may have a quarterly or yearly status, so we cannot freely set up the reference day.

16. When we combine data from different sources, the result is likely to be that the observation register no longer has a single reference day, but each variable may have its own reference day.

17. Another problem related to time is that we might receive data associated with a change of status (a transaction or event). But this day is in many cases not the day of the event but rather the day when the event was entered into the administrative system. In general statistical offices should be aware that an extract of a register on a certain day reflects the real world on an earlier day. It is important to examine and report this delay.

18. When we construct new variables, as described above, the problem of time should be carefully considered. If the data elements (variables) that are input to the algorithm have different reference dates, what will be the reference day of the new variable? For one 'observed' object, one variable with one specific reference date is used as a basis, for the next object another variable with another reference day is selected. So we might be able to record the individual reference date for each object. But it does not help us if we search for the reference date for a single variable across all the objects. As a consequence, we may have to live with 'blurred' reference dates for such variables.

VII. CLASSIFICATION MODULES

19. In the Danish model,[4] the term classification module is used to describe a data collection giving basic data about person. A number of these modules exist covering different themes such as employment status or education status, etc. and each module may be linked to a specific period or point in time.

20. The advantage of these modules is that they are used across the whole statistical system, acting as a common source. So the estimation of the variables is done only once, ensuring consistency across the 'surveys' utilising the modules.

---

[4] Statistics Denmark 1995: Statistics on Persons in Denmark, a Register-Based Approach. Eurostat, Luxembourg, 1995.

VIII. THE STATISTICAL OBJECT REGISTER

21. As a consequence of the above methods, a statistical object register is no longer linked directly to a single survey, where all information is gathered in a controlled way using a questionnaire. We find it useful, however, to keep the concept of the statistical object register. We still talk about carrying out a survey, i.e. collecting certain data about a certain population at a certain time. The (final) statistical object register is still considered the only valid input to creating statistical estimates.

22. What does change is the description of the 'survey' process. This becomes a description of the different source registers utilised and should include a description of the administrative source, the original purpose of collecting the data, possible problems with data quality in general, etc.

IX. THE VARIABLES

23. As stated above, the point of departure when creating statistics on administrative data should be the statistical concepts. As a consequence, the variables of the statistical observation register should reflect the statistical concepts rather than what was available at the source level. The focus of the description should be on quality. There are two aspects to this: one is reliability, i.e. does the data of the source registers reflect the reality or is it biased (for one reason or another)? The second is to what extent has it been possible to match the statistical concept through the estimation process creating the variable.

24. The source registers often contain a large number of variables required by different rules in the administrative procedure. It is tempting to copy all of these variables directly to the statistical observation register simply because they are available. In the actual situation in Statistics Denmark, we often find this to be the case. Each of the observation registers may include a large number of variables of which the majority have never been used for any production of statistics. Some of them have not even been used in the estimation of the statistical variables.

25. Careful consideration should be given as to whether these variables should be kept at all in the statistical observation register. If they are stored, they should be properly described, but why use costly resources on doing so, if nobody really cares about these variables? It may be argued that keeping the original variables would allow a future user to exploit the material, creating new variables, matching new statistical concepts, but then proper documentation of these variables would be required and must be maintained.

X. REDUNDANT DATA

26. In the implementation of a register-based statistical system, there is very likely going to be redundant data. As we have seen, data from the classification modules are copied into different observation registers and an observation register often delivers input to other registers. This creates a lot of redundant data in the system, but since the different modules and observation registers are static (not updated), this is not a real problem.

27. The real issue, however, is to make sure that the redundancy is controlled and that all instances of the same variable share the same description in the metadata system. The metadata system should not be static. Definition and value set may not change (unless they are corrected), but experience gained from using the data may lead to notes about quality, etc., and such information should be shared across the whole statistical system.

28. It should also be easy to find out whether two statistical tables from different surveys are using the same variable for a certain dimension or not, or to trace all surveys using a specific variable, etc.

29. An important part of the metadata system is to keep track of the redundancy of data and to avoid redundancy of metadata.