WP. 22
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 3-5 April 2006)

Topic (iii): Metadata and the Statistical Cycle

METADATA AS A CRUCIAL STARTING LINK IN NEW STATISTICAL CYCLES

Supporting Paper

Submitted by Statistics Netherlands [1]

I. INTRODUCTION

1. The urgency of using administrative data as the primary data source for statistics is becoming more and more important. Statistics Netherlands even has the legal obligation to first investigate all possibilities for using administrative data sources before sending out/setting up a survey. In the Netherlands now several large (national) projects are being implemented to develop unique basic registrations based upon the motto 'singular (data) collection, multiple use'. At the moment Statistics Netherlands is using more and more datasets from these registrations as primary source for statistics and some statistics are even already completely based upon administrative data sources.

2. In this context it is necessary to develop and implement a different and new start of the statistical cycle. Integrating administrative data in the statistical cycle as (partial) substitution for survey data is long and difficult process. Besides the cultural/emotional fact that often it is very difficult to cope with changes and the new developments, the highest threshold that has to be crossed is gathering maximum knowledge about the unknown new data sources. The biggest challenge is making the right match between the administrative data and the survey data that have to be replaced.

3. Metadata play an essential part in this process, it is the key to a succesful implementation of administrative data as the primary data source for your statistics. Looking at the statistical cycle from this perspective, there following three phases are most affected: a) data collection/input b) data transformation and c) output/using the prepared administrative data. It is crucial to gather the right metadata in each phase and then use it for the right purpose.

4. This paper focuses on various aspects and different roles/values of metadata as an essential element for the (right) use of metadata in the implementation of administrative data in the statistical cycle. The paper is organised as follows:

[1] Prepared by Harry Goossens, hgoo@cbs.nl

- In Section II we describe several problems and challenges we face in the implementation process related to metadata;
- Section III looks into various aspects of metadata, more generally;
- Section IV describes a method how to determine the right metadata for the various purposes;
- In Section V some results are presented;
- In Section VI finally some persisting difficulties and further developments are addressed to.

II. PROBLEMS & CHALLENGES

5. Although Statistics Netherlands already have successfully implemented several administrative datasets as primary source for statistics and even some statistics already are completely based on administrative data (for example the financial statistics on small enterprises / SFKO), it is still very difficult to realise new succesful implementations.

6. One of the main reasons for this fact is the growing complexity of the internal statistical process. Statistics Netherlands started in 1999 with the implementation of a new, uniform statistical process. This project, called IMPECT, aimed to redesign and standardize the various statistical processes. This resulted in new internal dependencies. Parallel to the change of processes also the organisational structure was changed, resulting in a more efficient structure. This was the end of the so-called 'stovepipes', specific organisational units for each individual statistic and branch of business (see also van Velzen, 2005).

7. The result of these changes was also that the clear often 1:1 relation between input and output transformed in at least 1:n. The bigger gap and the fact that link between input and output is no longer clear and direct make it much more difficult and complex to compare and match new input variables from external data sources with the output variables. Based upon the old structure with direct relations, the matching of variables mostly could be done from a strong focus on just the specific value of the single data elements, without the need of also looking at the process as a total. In the new situation the insufficient use of metadata is becoming a growing negative and aggravating factor in the implementation process(es).

8. Another factor is that in addition to the legal obligation for Statistics Netherlands to use as much as possible administrative data, there are large projects in progress aiming to develop a general system with an integrated set of national basic registrations for all governmental use. Allthough this makes much more datasets available and accessible, those datasets are part of a large and complex data infrastructure, which on the other hand makes use more difficult. Using this data demands very good knowledge of the data, process and subject in the broadest possible way.

9. In our modern with the all ICT possibilities regarding data management, the challenge of growing complexity is in fact a very common phenomenon that by the changes at Statistics Netherlands has only become a bit bigger and more specific. The last, also more common problem that must be noticed in this context is the fact that administrative data sources are almost always are collected for usage other than statistics. This means that you need to know that specific usage, including a certain expertise on that specialty.

10. Looking at all these challenges, we concluded if you successfully want to use administrative data it is inevitable to start with investigating the metadata first, before looking into the actual data. It is best doing this from an overview of the whole process that is concerned. And in addition to that, the more complex the process/environment is, the more it is necessary to do this in a structured way. This can be expressed in on central question: *'How can we first determine and then define the metadata we need to get the right information and knowledge about (new) administrative data sources that is needed for a succesful implementation of those data sources ?'* Since it is much easier to find something if you know what you are looking for…..

III. ASPECTS OF METADATA

A. Definition of metadata

11. Studying metadata often lead to discussions about the definition what is metadata and what is not. Starting from our central question we find this not really relevant. Much more it is important to know what information you want need and from there, how you can gather the required metadata. Therefore, it is essential to focus on the metadata that is essential for the specific process that is being handled. Mostly a massive package of often unstructured (meta)data is available, so first it is important to define the minimal set of metadata that is necessary and zoom in on that. After that, use the metadata together with the usage of the actual data, where this concerns both the use within the own internal process (such as data transformation for imputation, validation etc.) and the external use by clients (making statistics). Hereby it is important that you secure both the data as the corresponding metadata, as the meta reports the transformation of the actual data and sometimes also transforms.

B. Kinds of metadata

12. Although it seems to be very obvious, it is important to realise that there are different kinds of metadata and that there is not just one type of metadata. The more you study the metadata of a specific subject/dataset, the more differences you will discover. Therefore it is necessary to determine different categories to get a good overview of the specific metadata/information that is needed. In this context often the terms micro- and macro-metadata are used, but there are no standards. Depending on the dataset you are handling and what for you can just distinguish more or less (sub)categories that is convenient for you. It is mainly a question of using good common sense.

13. In the Dutch basic registrations project the following categories are used:

- *technical metadata*: specifications of variables, field length, data types etc., often defined in ICT specifications of information systems;
- *content metadata*: definitions en descriptions of the meaning of data elements, often defined in catalogues, as well as general knowledge about the specific subject. If for example using VAT data, you not only need to know the definition of ‘turnover’ that the owner of the data source uses in this dataset, it is also important that you have thorough knowledge of tax laws in general and VAT more specific
- *metadata about the process* (not to mix up with process data !):
  for good understanding and correct use of external data sources it is necessary to have good understandings of the process that leads to the data source that is delivered to you; not only what transformations, calculations etc. are performed by the supplier, but also the primary purpose for which the dataset is collected and which administration is held about that process.

C. Purposes of metadata

14. Before starting to use an administrative data source there a several aspects and criteria that need to be checked and approved. That can be more factual criteria as well as process related criteria. Either way you need information based upon which you can decide if a criterion is approved or not. From this point of view it is very helpful to first determine specific purposes you want to use the metadata for.

15. Here also there is no standard but much more it is predetermined within each specific process for which you are investigating the metadata. For example commonly used factual purposes are *quality of the content of dataset, completeness of the dataset* or *timelags between and reliability of deliveries.* Process related purposes often concern *matching input variables from the source with the expected output for statistics* or *monitoring process steps.*

D. Organisation

16. The organisation of the management of metadata is another crucial aspect for making the metadata accessible and useable. Therefore it is important to spend sufficient time and effort on it. Often the focus lies on technical solutions and developing systems for metadata management. Although this are important parts, they bring little advantage if not everyone involved is convinced of the importance of metadata . Awareness and discipline in following the procedures are minimal equally important as the tools. But this in fact is a theme for itself, so we will not further discus it in this paper. There is a lot of documentation already available see on this topic, for instance 'Wallgren & Wallgren' or the Swedish R&D report 2001.

IV. DETERMINATON METHOD

17. In spite of success in the past, recently it was problematic for us to give a quick and thorough answer on an apparently simple question. In a project concerning statistics of large enterprises, the goal was to define a dataset as a consistency matrix of 14 key variables on the level of enterprise groups. Therefore we had to match these 14 variables unambiguously with the source files of the underlying statistics (mainly based upon survey data) as well as with the basic statistic files build from tax data.

18. Explaining and discussing this with the (angry) project manager the reason why this action was so problematic became quiet obvious. Positive was that the management of the metadata is being done, but not in a structured and coordinated way. The fact that it being done at various places and in various ways, not systematically, using different codesets, with no coordination (= no organisation of the management of metadata) made it very difficult to gather the information we need to make that clear match.

19. Though the project of developing the system for national basic registrations focuses on a complete different process, namely the integration of various distinctive datasets into a overall system for generic usage, it faced the same problem. Participating in a special working group we had to answer the following question: '*What minimal set of metadata is necessary to make the system work.*' Trying to answer this question we experienced that it is very difficult to limit the huge amount of metadata to the really essential subset. To do so we first set up a structured way of working, partially based on the experiences from Statistics Netherlands.

20. We started with the distinction of the different main steps of the process and for each step we determined the various purposes metadata is needed for. All these different purposes did we combine to one set for the total process. Then we determined the various kinds of metadata we could find, looking at the total process. In this determination action it is important not trying to be complete, resulting in too much details, you must maintain a certain level of abstraction. The next step was to make a cross reference table, setting out the purposes facing the kinds of meta. For each cell we then described in common language what information we wanted to now and subsequently started to determine which metadata (elements) provide that information and where (at which registration) that meta is administered.

21. The first time we filled out such a table it was a vast struggle as we sometimes were being trapped in discussions whether a (set of) elements were metadata or actual data, or we had defined enough/the right purposes, but after a while we managed to stay in the right perspective and the method worked well. We experienced also that if you point out a metadata element in a cell it is beneficial to see which other cells (= purposes) that specific element is also needed and see if it perhaps changes during the process. And though at the beginning it is al lot of work that seems to take (too) much time, if being handy with this way of working and not going to detailed, results are coming up quickly.

22. For the basic registration project we defined the following cross reference table of 6 purposes facing 4 kinds of metadata. As it is now being worked out in detail, this is contains only examples that should illustrate how it is used.

PURPOSES / **KINDS**	Collecting data, Input in diverse registrations	Delivering data of the system, data exchange within system	Quality check and assurance	Usage of the delivered data	Viewing	Archivation
Technical Specs						
Catalogue Specs (definitions, general descriptions)	*What is the exact meaning of an element*			*Which services does the system provide*	*What information does the system provide*	
Description of the administration, how it is set up and managed	*What accuracy, timelags are used*		*What accuracy, timelags are used*			
Meta of the process of the administration, how is the progress, status information			*What status information is available*			

23. Looking at the process of implementing administrative data sources we determined 3 process steps :
- Input, gathering and basic preparing the data for use
- Developing a new method of statistics that must be applied
- The actual implementation in the statistical cycle

As mentioned before the definition of purposes and kinds is very related to the specific process when handling an administrative data source.

24. Based upon these experiences in applying administrative data, we are now setting up a new start of the statistical cycle, which aims to first determine and specify the metadata in a structured and more general way, before looking into the data itself. This way of working also provides a good and easier start to set up a clear and good organisation of the metadata management. Yet this way of working is just the beginning, there are still several aspects that now must be must be handled using this method.

V. RESULTS

25. Until the statistical year 2003 Statistics Netherlands made the annual structural business statistic completely based upon survey data. As a consequence of the new statistical law a project was started to develop a new method that makes it possible to use as much as tax data as possible.

26. This statistic is made for 6 branches, where each branch is split up in groups of enterprises with homogeneous main activities, the so called kernel cells. Further we use 9 size categories, based upon the number of employees. For the various combinations of branch and size there are 8 basic questionnaires, consisting of a generic part and a branch specific part. Depending on several possible combinations of branch, size and cell, in total there about 200 different questionnaires.

27. First the project focused on the small enterprises, i.e. having < 10 employees, being ca. 92 % of all enterprises and representing ca. 20% of the turnover. Based upon the FESS method (Fiscal Estimate Sbs System) the project made it possible that over the year 2004 (send out in 2005) Statistics Netherlands could realize a reduction of 16.000 questionnaires, on a total of 86.500 over all categories. For the small enterprises this means that 33 kernel cells can be completely based on tax data and 22 partially, resulting in only sending out 1.480 questionnaires, i.e. 8 % of the original statistical sample.

28. Recently the second phase of the investigation studied the middle class enterprises, i.e. up until a maximum of 100 employees (category 4-6). For this category, it was identified that 10 % – 15 % of the kernel cells can be based on tax data. Before it is possible to implement this method for other categories, it is necessary to investigate if it is stable over a minimum of 3 years.

29. For large enterprises it is not possible to use this method. The main reason is that as a result of the complex enterprise structures, the number of enterprises that can be unambiguously linked to the tax data is too small to be representative.

V. FUTURE DEVELOPMENTS

A. Persisting problems

30. One of the most persisting thresholds in the process of implementing administrative data is the cultural or emotional aspect. One of the main characteristics of making statistics is the fact that statistics must be vast and stable over a long period, so every new change is seen as a disturbance and threat. When investigating the possibilities of using administrative data as substitution for survey data they always first look at the ‘NOTS’, the things that are <u>not</u> possible, don’t match, instead of looking at the things that can be realised. This cultural aspect needs constant attention and just by coming up with practical results you can convince people.

31. Another problem is that if you want to use administrative data for statistical means that data source often is collected for complete different primary objectives. It is very desirable but also very difficult to gain some influence on the original primary goal and put in some statistical elements.

32. The last crucial issue that should shortly be mentioned is more organisational. If you start to use large external data source you have to devolop a business and data architecture that is suited for this new way of working (see the Swedish R&D report).

B. Future

33. At the moment within Statistics Netherlands, we are investigating which other possibilities we can find to use more administrative data in the broadest possible way. As mentioned above, the first results lead to reducing the amount of questionnaires being sent out for the business statistics. In addition to that there is now a project investigating the large enterprises. Further we now also are redesigning the questionnaire itself, so that enterprises have less questions to answer.

34. Another important development is that we participate in the National Taxonomy project, together with the tax collecting service and other ministries. The main goal is to define a general and harmonized set of data elements (variables) that are used in various processes in all governmental use. Based upon that harmonisation, it will then be possible for companies to deliver one uniform set of data to a central point of public authority, directly from their administrative software systems. All further governmental use must then be from that set.

35. The biggest difficulty here is that several different primary goals of collecting data (for instance tax versus statistics) must be combined and used the same definitions, which could lead to a lot of overlap. So every participating party must be willing to compromise. But who really wants to change?

References

J. van Velzen (2005), *The embedding of a uniform statistical process;*
UN/ECE Work session on Statistical Data Editing, Ottawa

Statistics Sweden (2001), *The future developement of the Swedish register system;*
Final R&D Report of the Register Project

A. Wallgren & B. Wallgren, *Register Based Statistics – administrative data for statistical purposes*.