2004/11

June 2004

Documents

Statistics Norway

Statistisk sentralbyrå

Statistics Norway

Anne Gro Hustoft, Jenny Linnerud
and Hans Viggo Sæbø

Quality and metadata in Statistics Norway

European Conference on Quality and Methodology in Official Statistics (Q2004)
Mainz, 24 - 26 May 2004

Quality and metadata in Statistics Norway

Anne Gro Hustoft, Jenny Linnerud and Hans Viggo Sæbø[1]

Abstract

Many statistical institutions have launched systematic quality work, with user needs as the point of departure for defining quality. One of the main quality dimensions of statistics is accessibility and clarity of statistics. Statistical metadata defined as systematic documentation of statistics are necessary for users to find, understand and use the statistics. In addition metadata linked to production processes are necessary to improve these. Hence, metadata has become an important issue in the work on improving quality and efficiency in statistical institutions. The paper considers the role of metadata within the framework of systematic quality work. Typical challenges a statistical institution faces in this field are how to provide different metadata to different external and internal users on different levels of detail, and how to implement and link various metadata systems to promote efficiency in production processes and dissemination.

Statistics Norway has developed many different metadata systems to serve different purposes and different user groups. These systems have often been developed in isolation from each other in a traditional "stovepipe" environment. This has led to the same information being stored several times in several places making the availability of updated and consistent information difficult. In the last two years, there has been a strong focus on the need to link existing systems and a requirement that new metadata systems should not be built in isolation. Our aim is that metadata should be updated in one place and accessible everywhere. Our metadata systems should also be useful as tools for the harmonisation and standardisation of our documentation. The paper will discuss the linking of different metadata systems, e.g. systems for data and variables data documentation, classification database and facts about the statistics or quality declarations for end users.

[1] Anne Gro Hustoft (agt@ssb.no), Jenny Linnerud (jal@ssb.no) and Hans Viggo Sæbø (hvs@ssb.no) are all Senior Advisers in Statistics Norway.

1. Introduction

Many statistical institutions have launched systematic quality work, with user needs as the point of departure for defining and improving quality. Statistical metadata are necessary for users to find, understand and use the statistics. In addition metadata linked to production processes are necessary to improve these. Hence, metadata has become an important issue in the work on improving quality and efficiency in statistical institutions.

The paper considers the role of metadata within the framework of systematic quality work. Typical challenges a statistical institution faces in this field are how to provide different metadata to different external and internal users on different levels of detail, and how to implement and link various metadata systems to promote efficiency in production processes and dissemination.

Statistics Norway (SSB) has developed many different metadata systems to serve different purposes and different user groups. These systems have often been developed in isolation from each other in a traditional "stovepipe" environment. This has led to the same information being stored several times in several places making the availability of updated and consistent information difficult. This is the main challenge for Statistics Norway's metadata systems today.

The paper describes briefly some of the metadata systems in Statistics Norway today, according to users, uses, and steps in the statistical production chain. Some organisational challenges are also considered. Our aim is that metadata should be updated in one place and accessible everywhere. The metadata systems should also be useful as tools for the harmonisation and standardisation of our documentation. We have recently launched a project to develop a strategy and plan for metadata in Statistics Norway, including the linkage of existing systems and rules for metadata management.

Two recent papers from Statistics Norway are particularly relevant for parts of this paper. Statistics Norway (2004a) discusses metadata needed when searching for and using statistics available on National Statistical Institutes' web services. Main conclusions from this paper are presented in chapter 3. Statistics Norway (2004b) presents the work on developing a system for storing and retrieving variable definitions, which is one of the key systems to be considered when discussing a future system based on linkage between different metadata systems in Statistics Norway. This system is briefly described among others in chapter 4.

2. Framework

2.1. What is metadata?

The term *metadata* is often used synonymously with "documentation". However, *structure* is a key word when it comes to producing and presenting statistics today. At least this concerns metadata used by IT-systems, but also metadata used by humans to find and understand statistics available on Internet among a huge amount of information and hypertext links. Documentation for a wide variety of users without a high degree of harmonisation and structure will tend to be useless. Hence, in this paper we will use the term statistical metadata to mean structured or systematic information about statistics.

A *statistical metadata system* is a data processing system that uses, stores and produces statistical metadata (UNECE 2000a). This means that such systems have some kind of functionality in addition to being a metadata storage. Most databases containing data or statistics also use, store and produce some metadata. Hence, they are metadata systems in addition to more general data systems, according to the wide definition above. However, in this paper the discussion on metadata systems will focus on systems that have storage and handling of metadata as their main purpose.

2.2. Quality and metadata

Many National Statistical Institutes (NSIs) have started a systematic quality work. Statistics Norway's work on this is described by Sæbø, Byfuglien and Johannessen (2003). Quality can most simply be defined as *fitness for use* in terms of user needs. The dimensions of product quality for statistics are often described according to Eurostat's criteria (Eurostat 1998):

- Relevance and completeness
- Accuracy
- Timeliness and punctuality
- Comparability and coherence
- Accessibility and clarity

Cost constraints are important, and costs always have to be considered in connection with quality indicators. Statistics must also be objective, and personal privacy must be protected. Good product quality is necessary to satisfy user needs, but improving processes (in the production of statistics) is the key for better product quality at an acceptable cost.

Systematic information about statistics, or metadata, is a precondition for proper use of statistics. Metadata are necessary for the clarity of statistics, and metadata assisting search for statistics and further processing also contribute to relevance and accessibility. Information about production processes is often needed in order for the users to understand the statistics as well. Such information is also crucial for improving production processes.

2.3. Types of metadata

In the paper Statistics Norway (2004a), statistical metadata are classified according to different user groups:

- Metadata for the users of statistical information
- Process documentation for internal users
- Metadata for external data providers
- Quality information for both external and internal users

Metadata for users comprise metadata for finding statistics and navigation on web pages ("discovery metadata"), and metadata for understanding and post-processing statistics. They include metadata accompanying statistical data submitted from NSIs to international organisations such as Eurostat.

Process documentation is needed for efficient statistics production, monitoring and control, training of new staff members and improvements. The users of such documentation are primarily the statistics producers, but also NSI management need such information for monitoring and control.

External data providers need metadata to provide correct data. These metadata are often linked to questionnaires (on paper or electronic).

Quality metadata normally refer to the quality dimensions of statistics listed above, and are needed by both users and producers of statistics. For funding agencies and management they are crucial.

Sundgren points out that very often the same persons may perform a combination of users/producer roles, so it may be better to talk about usages rather than users of metadata, see Statistics Sweden (2004). He also adds that software products and computerised data processing systems need metadata in order to function properly. Such systems can have both internal producers and external users of statistics as their users. We could talk about discovery, explanatory (including information on quality) and technical metadata. However, there will still be overlapping between the groups.

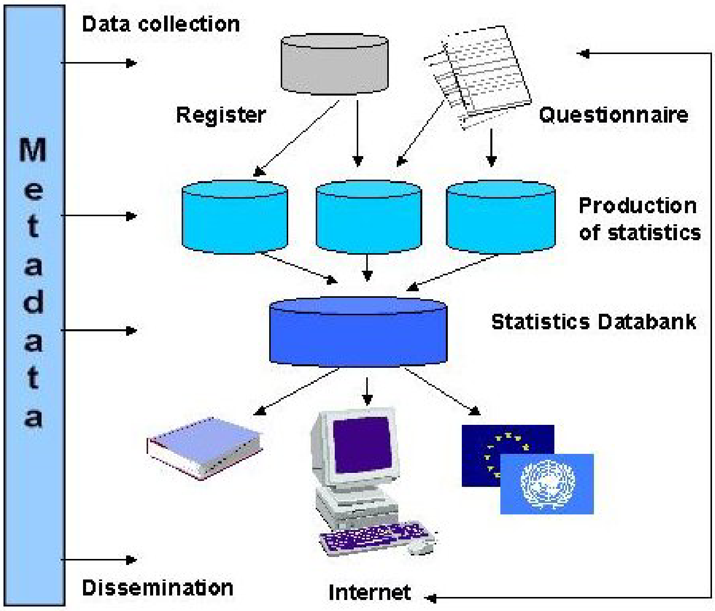
Metadata requirements will vary in detail and complexity within each group of users. This calls for metadata at different levels of detail and solutions such as linkages between different levels of metadata on web.

The classification by external data providers, internal producers and users of statistics roughly corresponds to different steps in the statistical production chain: Data collection, production of statistics, and dissemination of statistics to users.

Figure 1 is a simplified model of the data flow through a typical national statistical institute, with arrows that illustrate steps or processes with important metadata requirements.

As mentioned, there is some overlap between the metadata user groups. Users of statistics often need to know something about the production processes and certainly something about the quality. Definitions of statistical terms are or should be common for external data providers, producers of statistics and users. Lack of harmonisation of metadata and metadata systems across user groups and throughout the statistical production chain becomes an issue.

Figure 1. Data flow and important metadata requirements in a statistical institute



3. Metadata on the Internet in general

Statistics Norway (2004a) discusses metadata requirements for users of statistics available on the Internet, i.e. for finding and understanding statistics (discovery and explanatory metadata). Most NSIs have good basic facilities for searching and navigation in their web-services, and some explanations of statistical variables and classifications in addition to possibilities for downloading and further processing of statistics, as recommended in UNECE (2000b). However, the volume and complexity of the web requires easy access (to metadata) and a good metadata structure. Many NSIs do not have a systematic description of their statistics linked directly to the relevant statistics (such as information on background and purpose, production methods, definition of concepts and standards, errors and accuracy, comparability and alternative sources). This is an important challenge for the NSI web services today in addition to the quality of the metadata themselves. Regular updating is necessary for all metadata, and solutions ensuring that metadata can easily be updated together with the data, and only in one place (e.g. database) is a precondition for an efficient metadata system.

4. Metadata systems in Statistics Norway

Figure 2 shows existing metadata systems in Statistics Norway. They can be characterised according to main user group or production chain step: External data providers/data collection, internal users/production and external users/dissemination. Some of the most important systems are considered in the following, in the same order as in the figure.

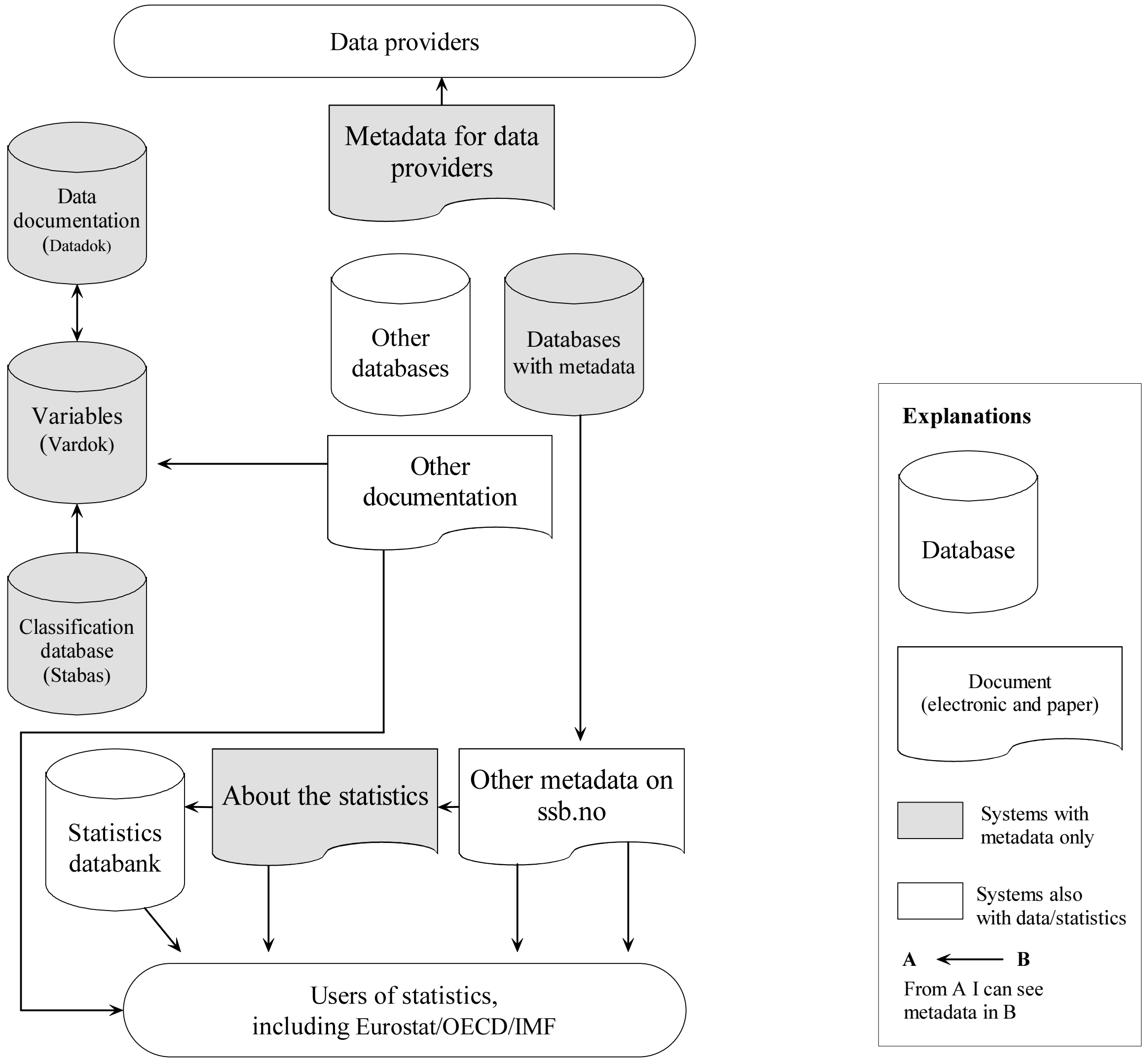
4.1. Data documentation system (Datadok)

The data documentation system includes:

- Technical documentation of data, so that the data can be processed in other systems, i.e. file descriptions
- Administrative documentation of data so that they can be located. Data are clearly connected to the appropriate concessions granted by the Data Inspectorate in order to restrict their access in accordance with these concessions

Everyone in SSB has access to this system. There are no plans for any external access to the system. Since the contents of the system are mainly technical and administrative metadata the main improvement for quality is in increased accessibility.

Figure 2. Metadata systems in Statistics Norway


Data providers
Metadata for data providers
Data documentation (Datadok)
Other databases
Databases with metadata
Variables (Vardok)
Other documentation
Classification database (Stabas)
Statistics databank
About the statistics
Other metadata on ssb.no
Users of statistics, including Eurostat/OECD/IMF
Explanations
Database
Document (electronic and paper)
Systems with metadata only
Systems also with data/statistics
A
B
From A I can see metadata in B

4.2. Variables documentation system (Vardok)

In Statistics Norway information about variables can be found in different documents and systems, which makes accessibility difficult. The decentralised storage also results in the same variable name being defined in different ways in different parts of the organisation, and one can also find different names being used for the same variable.

The variables documentation system is a central system for documenting variables (e.g. definition, validity periods, classifications used) and a tool for harmonisation of names and definitions of variables. The central storage will help us improve accessibility considerably and harmonisation should result in improved clarity and comparability of our statistics.

At present Vardok can be accessed by everybody in SSB, and there has been a strong user involvement in the development of the system. In the future external users will also, to some extent, have access to the information in Vardok, and it is relevant to differentiate between the needs of ordinary and advanced users. The variables definitions in Vardok do not contain too much detail, but it is possible to link a variable to more detailed reports if that is considered relevant. In this way one avoids burdening the ordinary user with lots of details, while at the same time giving the advanced users access to more information.

4.3. Classification database (Stabas)

Two important aims for our work related to classifications are:

- To make accessibility and maintenance of classifications easier
- To ensure common use of classifications across different subject matter areas

These aims especially focus on the quality criteria accessibility, comparability and coherence.

To achieve the goals, we decided to collect all classifications in Statistics Norway in one database (Stabas - Standard database). This database has been developed in cooperation with Statistics Denmark, based on the Neuchâtel terminology[2]. All classifications are not yet contained in Stabas, it will still take some time to achieve this goal. The challenges from now on are mainly organisational. Everybody in SSB can access the database. At present we are working to make it accessible for external users on the web. This will be specifically useful for institutions supplying data to SSB, but also for the ordinary user wanting to learn more about a certain classification.

4.4. About the statistics

About the statistics is a systematic description of statistics on Statistics Norway's website (www.ssb.no), linked to the daily releases of new statistics and to the tables in the dissemination database (Statistics databank). These metadata are aimed at the general user who wishes to learn more about the statistics (e.g. to check if it is comparable to certain other statistics), but who doesn't have the advanced users need for details. To supply the advanced user with more information, we want to establish links to more extensive documentation, e.g. detailed sampling reports and calculations. The box below shows the types of metadata contained in About the statistics.

After using About the statistics for 5 years, we have recently evaluated the need for adjustments. All in all, the subject matter divisions have invested a great deal of effort in making About the statistics, but there is still room for improvement. We found large variability in the way the template is filled in, varying both between types of releases and subject matter divisions. This variability may be caused by different factors; e.g. the guidelines for filling in the template may at some points have been too imprecise. In addition there is the constant competition for resources between the documentation work

[2] The Neuchâtel group working with terminology models for classification databases was established in 1999 and consisted of representatives from Statistics Denmark, Statistics Sweden, Statistics Switzerland, Statistics Norway and run Software-Werkstatt.

and the demand to comply with the publishing deadlines. One might also suspect that today's metadata situation, where the producers have to document the same metadata in several places, leads to some frustration. This is certainly the case for About the statistics where a major part of the information is also documented elsewhere. To improve the working situation for our producers, we have adjusted our guidelines to make them more precise, and we will also make standard descriptions for some of the sections (related to methodological issues) that have proved difficult to fill in. This will make the work easier for the producers, and at the same time help us harmonising the descriptions. Because of the overlapping information in About the statistics and other metadata systems, About the statistics will be one of our case studies when discussing further integration of metadata systems in Statistics Norway later this year.

Box 1. About the statistics

1. **Administrative information**
   - 1.1. Name
   - 1.2. Frequency
   - 1.3. Regional level
   - 1.4. Subject group
   - 1.5. Responsible division
   - 1.6. Authority
   - 1.7. EU regulation (if relevant)
   - 1.8. International reporting
2. **Background and purpose**
   - 2.1. Purpose and history
   - 2.2. Users and applications
3. **Statistics production**
   - 3.1. Population
   - 3.2. Dat a sources
   - 3.3. Sampling
   - 3.4. Collection of data
   - 3.5. Control and editing
   - 3.6. Calculations
   - 3.7. Confidentiality
4. **Concepts, variables and classifications**
   - 4.1. Definition of the main concepts and variables
   - 4.2. Standard classifications
5. **Sources of error and uncertainty**
   - 5.1. Measurement errors
   - 5.2. Non response
   - 5.3. Sampling errors
   - 5.4. Other errors
6. **Comparability and coherence**
   - 6.1. Spatial comparability and comparability over time
   - 6.2. Coherence with other statistics
7. **Accessibility**
   - 7.1. Other Internet addresses and publications
   - 7.2. Storing and use of basic material

4.5. Other systems

There are several other systems that are used to generate metadata for external data providers, e.g. electronic questionnaires, and production and dissemination systems also containing classifications and variable definitions, but these do not have handling and storage of such metadata as their main purpose. These include databases containing microdata in several areas and the Statistics databank with more aggregated statistics that is accessible by external users.

4.6. Linking of metadata systems

As mentioned in the introduction, a project to develop a strategy and plan for metadata in Statistics Norway has recently been launched. However, the linking of existing systems has already started. This has even now improved accessibility and should contribute considerably to the reduction of duplicate information in the future.

Today there is a link between Vardok and Stabas that allows the user to access a classification connected to a variable (e.g. the Classification of Occupations which is connected to the variable Occupation) directly from Vardok.

As mentioned earlier, it is possible to link a variable in Vardok to a report stored on an internal server and/or on the Intra- or Internet. This gives the advanced users access to more extensive information.

It is also possible to link a variable in Vardok to the files where it is used (through the file descriptions in Datadok). When a variable is linked in this way, you can have a view of the variable definition (stored and updated in Vardok) when working on a file description in Datadok.

In the future we also consider having a link from Vardok to our dissemination database (Statistics databank) in order to give external users access to our variables definitions. So far we have only made a demo of the link to show how it might work.

Today we have definitions of variables both in Vardok and About the statistics. This necessitates updating the same information in two places, which is a situation we would like to eliminate. In the future we want to make a link between variables definitions in Vardok and About the statistics to achieve automatic updating of the variables definitions in About the statistics.

There is always the danger that without a metadata strategy each new system will find a metadata solution of its own. This has been the case for some of the major systems in Statistics Norway (specially the ones aimed at data providers, fig. 2), and one of the future challenges will be to fit these metadatabases into the over all system without losing the benefits and routines that have been created connected to them.

5. Future work

In the recently launched strategy project we will set up an action plan for the specific development projects needed. In addition, guidelines for the future metadata work and management will be established. The overall aim is to establish an integrated metadata system that will contribute to better quality and more efficient statistics production.

Based on their experience with interviewing in the development of the variables documentation system, the project group has decided to use this method in the metadata strategy project. One of the strengths of the method is the user focus, and in the first step the project group has identified different groups of users who will need to be represented. The next step will consist of interviewing these users in their own offices. Two people will conduct each interview. One will have a dialogue with the user and the other will observe and take detailed notes. During the interview we will capture background information for the interviewee, try to map the metadata flow relevant to this specific interviewee (hoping that the chosen collection of interviewees will give quite a good overview of the important metadata flows in Statistics Norway), and discuss present and future links between metadata systems and organisational issues. Figure 2 or more detailed versions of it could be a point of departure and a tool for development of ideas. Our aim is that the collected information will give us a common vision for an integrated metadata system.

Our approach will be practical. We will not spend time trying to make exact theoretical definitions for specific terms, or doing research to create one all encompassing metadata system to take care of our metadata needs. As the resources are limited, our starting point will be the systems that already exist. New systems will be created only for important functionality that cannot be covered through existing systems.

To avoid duplicate information, we will promote central systems for the updating of specific types of metadata (e.g. variables and classifications), and exchange metadata with the rest of the systems that need this kind of information. This will improve standardisation of metadata, which will also be of help in another recently initiated work, i.e. the coordination of metadata between SSB and other Norwegian institutions producing statistics.

6. Conclusions

Most of the improvement in quality due to more integrated metadata systems will be in the areas of accessibility, clarity, comparability and coherence. All in all, this will help us make the most of our resources inside SSB, and also give our external users easier access to, and better understanding of, our statistics.

There will of course be technical challenges connected to the implementation of an integrated metadata system, but we believe that the most challenging problems will be related to human and organisational issues. This is supported by the work on adoption issues within the MetaNet project, see Byfuglien and Linnerud (2003). It is challenging for IT and methodology people to communicate with the producers and users of metadata. The terms used in metadatafora do not always reflect the language of the user, and it will therefore be important to find a communication platform where metadata experts and statistical experts can understand each other. It may for example be better to use the term documentation than metadata in this communication.

The resources allocated for documentation are often scarce, and with tight production schedules, it is difficult to prioritise documentation. Documentation work has to be an integrated part of statistics production. In order to obtain this, the following issues will be addressed in addition to good technical solutions and procedures for handling metadata:

- Leadership commitment for solutions and rules/procedures
- Integration of documentation issues into all relevant planning and follow up procedures in the institution
- Promotion of documentation through other initiatives such as the systematic quality work
- Training and human resource development
- Handbooks and best practices, on statistics production processes and on project work

The new metadata strategy may also result in changes in the organisation of statistics production. Organisational changes are often difficult and time consuming, and it is important that the benefits of the changes are not too far ahead in time. Without losing the vision of an integrated metadata system that will contribute to better quality and more efficient statistics production, we will try to follow a stepwise development where each step results in specific improvements for the people involved.

References

Byfuglien, J. and Linnerud, J.A. (2003): "MetaNet Survey on Statistical Metadata". Eurostat Metadata Production and Exchange Workshop, Luxembourg, April 3-4 2003.

Eurostat (1998): "Definition of quality in statistics". Doc. Eurostat/A4/Quality/98/General Definition.

Statistics Norway (2004a): "Statistical metadata on the Internet revisited". Invited paper to the Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, February 9-11 2004.

Statistics Norway (2004b): "Variables documentation system in Statistics Norway". Contributed paper to the Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, February 9-11 2004.

Statistics Sweden (2004): "Metadata systems in statistical production processes - for which purposes are they needed, and how can they best be organised?" Invited paper to the Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, February 9-11 2004.

Sæbø, H.V., Byfuglien, J. and Johannessen, R. (2003): "Quality Issues at Statistics Norway". Journal of Official Statistics, Vol. 19, No. 3, 2003, pp. 287-303.

UNECE (2000a): "Terminology on Statistical Metadata", Conference of European Statisticians - Statistical Standards and Studies No 53.

UNECE (2000b): "Guidelines for Statistical Metadata on the Internet", Conference of European Statisticians - Statistical Standards and Studies No 52.

Recent publications in the series Documents

2002/5 P. Boug, Å. Cappelen and A. Rygh Swensen: Expectations and Regime Robustness in Price Formation: Evidence from VAR Models and Recursive Methods

2002/6 B.J. Eriksson, A.B. Dahle, R. Haugan, L. E. Legernes, J. Myklebust and E. Skauen: Price Indices for Capital Goods. Part 2 - A Status Report

2002/7 R. Kjeldstad and M. Rønsen: Welfare, Rules, Business Cycles and the Employment of Single Parents

2002/8 B.K. Wold, I.T. Olsen and S. Opdahl: Basic Social Policy Data. Basic Data to Monitor Status & Intended Policy Effects with Focus on Social Sectors incorporating Millennium Development Goals and Indicators

2002/9 T.A. Bye: Climate Change and Energy Consequenses.

2002/10 B. Halvorsen: Philosophical Issues Concerning Applied Cost-Benefit Analysis

2002/11 E. Røed Larsen: An Introductory Guide to the Economics of Sustainable Tourism

2002/12 B. Halvorsen and R. Nesbakken: Distributional Effects of Household Electricity Taxation

2002/13 H. Hungnes: Private Investments in Norway and the User Cost of Capital

2002/14 H. Hungnes: Causality of Macroeconomics: Identifying Causal Relationships from Policy Instruments to Target Variables

2002/15 J.L. Hass, K.Ø. Sørensen and K. Erlandsen: Norwegian Economic and Environment Accounts (NOREEA) Project Report -2001

2002/16 E.H. Nymoen: Influence of Migrants on Regional Varations of Cerebrovascular Disease Mortality in Norway. 1991-1994

2002/17 H.V. Sæbø, R. Glørsen and D. Sve: Electronic Data Collection in Statistics Norway

2002/18 T. Lappegård: Education attainment and fertility pattern among Norwegian women.

2003/1 A. Andersen, T.M. Normann og E. Ugreninov: EU - SILC. Pilot Survey. Quality Report from Staistics Norway.

2003/2 O. Ljones: Implementation of a Certificate in Official Statistics - A tool for Human Resource Management in a National Statistical Institute

2003/3 J. Aasness, E. Biørn and T. Skjerpen: Supplement to Distribution of Preferences and Measurement Errors in a Disaggregated Expenditure System

2003/4 H. Brunborg, S. Gåsemyr, G. Rygh and J.K. Tønder: Development of Registers of People, Companies and Properties in Uganda Report from a Norwegian Mission

2003/5 J. Ramm, E.Wedde and H. Bævre: World health survey. Survey report.

2003/6 B. Møller and L. Belsby: Use of HBS-data for estimating Household Final Consuption Final paper from the project. Paper building on the work done in the Eurostat Task Force 2002

2003/7 B.A. Holth, T. Risberg, E. Wedde and H. Degerdal: Continuing Vocational Training Survey (CVTS2). Quality Report for Norway.

2003/8 P.M. Bergh and A.S. Abrahamsen: Energy consumption in the services sector. 2000

2003/9 K-G. Lindquist and T. Skjerpen: Exploring the Change in Skill Structure of Labour Demand in Norwegian Manufacturing

2004/1 S. Longva: Indicators for Democratic Debate - Informing the Public at General Elections.

2004/2 H. Skiri: Selected documents on the modernisation of the civil registration system in Albania.

2004/3 J.H. Wang: Non-response in the Norwegian Business Tendency Survey.

2004/4 E. Gulløy and B.K Wold: Statistics for Development, Policy and Democracy. Successful Experience and Lessons Learned through 10 years of statistical and institutional development assistance and cooperation by Statistics Norway

2004/5 S. Glomsrød and L. Lindholt: The petroleum business environment.

2004/6 H.V. Sæbø: Statistical Metadata on the Internet Revised.

2004/7 M. Bråthen: Collecting data on wages for the Labour Force Survey – a pilot

2004/8 A.L. Brathaug and E. Fløttum: Norwegian Experiences on Treatment of Changes in Methodologies and Classifications when Compiling Long Time Series of National Accounts.

2004/9 L. Røgeberg, T. Skoglund and S. Todsen: Report on the project Quality adjusted input price indices for collective services in the Norwegian national accounts. Report from a project co-financed by Eurostat.