Statistics Norway
Statistisk sentralbyrå

Statistics Norway
Director General

Hans Viggo Sæbø

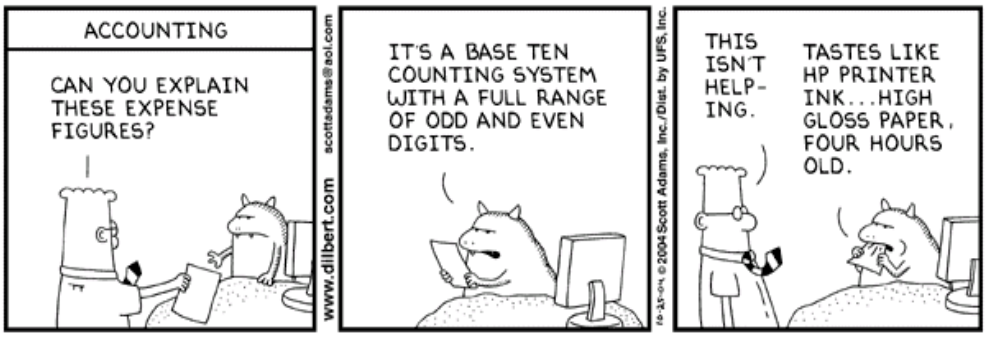
Documents

Metadata strategy in Statistics Norway

Eurostat Metadata Working Group
Luxembourg, 6-7 June 2005

Abstract

Statistics Norway has developed many different metadata systems. 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 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. To facilitate this Statistics Norway has developed a metadata strategy. Our aim is that metadata should be updated in one place and accessible everywhere. The paper describes the work on the main elements of Statistics Norway's metadata strategy, which has been approved early in 2005. In short, the strategy focuses on establishing a conceptual framework, clear roles and responsibilities, and a stepwise development involving integration and linkage of systems. Linkage to international work on metadata standards (e.g. SDMX) is emphasized in the paper, and issues linked to the implementation of the strategy are discussed.



1. Introduction

The following paragraph is from Statistics Norway's current strategy (Strategy 2002-):
" In itself data are not knowledge, but must be summarised and interpreted in order to give knowledge. Statistics imply a certain summing up, but may be useless if not accompanied with definitions and information on how the statistics are produced and can be used."

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. 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, and has led to a need for a metadata strategy. Such a strategy followed by a metadata plan was worked out in 2004 and approved early in 2005.

This paper describes briefly the process towards and the main elements of the metadata strategy of Statistics Norway. 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.

Several recent papers from Statistics Norway are particularly relevant for parts of this paper. Statistics Norway (2004a) discusses metadata that are needed when searching for and using statistics available on National Statistical Institutes' web services. Statistics Norway (2004b) and (2004c) present 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. The work on metadata and metadata systems in Statistics Norway is discussed within a framework of quality and systematic quality work in Statistics Norway (2004d). Our metadata work has also been described in a paper presented at the Eurostat EPROS meeting in 2004, see Statistics Norway (2004e).

2. The metadata strategy

Objectives
The objective of Statistics Norway's work on metadata is to develop an integrated metadata system that shall contribute to effective statistics production and dissemination, in addition to improved quality of statistics.

The short-term objective was to develop a metadata strategy with a plan for the next few years.

The strategy shall ensure that different metadata systems can be linked together in an integrated system, to make the metadata easily accessible for all users and updated in one place. Harmonisation and standardisation of our documentation are key words in this context. The metadata plan comprises almost all Statistics Norway's common metadata and contains several project proposals.

Definitions
The strategy has used the following definitions for the main concepts:

- *Statistical metadata* is structured or systematic information that is used for the production, dissemination, understanding, finding and (re-)use of statistics
- A *metadata system* is a processing system that uses, stores and produces metadata

Both definitions are based on definitions given in SDMX (2005)[1]. One should note that the definition of metadata is broad, covering metadata for users of statistics, producers and data providers, as well as metadata necessary for machine-to-machine communication. These metadata user groups will have different requirements for metadata, but some of these metadata will be common for several users and uses.

The strategy process

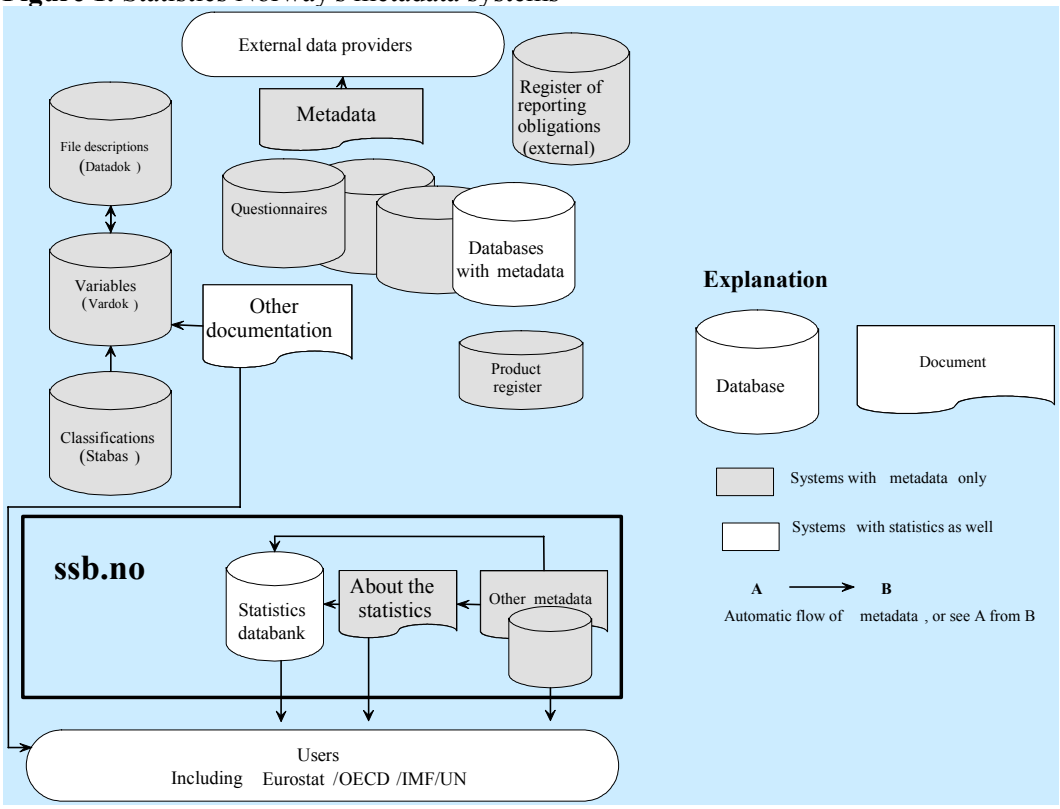
The metadata strategy has been worked out by a project group[2] reporting through a steering committee headed by the head of the Department for IT and Data Collection. However, the needs of many different metadata users and producers have been identified through interviews with about 25 persons in Statistics Norway: Managers, statistics producers, disseminators, IT and methodological specialists. The work in the project was concentrated on Statistics Norway's common metadata and systems, but this mapping also identified possible improvements with regard to the handling of all metadata in Statistics Norway. All respondents were positive to the interview process and contributed constructively during the talks.

The project has also been based on information acquired through written sources (on paper or on web). International cooperation has to be mentioned as a source for inspiration and ideas (such as the work in SDMX, Eurostat Quality LEG, the FASTER, MetaNet and CODACMOS projects and the Neuchâtel group). Statistics Norway's relevant committees (Metadata forum, the IT and Standards Committees) have given advice.

Status

Figure 1 shows Statistics Norway's main metadata systems when the new strategy was adopted early in 2005.

Figure 1. Statistics Norway's metadata systems



[1] Earlier versions of SDMX (2005) were used in the project, but these definitions have not been changed

[2] Project group: Hans Viggo Sæbø (project manager), Anne Gro Hustoft, Jenny Linnerud, Morten Q. Andresen, Thomas Hoel and Hans-Kristian Torvbråten.

Statistics Norway's main metadata systems are

- About the statistics - a systematic description of the statistics published on the web
- Classification database (Stabas) - available on the web (www.ssb.no/english/stabas)
- Variables documentation system (Vardok)
- Data documentation system (Datadok)
- Systems to generate metadata for external data providers

The Statistics databank (StatBank) should also be mentioned in this connection since metadata is a central part of this database (example in figure 3).

Main elements in a metadata strategy

The main elements in the metadata strategy are linked to three areas:

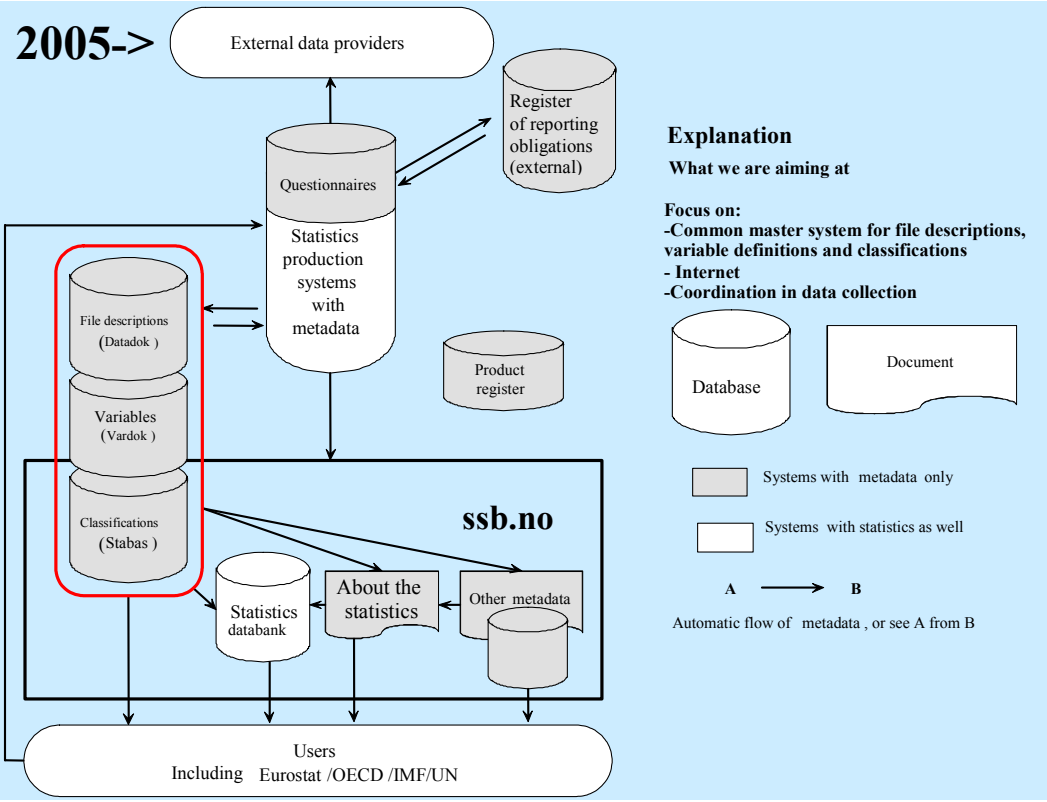
- Common understanding through establishment of documentation and concepts linked to metadata
- Clear roles and responsibilities
- Stepwise development of content and functionality in master systems for metadata

The strategy comprises a number of activities. Some of these started before the strategy was worked out, but they will have to take the strategy into account in further development:

- Definitions of key concepts linked to metadata
- Development of master systems
    - Further development of functionality and content of the master systems for standard classifications (Stabas), variable definitions (Vardok) and file descriptions (Datadok)
    - Coordination and linkage of the master systems and other data and metadata systems
- Other metadata systems and activities
    - About the statistics
    - Information about micro-data
    - Data collection metadata systems
    - Systems for search
    - Framework for metadata
    - Intranet
    - Administrative systems
    - Links to external systems (including systems in Eurostat, OECD, IMF)
- Other measures
    - Training
    - Measurements/indicators of performance (to facilitate improvements)

Figure 2 shows important elements of Statistics Norway's metadata systems that we are moving towards in 2005. Focus has been put on the linkage of master systems for classifications, variable definitions and file descriptions, that more systems and metadata will be available on the external web and on coordination of data collection. The standard classification system (Stabas) is already available with central classifications on the external web, but more classifications and variable definitions are expected to be available here during 2005. Some of the activities that are described in the plan will last longer than 2005, and the figure is meant as a goal we should aim for.

Figure 2. Development of metadata systems in Statistics Norway



3. Metadata on a national and international level

We believe that the challenges linked to metadata in Statistics Norway are typical for those of a national statistical institute (NSI), with lack of common concepts and fragmentation of systems resulting in double work and poor quality. On an international level there are similar challenges linked to different data and metadata systems, and requirements from different international organisations. Requirements and solutions for data and metadata transfer are relevant issues on both national and international level. The SDMX initiative is important in this setting, but elements of this work should also benefit (and benefit from) the national work and solutions. Examples are the work on common frameworks (such as quality frameworks) and vocabularies, in addition to technologies and formats that can be used in data transfer between NSIs and international or other national organisations, and within NSIs. In Statistics Norway we will consider SDMX developments for both external data and metadata transfer purposes, as well as for internal solutions within the framework of the metadata strategy. However, when technology has been developed and adopted, it is crucial that the new solutions are actually used in practice, for example that international organisations are able and willing to receive data and metadata in the recommended formats.

4. Implementation of the metadata strategy

Several of the projects and activities proposed in the metadata strategy and plan have already started. This applies to the content and further development of the metadata master systems Stabas, Vardok and Datadok, as well as the linking of these and other existing systems. The organisation and content of this and further work are considered in the following.

Organisation

Several of the activities proposed in the metadata strategy and plan are organised as projects with steering committees. However, to ensure coordination and to put more pressure on the implementation, it has been decided to set up a separate steering committee in order to follow up the

metadata strategy. In practice, this committee is the same as the steering committee for the strategy project.

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 (especially the ones aimed at data providers, figure 1), and one of the future challenges will be to fit these metadatabases into the overall system without losing the benefits and routines that are connected with them. In order to ensure this, the overall steering committee will require information and follow-up on all projects related to metadata regularly.

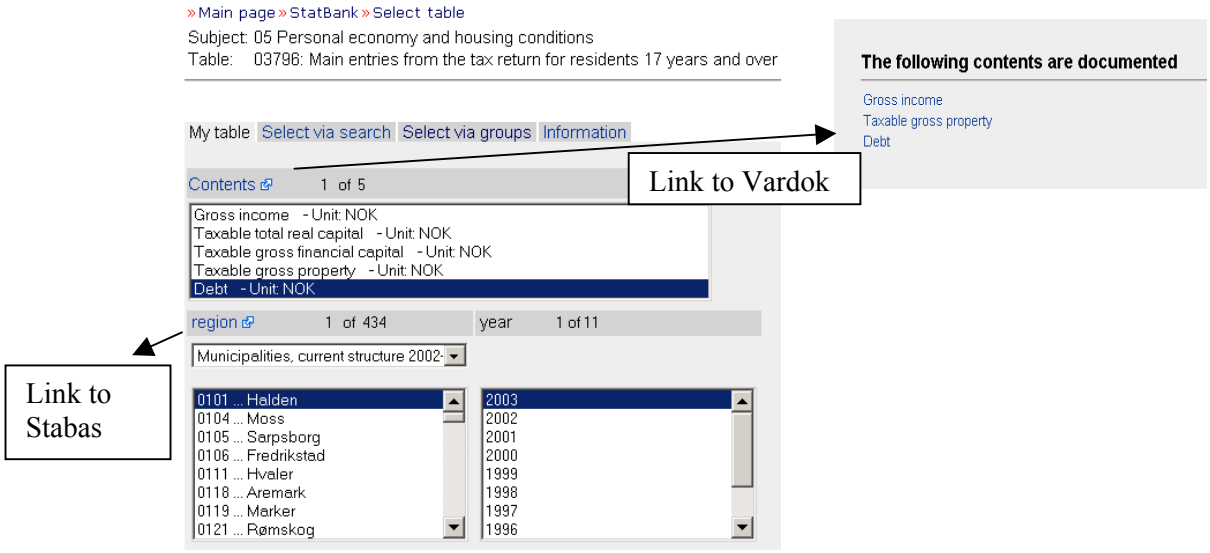
Services and linking of metadata systems
Development of functionality and services providing users easy access to and use of the metadata systems is central in the metadata strategy. As mentioned, the linking of existing systems has already started. This has even now improved accessibility, and should 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 directly from Vardok. 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. You can also have a view of the variable definition stored and updated in Vardok when working on a file description in Datadok.

Our dissemination database (StatBank) has become central in all our dissemination. Users can construct, select and download statistical tables. StatBank is based on cubes with variables grouped by different classifications. The databank can be reached from our daily statistics: "More tables in StatBank". From the databank there is a link to "About the statistics" and links to "Vardok" and "Stabas" are being established. Figure 3 shows an example of the user interface of the selection page of StatBank, with some of the links under development.

Figure 3. Link from StatBank to Vardok



Efficient exchange of data and metadata across different system platforms is a major issue. In 2005 Statistics Norway will prioritise the building of an IT platform based upon principals for service

oriented architecture, so that in the next round we can build new solutions and services in a uniform and well thought through manner.

Conceptual framework
The establishment of concepts linked to metadata is crucial both for statisticians and others to know what, how and where to put metadata in the different systems. One example is a confusion we have experienced about the concepts "variable" and "classification". What should be included and explained in Vardok and what should be stored in Stabas? Which concepts should be used in the Statistics databank to facilitate a simple and unified structure of cubes and tables? A small group has been set up to define the key quality and metadata concepts, while a project on the metadata framework will include relationships between and more details on these concepts. The concepts will be based on the definition in the SDMX metadata common vocabulary, and existing Norwegian concepts will be related to these. The set of key concepts will be limited, but include the concepts quality, quality (dimensions) in statistics, statistics, register, variable (statistical, classification and identifying variable), classification, code, statistical unit (observation, reporting and analysis unit), measurement unit, population and frame. It has already been experienced that even if it might be easy to agree on some basic definitions such as those given by SDMX, the challenge is to apply these in practice, in a statistical table or on a micro level. One of the reasons for this is that concepts can be interpreted differently on these levels.

Metadata and quality
Most of the improvement in quality due to more integrated metadata systems will be for internal users, who will experience improved accessibility, clarity, comparability and coherence of data. This, in addition to elimination of double work, will improve the efficiency of our production processes. External users should also benefit by easier access and better understanding (clarity) of our statistics. This should also benefit international organisations.

One project aims particularly at an external user group: Information about micro-data. Sensitive micro data may be made available for planning and research purposes, given that the researchers have a licence from the Data Inspectorate, and that the data are de-identified. The information about and hence the accessibility and clarity of micro-data has been poor, and in this case better metadata should provide a considerable quality gain. However, in this project as in others, it is crucial that information and links are based on metadata in existing systems or developed within the frame of the metadata strategy.

5. Conclusions

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. It is challenging for IT and methodology people both to communicate with each other and with the producers. The terms used in metadata discussions do not always reflect the language of the producer, and the work on concepts is therefore crucial.

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

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

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 lie 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.

In a longer term we will aim at more integration and a few more metadata systems (that will be a part of the total metadata system), and more transfer of metadata between systems with metadata and statistics. Compared to the 2005 figure a complete and integrated metadata system will also comprise:

- Systems with administrative metadata
- All systems linked to data collection
- Metadata linked to all registers and databases
- Systematic documentation of all concepts (for example the different statistical units, in addition to variables and classifications)
- Better links to external systems (such as the systems of Eurostat and other international organisations)

References

Statistics Norway (2004a): "Statistical metadata on the Internet revisited". Prepared by Hans Viggo Sæbø. 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". Anne Gro Hustoft and Jenny Linnerud. Contributed paper to the Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, February 9-11 2004

Statistics Norway (2004c): "Development of a variables documentation system in Statistics Norway". International Statistical Conference "Investment in the future". Anne Gro Hustoft and Jenny Linnerud. Prague, September 6-7 2004

Statistics Norway (2004d): "Quality and metadata in Statistics Norway". Anne Gro Hustoft, Jenny Linnerud and Hans Viggo Sæbø. Quality 2004, Mainz, May 24-26 2004

Statistics Norway (2004e): "Norwegian experiences and future challenges to improve data collection and metadata management". Rune Gløersen and Hans Viggo Sæbø. Eurostat EPROS meeting November 9 2004

SDMX (2005): "SDMX Metadata Common Vocabulary". Draft 2nd release, February 2005

Recent publications in the series Documents

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 og 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 - 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 Indicies for Collective Services in the Norwegian National Accounts. Report from a Project Co-financed by Eurostat.

2004/10 A-K. Mevik: Uncertainty in the Norwegian Business Tendency Survey.

2004/11 A.G. Hustoft, J. Linnerud and H.V. Sæbø: Quality and metadata in Statistics Norway.

2004/12 E. Engelien, R. Klæboe and Margrete Steinnes: Neighbourhood sonoscapes. Context sensitive noise impact mapping .

2004/13 Petter Vegard Hansen: Regional electricity spot price responses in Norway.

2004/14 A.G. Hustoft and J. Linnerud: Development of a variables documentation system in Statistics Norway. International Statistical Conference "Investment in the future", Prague, Czech Republic, 6-7 September 2004.

2004/15 J.L. Hass: Compilation of data on expenditure in Environmental protection by businesses. Report to the European Commission DG for Environment.

2004/16 A. Raknerud, J. Rønningen og T. Skjerpen: Documentation of the capital database. A database with data for tagible fixed assets and economic data at the firm level.

2004/17 B.K. Wold D. Roll-Hansen A. Mathiassen and S. Opdahl: A Sustainable Household Survey Based Poverty Monitoring System. A Poverty Monitoring System Based upon Household Survey Estimation of Total Consumption. A Preliminary Paper Asking for Cooperation

2004/18 T. Karlsen, D. Quang Pham and T. Skjerpen: Seasonal adjustment and smoothing of manufacturing investments series from the quartely Norwegian national accounts

2005/1 V. Skirbekk: The Impact of a Lower School Leaving Age and a Later Retirement on the Financing of the Norwegian Public Pension System.

2005/2 H. Utne: The Population and Housing Censushandbook 2001.

2005/3 J. L.Hass and R. Straumann: Environmental Protection Expenditure: Methodological work for the Oil and Gas Extraction Industry. Report to Eurostat.

2005/4 L. Hobbelstad Simpson: National Accounts Supply and Use Tables (SUT) in Constant Prices SNA-NT "SUT/CONSTANT"

2005/5 L. Hobbelstad Simpson: National Accounts Supply and Use Tables (SUT) in Current Prices. SNA-NT "SUT/STARTER"

2005/6 S. Todsen: SNA-NT User's Guide for Supply and Use Tables in Current and Constant Prices.