WP.28
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 (iv): Using metadata for searching and finding statistical data in websites and portals

STATISTICAL METADATA IN STATISTICS NORWAY
Contributed paper

Submitted by Statistics Norway[1]

I. INTRODUCTION

1. Development of functionality and services providing users easy access to and use of the metadata systems is central in Statistics Norway's (SSB) metadata strategy. This paper will focus on the content and development of a dedicated area for statistical metadata on SSB's website. The overall aim of this development is to make the contents of all our metadata systems more accessible and easier to use. The aim for 2006 is to test the statistical metadata area with users inside SSB. The contents of our variables, classifications and file descriptions servers will be displayed in this area. The design will be flexible so that the contents of other metadata systems can be added when appropriate.

II. METADATA STRATEGY

2. Statistics Norway has developed many different metadata systems to serve different purposes and different user groups. 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, which was approved early in 2005, together with a metadata plan for the next few years [1].

3. The strategy focuses on establishing a common understanding through establishment of documentation and concepts linked to metadata, clear roles and responsibilities, and a stepwise development of content, integration and linkage of master systems for metadata. Our aim is that metadata should be updated in one place and accessible everywhere. Our metadata systems should also serve as tools for the harmonisation and standardisation of our documentation.

4. The metadata plan comprises almost all Statistics Norway's metadata and contains several project proposals. Some of these projects started before the strategy was finalised, but they are now taking the strategy into account in further development:

- Definitions of key concepts linked to metadata

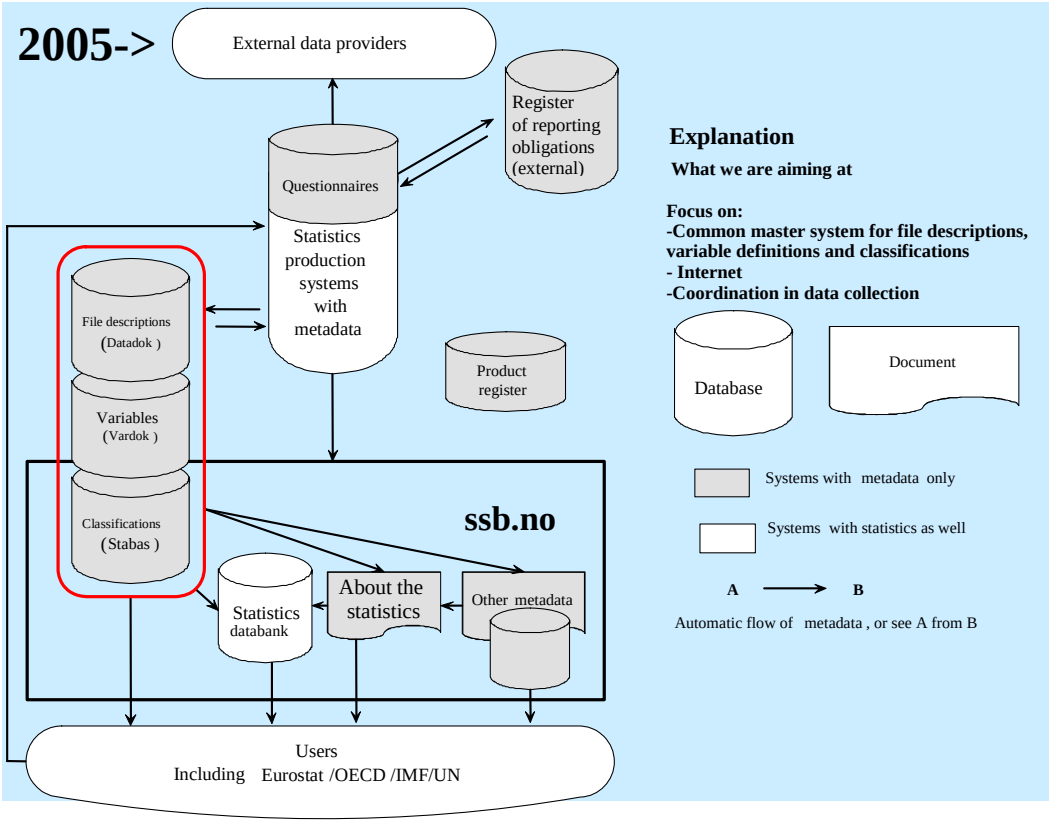
[1] Prepared by Anne Gro Hustoft (agt@ssb.no) and Jenny Linnerud (jal@ssb.no)

- Development of master systems
  - Further development of functionality and content of the master systems for standard classifications (Stabas), variables 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
  - For more proposals see [1].

5. A small group was set up to define key quality and metadata concepts. The set of key concepts was limited, but included the concepts quality, quality (dimensions) in statistics, statistics, register, variable, indicator, indicator, classification, code list, statistical unit (observation, reporting and analysis unit), measurement unit and population. It has already been experienced that even if it might be easy to agree on some basic definitions such as those given by the SDMX Metadata Common Vocabulary and existing Norwegian concepts related to these, the challenge is to apply these in practice, in a statistical table or on a micro level. 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.

6. Figure 1 shows important elements of Statistics Norway's metadata systems that we are moving towards. Focus has been put on the linkage of master systems for classifications, variables definitions and file descriptions, that more systems and metadata will be available on the external web and on coordination of data collection.

Figure 1. Development of metadata systems in Statistics Norway



7. Efficient exchange of data and metadata across different systems and platforms is a major issue. In 2005 Statistics Norway prioritised a service oriented architecture. 9 services for variables, 3 services for file descriptions and 3 services for classifications were prototyped. The introduction of these services has been so well taken up by the organisation that the demand is currently exceeding the supply. Further development of services is planned to continue until 2008. We aim to link all our metadata systems by services thereby improving the availability of updated information.

III. STATISTICAL METADATA WEB PAGE

8. The overall purpose of the Statistical metadata web page is to make Statistics Norway's metadata systems more accessible and easier to use. Both internal and external users will get easier access to the metadata by displaying the contents of these systems in a common web page. Our work within this area has been inspired by the corresponding web pages of Statistics Canada (www.statcan.ca/english/concepts) and Statistics New Zealand (www.stats.govt.nz/statistical-methods).

9. The main purpose of the web page is to give access to information stored in the metadata systems and delivered by web services but the page will also contain links to other relevant metadata. This year we will establish links to the metadata connected to each published statistics ("About the statistics"), metadata about our data collections ("About the data collection"), questionnaires and other relevant documentation. The contents of the web page will be extended as other relevant metadata becomes available.

10. The figure below gives a picture of the links we intend to include on our web page. All these metadata exist at present, but they are stored in different systems and different places on the web. They are difficult to find for the inexperienced user.

Figure 2. Links in the statistical metadata web page

About the statistics	Standards
• Listed by division	• Classifications (Stabas)
• Listed by subjects	• Nomenclature and documents
• Advance release calendar	
	File descriptions and registers
About the data collection	• File descriptions (Datadok)
	• Registers (Datadok)
Definitions	
• Statistical units	Statistical methods
• Variables (Vardok)	• Articles
• Terminology	• Reports
Questionnaires	International links
	• Definitions
	• Standard classifications
	• Statistical methods

11. A project group consisting of three IT-specialists and one standards/metadata specialist is in charge of developing the web page. In addition specialists within different subject matter areas will test the system during the development process. The project group will also forward a plan for organisation and quality assurance connected to the operation of the underlying metadata systems.

12. Our aim is to make the statistical metadata web page accessible to users within Statistics Norway in 2006, and to make a version for external users in 2007. The external version of the statistical metadata web page (2007) will also be available in English. Planned resources for the development of an internal statistical metadata web page (2006) is 1300 man-hours for the IT-specialists and 300 man-hours for the standards/metadata specialist who is also the project manager.

IV. STATISTICAL METADATA

13. In this section we give some more information about the underlying metadata systems and other relevant metadata that will be made available through our web page.

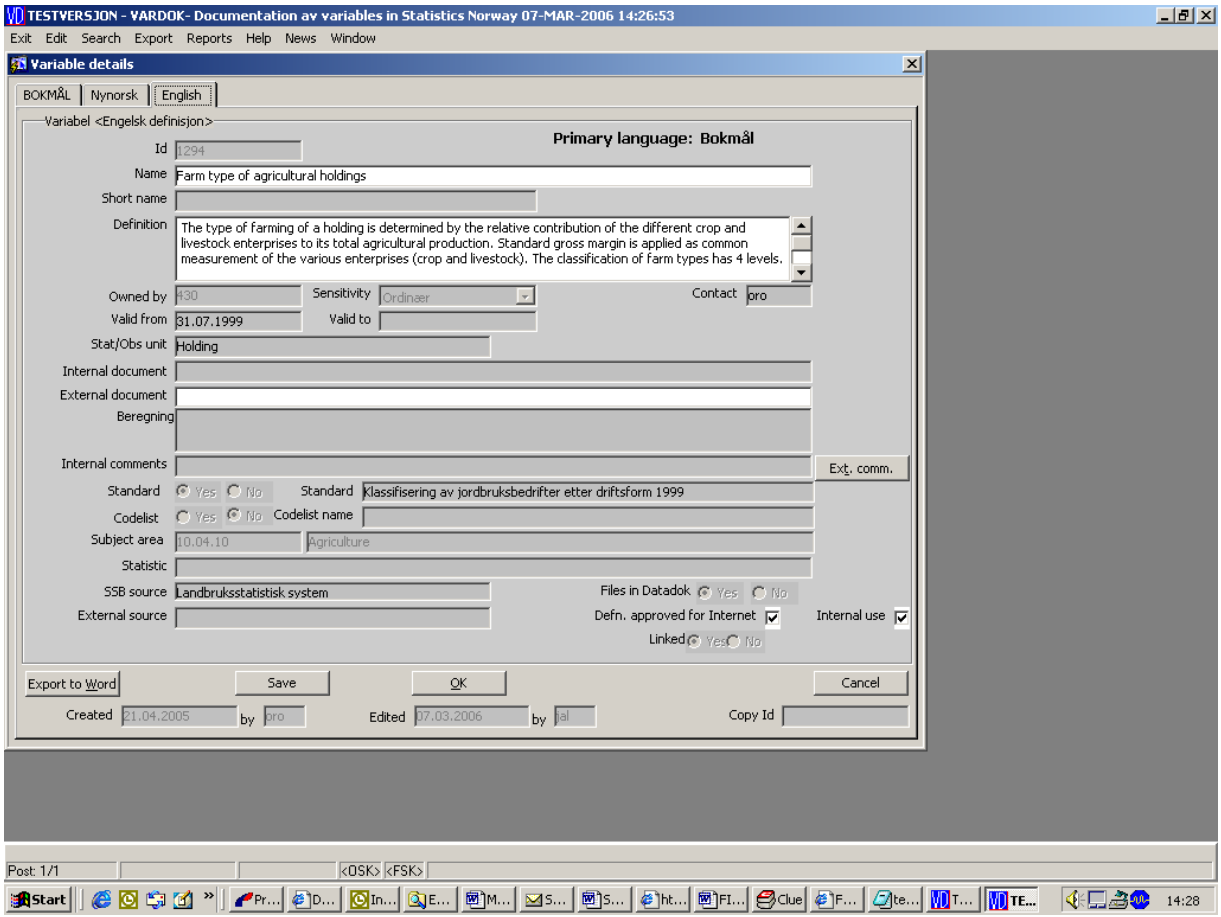
A. Documentation of variables (Vardok)

14. At the METIS-meeting in 2004 Statistics Norway presented a system for documentation of variables (Vardok) [2]. This is a central system for documenting variables (e.g. definition, validity periods, classifications used) and also a tool for harmonisation of names and definitions of variables.

15. Since 2004 the Vardok system has been further developed by introducing the possibility of multilingual functionality (English and two different versions of Norwegian), of describing variables by equations and of linking different variables (if variable A is the sum of variable B and variable C, you don't have to define B and C when you define A, you just link to the definitions of B and C).

16. This year Vardok will be linked to our event-history database and About the statistics. Last year it was linked to our system for dissemination of statistics (Statbank). The screenshot below shows the current situation for Vardok regarding information fields. Only four information fields are available for translation to English. The rest should be automatically translated or are deemed unnecessary.

Figure 3. Information fields in Vardok



17. We are also about to finish a web-version of the variables documentation system.

18. An important part of our work the last two years has been documentation and quality assurance of new variables. At present (March 2006) 1200 variables are documented in Vardok. 1104 of these are approved for dissemination within Statistics Norway, 421 are approved for dissemination outside Statistics Norway (the figure below shows the development in the number of variables documented). The variables that are not approved for dissemination are still being discussed within the subject matter divisions that are responsible for them. As soon as the variables are approved for dissemination outside Statistics Norway, they can be displayed on the Internet through Statbank, About the statistics and About the data collection.

Figure 4. Progress in Vardok

Year	Number of variables documented per year	Number of divisions with access per year
2002	157	5
2003	352	5
2004	323	6
2005	284	11
2006	84 so far	12
Total	**1200**	**-**

19. The number of divisions with access to Vardok has been restricted (12 of maximum 18 subject matter divisions in 2006) but as soon as the division flags their variables as approved for internal use, all other divisions can give feedback on the definitions.

20. Within the Vardok-project we have made a special effort to start harmonising the accounts statistics variables. The different subject matter divisions in many cases use the same names for their accounts statistics variables, but define them a bit differently, because they are subject to different laws and regulations. The treatment of these variables therefore requires a tighter cooperation between the involved divisions than the documentation of other variables.

21. Parallel to the development of the variables documentation system, SSB participates in the Neuchâtel group[2] where terminology and models connected to variables are discussed.

22. 2006 is the last year in the development phase for the Vardok-project. Next year all 18 divisions will be able to document their variables definitions in Vardok.

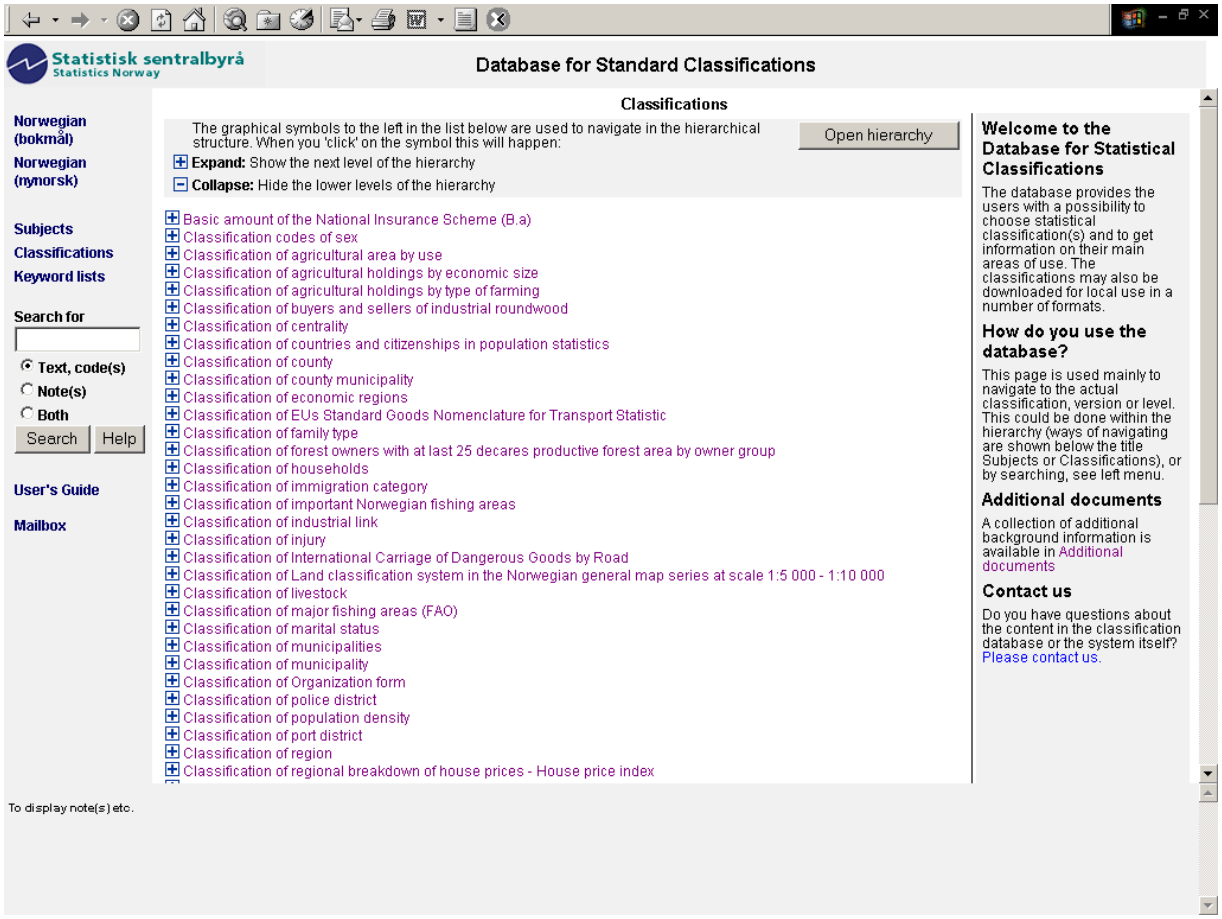
B. Classifications (Stabas)

23. Statistics Denmark and Statistics Norway started a joint project in 2001 to implement the Neuchâtel model in a multilingual standard relational database. The project implemented Version 2.0 of the Neuchâtel terminology for Classification database object types and their attributes with a few attributes being left out. Production started in Statistics Norway in 2002.

24. The model was implemented using Microsoft Access Database for demonstration and testing purposes and Oracle for production purposes. An application in Visual Basic was developed for maintaining the database (creating and updating classifications including variants, indexes and correspondence tables). Import and export functions have been defined to a preliminary XML format. In addition a web-based browser was developed and is used for the dissemination of selected classifications on the Internet.

[2] The Neuchâtel group working with terminology models for classification databases was established in 1999 and consisted of Statistics Denmark, Statistics Sweden, Statistics Switzerland, Statistics Norway and run Software-Werkstatt. Statistics Netherlands and the Bureau of Labor Statistics (USA) have joined the group for the work on variables.

Figure 5. Database for Standard Classifications (www.ssb.no/english/stabas).



25. At present (March 2006) we have 51 current classification versions on our Intranet, 92 older versions and variants on our Intranet and 49 current classification versions on SSB's website. With very few exceptions these classifications are available in three languages (English and two different versions of Norwegian). A one-way link from Vardok to Stabas was established in 2003 and from Statbank to Stabas in 2005.

26. The statistical metadata web page will link up directly to our web-based version of Stabas and provide some administrative reports to be used in keeping the system updated and the quality of the information high.

C. File descriptions (Datadok)

27. We document all permanent data files in our file documentation database Datadok. The database was built in 1998 but wasn't mandatory until 2002. At present (March 2006) we have over 6 500 file descriptions stored there in Norwegian. There is a two-way link between Vardok and Datadok. We are currently running 3 prototype services for Datadok to provide the statistical metadata web page with more easily available and searchable file descriptions.

D. About the statistics

28. About the statistics is metadata that describes each statistics that is published by Statistics Norway. It contains administrative information, information about statistics production, variables, concepts, sources of errors and uncertainty, comparability, coherence and availability.

29. Last year we started implementing About the statistics using an CMS (Content Management system)-platform. There were two important reasons for doing this. The first was that use of CMS will make general updating of About the statistics more easy for the subject matter divisions. The other reason is that CMS will make it possible to link About the statistics to Vardok and Stabas. Then we will no longer define variables in

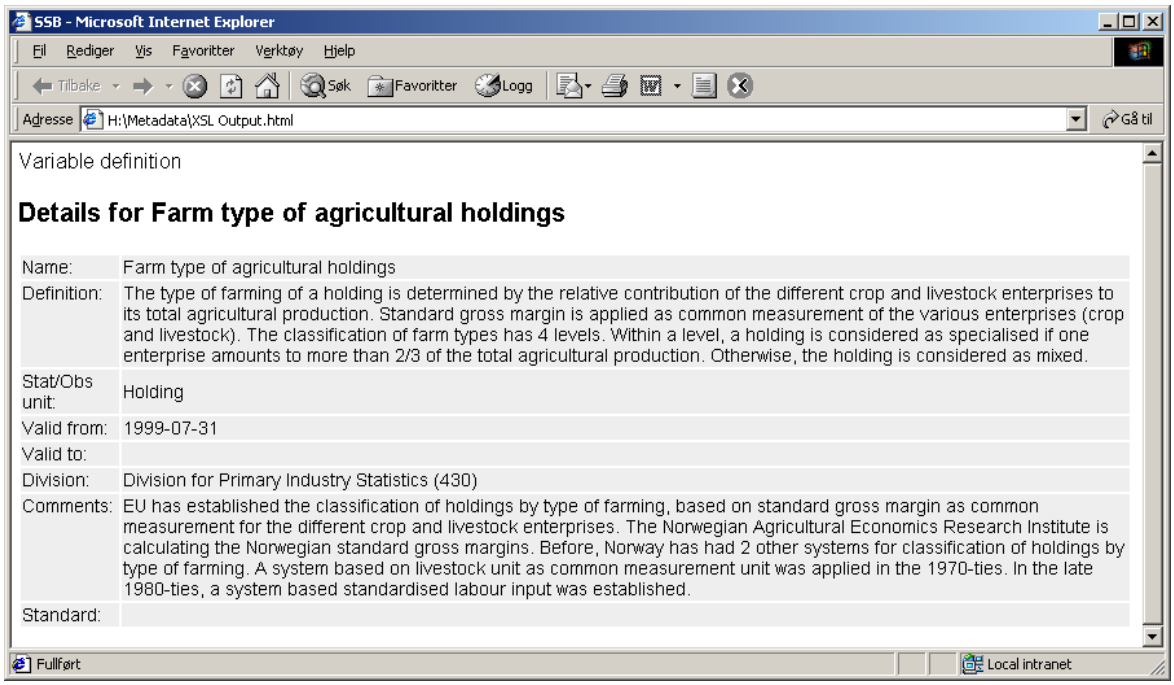
About the statistics, but link to the relevant variables in Vardok. In the same way we will link to the relevant classifications in Stabas. This will bring us closer to the goal that all definitions of variables should be documented and updated in one place (Vardok) and linked to other places where they are needed.

E. About the data collection

30. Researchers frequently use data collections from Statistics Norway for their research. However, the process from finding out what you need, to actually getting the data, may be long and troublesome, especially for inexperienced researchers. Statistics Norway has therefore (with support from the Research Council of Norway) developed a website (soon to be translated into English) to make information about this process more easily available. Among other things, this page provides the users with documentation of several data collections. Each data collection has a general description e.g. of data quality, and it also contains a list of relevant variables, including variable documentation from Vardok.

31. The next figure shows how the variable documentation is displayed through About the data collection on the Internet. If you compare it to fig. 3, you see that About the data collection has chosen to collect only those Vardok-fields that were found relevant for this specific group of users (researchers). At present Statbank only displays the definition information field. This is all that Statbank has found relevant for their specific group of users (mainly members of the public).

Figure 6. Variables documentation in About the data collection



F. Questionnaires

32. In SSB we have developed one online data collection process for businesses (IDUN) and one offline data collection process for reporting from local to central government (KOSTRA). In 2006 we are developing one metadatabase for questionnaires to support data collection from both businesses and local government. We are also re-examining the tools and processes with which we make questionnaires (paper, electronic, computer assisted interviews etc). In 2006 our ambition level is only to make all our questionnaires available on the statistical metadata web page in pdf format. In the future we would like to connect here to a question and questionnaire bank.

G. Other documentation

33. Other documentation will consist of links to international websites (e.g. SDMX Metadata Common Vocabulary, Eurostat's standard classifications in Ramon and Eurostat's definitions in Coded), publications related to statistical methods and other metadata documents (e.g. terminology).

IV. METADATA MANAGEMENT

34. Our plans for organising metadata management are still developing, but we have identified some points that we think will be important to achieve a successful future for the statistical metadata web page:

- All subject matter departments should appoint one person responsible for the topic of metadata. These metadata managers must have thorough knowledge of the metadata systems linked to the page, and should act as advisers to other people in their department. The responsible persons should also work together to solve metadata questions involving cross-departmental subjects.
- All subject matter divisions should have a contact person for Vardok, Stabas and Datadok to provide training and support to new users.
- The Department of dissemination should be involved and secure a unified presentation of the metadata disseminated to external users.

35. This preliminary list of points will be extended during 2006, and will result in a concrete plan for metadata management.

REFERENCES

[1] Metadata strategy in Statistics Norway. Hans Viggo Sæbø. Eurostat Metadata Working Group meeting, Luxembourg, June 6-7 2005.
[2] 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
[3] Neuchâtel v2.1 http://europa.eu.int/comm/eurostat/ramon/miscellaneous/index.cfm?TargetUrl=D