Statistics Norway

Statistisk sentralbyrå

Jenny Linnerud and Anne Gro Hustoft

Documents

Development of a variable documentation system in Statistics Norway

Eurostat, Metadata Working Group
Luxembourg, 7-8 June

Statistics Norway/Department of IT and Data Collection/
Department of Management Support

Abstract
Development of functionality and services providing users easy access to and use of metadata systems is central in Statistics Norway's (SSB) metadata strategy. In addition, SSB's IT strategy emphasises the importance of metadata in the overall information architecture and the use of principals of service oriented architecture for technical metadata solutions. This paper will focus on our master system for documentation of variables that occupies a central place in a dedicated area for statistical metadata on SSB's intranet. The contents of our classifications and file descriptions servers are also displayed in this area and the contents of other metadata systems will be added when appropriate. An almost identical copy will be displayed on our Internet site giving greater access to statistical metadata for external users. Parallel to the development of the variables documentation system, SSB has participated in the Neuchâtel group where terminology and models connected to variables have been discussed.

Introduction
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. There has also been an increasing focus on the need for interoperability with external systems both nationally and internationally.

Metadata strategy (2005 -)
Purpose
The overall aim of SSBs metadata strategy is to create a comprehensive metadata system that will contribute to an efficient production and dissemination of statistics, and at the same time improve the quality of the statistics. Our metadata should be updated in one place and accessible everywhere.

Definitions
In our metadata strategy work we found the following working definitions useful:

- A *metadata system* is a processing system that uses, stores and produces metadata.
- *Statistical metadata* is structured or systematic information that is used for the production, dissemination,

understanding, finding and (re-)use of statistics.

- A *mastersystem* is the system that contains the standard for this type of metadata (e.g. Vardok contains definitions of variables approved for use in SSB).

Project organisation
The project group consisted of 6 people with expertise in standards, total quality management, statistic production, IT development and metadata systems and experience from contextual design. The project started and finished in 2004 and used 1150 man-hours.

Method
The metadata strategy has been worked out by a project group 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.

Current status
The metadata strategy was approved early 2005 and resulted amongst other things in continued support for our variables documentation system and approval of resources for several new projects including definitions of central statistical concepts for use in metadata systems (available only in Norwegian at the present time), the service library for metadata systems and the statistical metadata portal. For more information see reference [1].

Variables documentation system

Purpose
The overall purpose of the variables documentation system is to document variables in a central location, accessible by all, and to function as a tool for harmonising names and definitions.

Project organisation
The development project began in 2000 and ended in 2006. The core project group consisted of 3 people with expertise from standards, metadata systems and IT development. In the first year expertise from the population census was also involved and in the last year dissemination expertise has been involved. A total of 12690 man-hours have been used with ca. 70% of resources from IT.

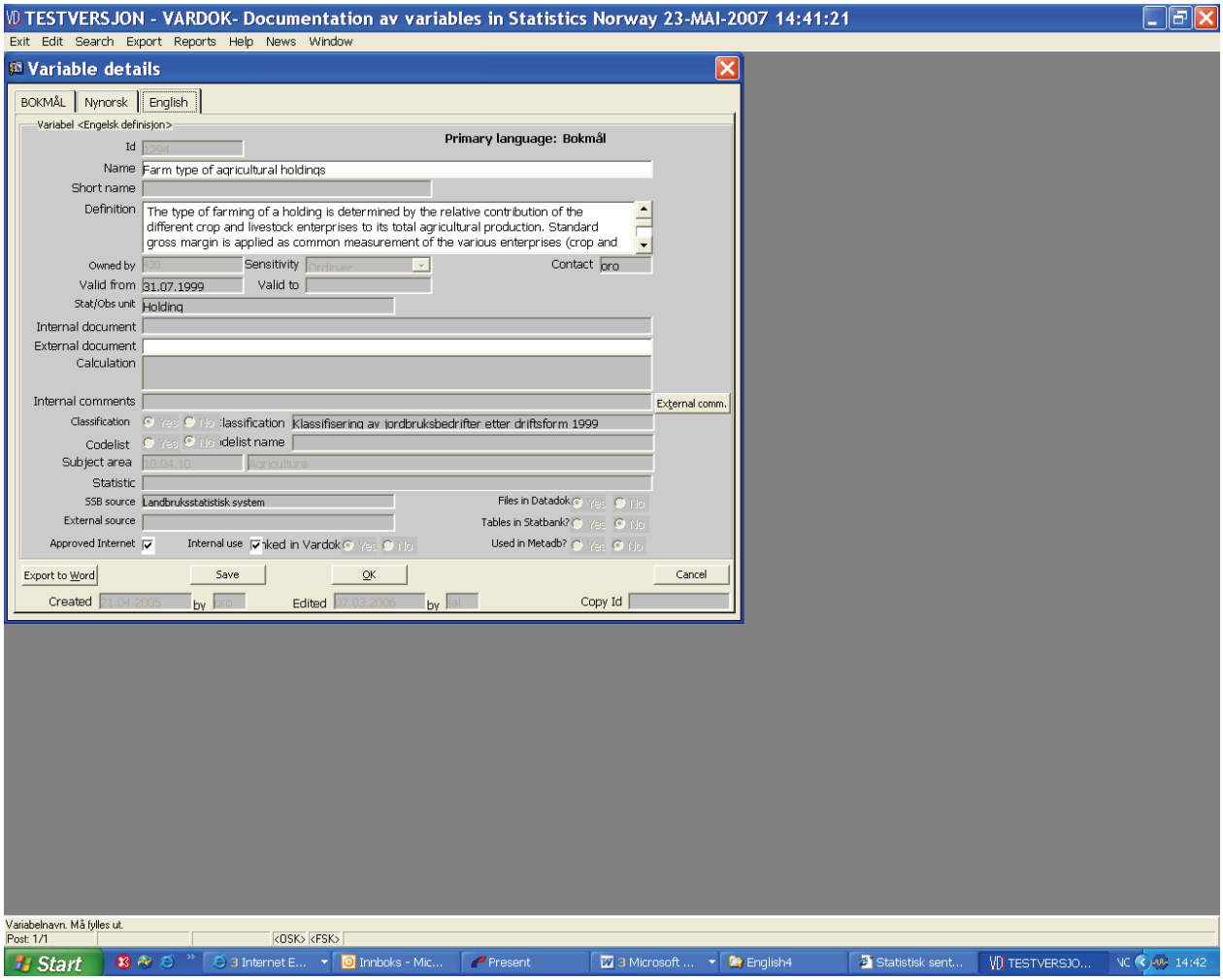
Method
Based on their experience in the European Commissions Information Society Framework V project FASTER (Flexible Access of Statistics, Tables and Electronic Resources), the project group decided to use contextual design [2] for the development of the variables documentation system. User involvement and stepwise development have been key success factors. 15 divisions and 42 people were involved in the design phase (2001-2002).

From 2003 to 2006 we concentrated our development efforts on linking Vardok to other metadata systems:

- A two way link between Vardok and Datadok (file descriptions database) was established in 2002
- A one-way link from Vardok to Stabas (standard classifications database) was established in 2003.
- A two way link between Vardok and StatBank (dissemination database) was established in 2005
- A two way link between Vardok and Metadb (system for documentation of event history data) was established in 2006.
- A one way link from About the statistics, About the data collection and the statistical metadata portal to Vardok, via services, was established in 2006.

Current status

The screenshot below shows the current situation for Vardok regarding information fields.



The following table shows the number of variables documented and the number of divisions involved.

Year	Number of variables documented per year	Number of divisions involved per year
2002	157	5
2003	352	5
2004	323	6
2005	284	11
2006	287	12
2007	8 so far this year	18
Total	**1411**	**All 18 statistics divisions**

At present 1411 variables are documented in Vardok in Norwegian. 1307 of these are approved for dissemination within Statistics Norway while 652 are approved for dissemination outside Statistics Norway.

110 variables are documented in Vardok in English and approved for dissemination. 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. For more information on the development of variables documentation system from 2000 to 2004 see reference [3].

2006 was the last year in the development phase for the Vardok-project. A total of 600 man-hours from standards and dissemination are planned in 2007 for continued harmonisation of names and definitions, and training of personnel in the six new divisions. 200 IT man-hours are planned for maintenance and any minor changes to the system.

Comparison with some other international work
The closest ISO/IEC 11179 [4] concept to our variables concept in Vardok is the Data Element Concept.
The closest Neuchâtel [5] concept to our variables concept in Vardok is the Object variable.

Technical solution
Programming language: PL/SQL
Database: Oracle 9.2
Development tools:

- Oracle Designer to design the database
- Oracle Developer (Oracle Forms 6i) for application development
- TOAD for PL/SQL packages in Oracle

Datadok - File descriptions

Purpose
We document all permanent archive data files in our file documentation database Datadok. The database was built in 1998 but wasn't mandatory until 2002.

Current status

- At present we have over 30 500 file descriptions stored there in Norwegian.
- There is a two-way link between Vardok and Datadok

Stabas - Standard classifications database

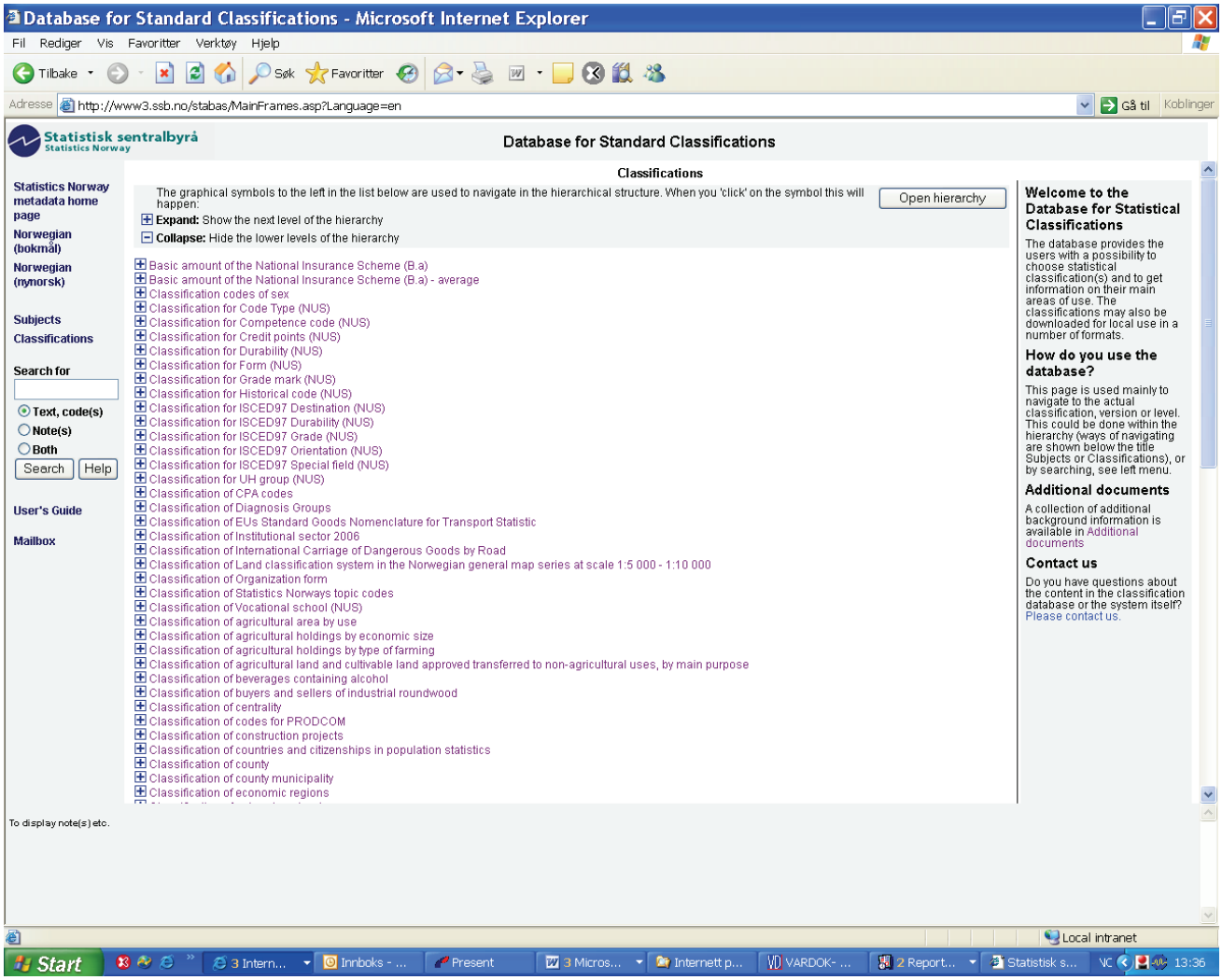
Purpose
The overall aim of Stabas is:

- To make work with and the use of standards simpler and more efficient
- To ensure systematic use of standards across different statistical areas

One main task is to make approved versions of the central statistical standards available in a database system where they can be taken out at different aggregation levels, together with texts in diferent languages and relevant documentation, and where the standards can be exported to other IT tools.

Current status
At present we have 68 current versions on our Internet, 32 older versions on our Internet and 145 versions on our Intranet. There is a one-way link from both Vardok and StatBank to Stabas. Stabas is based upon the Neuchâtel Terminology Model, Classifications database object types and their attributes [6]

The next screenshot shows the current situation for Stabas.



Technical solution

- Oracle database v. 9.2.0 or higher.
  - Including PL/SQL routines
- Maintenance routines, made by Statistics Denmark, using Visual Basic v. 6.0
  - Update application
  - Import/export application, data transfer using XML structures
    - Using perl to generate XML structures from documents for first time import
  - Administration application, internal security.
- Intranet/Internet presentation routines, made by Statistics Norway, using Microsoft ASP (Application Server Pages)
  - Same application for both platforms.
  - Data transfer through firewalls using Oracle export format file..

IT strategy (2007 -)

Statistics Norway has revised it's IT strategy for the next five years [7]. The sections on master systems for metadata, service oriented architecture and system architects are of relevance for this article and are summarised here.

Mastersystems for metadata
Statistics Norway has master systems for the following types of metadata: definitions of variables, classifications, file descriptions for archive data and event history data descriptions. Master systems for questionnaires and rules and planning information for statistics products are under development. These systems will communicate with all phases of statistics production, from data collection through data editing and analysis to dissemination, via metadata services.

Service-oriented architecture
Statistics Norway's technical solutions shall be built mainly upon the principles of service-oriented architecture. Guidelines on this are presented in *Norway's eGovernment plan*. All solutions for external users and most solutions for internal users shall:

- Have support for *open standards*.
- Be *platform independent*.
- Be *component based*.
- Have support for the packing in of data and functions in the form of *services* (web services).

These are central principals in service-oriented architecture. By applying these principles, applications and services can reuse existing functionality/components completely independent of the system they were developed in. In addition, by use of this technique, we can extend the lifetime of older applications, which have important functionality we wish to expose, just by creating a service layer on top of these. This increases the possibilities for collaboration between old and new applications in a completely new way, which gives benefits in the form of shorter development time, increased reuse and more consistent systems. This also enables us to replace systems behind the scenes, because communication with these is not directly exposed to the users.

System architects
System architects will be introduced for each of the areas in the top-level information architecture: data collection, metadata, dissemination, population management, data editing and analysis. The mandate for this role will be made and will support the system architect's responsibility to ensure that IT development projects are in line with the IT strategy.

Service library for metadata systems
Purpose
The purpose of this project is to

- Create a library of services for the master systems Vardok, Datadok, Metadb and Stabas.
- Define a framework for the description and formulation of SSB's metadata based on international metadata models (e.g. Neuchâtel) and standards (e.g. ISO/IEC 11179).
- Investigate how RDF (Resource Description Framework) can be integrated into SSB's data communication.

Project organisation
The project began in 2005 and is expected to end in 2008. The project began with 5 IT developers and at currently includes 6 IT developers and two system architects (metadata systems and dissemination systems). 840 man-hours were used in 2005, 1060 man-hours were used in 2006 and 2050 man-hours are planned for 2007. The participants are building up experience with services and gradually widening the project group to transfer this knowledge.

Current status
The services have been quickly taken into use in dissemination through About the statistics, About the data collections (descriptions of data for researchers) and Statistical metadata portal:

- 5 services for Vardok to provide definitions of variables.
- 4 services for Stabas to provide standard classifications.
- 7 services for Datadok to provide more easily available and searchable file descriptions. These services can also be used for quality assurance.
- 5 services for Metadb to provide event history data descriptions.

Further work on taking these services into use throughout the statistical production cycle will be undertaken in 2007 and 2008.

Technical solution

- PL/SQL (built into Oracle database) + TOAD (compiling, testing etc.)
- Altova XMLSpy (XML-schema og XML-editing)
- Programing language C# and Visual Studio to make web services -> WSDL (Web Service Description Language)
- Altova Semantic Works -> RDF

Statistical metadata portal

Purpose

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.

Project organisation

	2005 (used)	2006 (used)	2007 (planned)	Total
Statistical adviser in standards	200	300	300	800
System architect for metadata systems	200	200	200	600
IT developer	150	1310	800	2260
Web designer	-	390	300	690
Total	550	2200	1600	4350

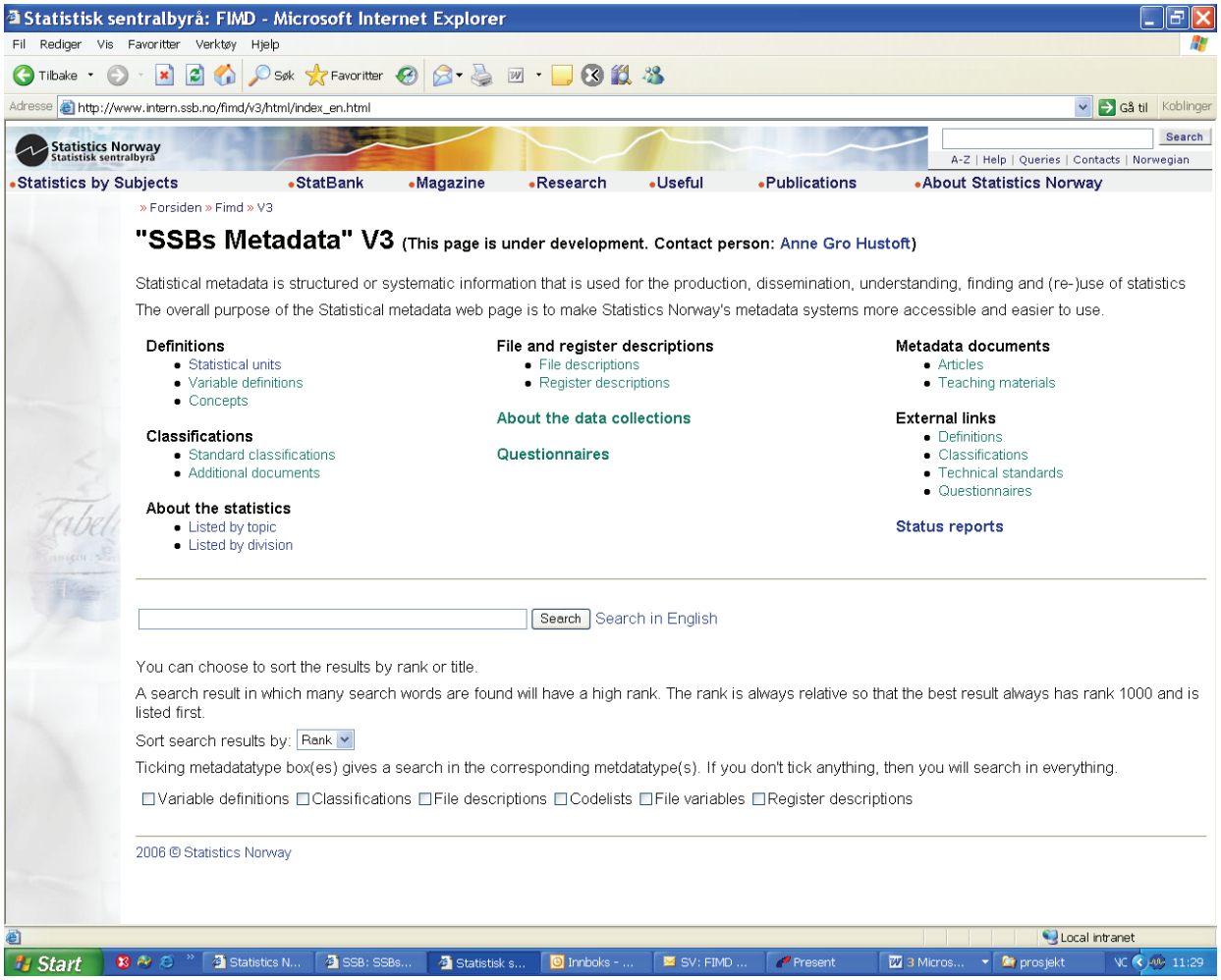
Method

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/statisticalmethods).

For more information on the statistical metadata portal and underlying metadata systems see reference [8].

Current status

The screenshot below shows the current situation for our statistical metadata portal.



SSB employees have access to the Statistical metadata portal via the intranet. User testing of functionality in the opening page shown above and for definitions of variables has been carried out. Further work is required on displaying file descriptions and on search functionality. We expect to have a fully functioning intranet version this year (2007).

Technical solution

- Routines for extracting data from master systems (using services from the service library for metadata systems)
- Store extracted metadata as single files in XML format (eXtensible Markup Language)
- Routines for making different presentations (XML ➔ HTML) (Hyper Text Markup Language)
- Indexing of XML filer for use in search routines.
- Technology solutions are close to those used in ssb.no ➔ CAN include (parts of) the statistical metadata web page in ssb.no.
- Wish to make pages in Internet also searchable from Google, Yahoo etc.

Conclusion

We have made progress in several areas covered by our metadata strategy particularly those nearest our external users. We are still just beginning to see how we can use and re-use our metadata throughout the entire statistical production process. An important step in this direction will be the current redesign of our questionnaire database. Our new IT strategy will support us in the process.

References

[1] Metadata strategy in Statistics Norway, *Hans Viggo Sæbø,* Eurostat Metadata Working Group Luxembourg, 6-7 June 2005.

[2] Hugh Byer and Karen Holtzblatt, Contextual Design - Defining Customer-Centred Systems, Morgan Kaufmann Publishers, 1998.

[3] Variables documentation system in Statistics Norway, *Anne Gro Hustoft and Jenny Linnerud* ,Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, 9 - 11 February 2004.

[4] ISO/IEC 11179: 2005 Metadata registries (All parts). Geneva: International Organization for Standardization and International Electrotechnical Commission.

[5] Neuchâtel Terminology Model, Classification database object types and their attributes, Version 2.1 http://www.unece.org/stats/cmf/PartB.html

[6] Neuchâtel Terminology Model, PART II: Variables and related concepts object types and their attributes, Version 1.0 http://www.unece.org/stats/cmf/PartB.html

[7] New IT Strategy for Statistics Norway*, Rune Gløersen,* UNECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS 2007),Geneva, 8-10 May 2007.

[8] Statistical Metadata in Statistics Norway, *Anne Gro Hustoft and Jenny Linnerud*, Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, 3 - 5 April 2006.