

**UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Work Session on Statistical Metadata (METIS)
(Geneva, Switzerland, 10-12 March 2010)

METADATA PROJECTS AND TASKS AT STATISTICS FINLAND

Saija Ylönen, Statistics Finland

I. ABSTRACT

1. The paper presents the Metis-wiki case study of Statistics Finland. It discusses recent developments of metadata systems and future plans for building and implementing a centralised, common metadata system to serve the whole statistical process.
2. The themes of the paper are the ongoing construction of a variable editor and renewal of the system of classifications.
3. The paper will also present the general goals and the organisation of metadata work at Statistics Finland.

II. ORGANISATION OF METADATA WORK AT STATISTICS FINLAND

4. Statistics Finland's activity and finances are steered by the Director General. The activity is organised into six statistics departments, three support function departments, and the Secretariat of the Director General. (The overall organisation chart of Statistics Finland can be seen at: http://tilastokeskus.fi/org/tilastokeskus/organisaatio_en.html).
5. The main guidelines for the development of the metadata systems at Statistics Finland are co-ordinated and processed in co-operation between the IT Management, Dissemination Services, and Classification and Metadata Services units. The IT Management unit is situated in the Secretariat of the Director General and the Dissemination Services, and Classification and Metadata Services in the IT and Statistical Methods department.
6. The Classification and Metadata Services unit maintains classification standards, concepts and the archiving metadata systems involving both methodological matters and applications management. The statistics departments maintain their own classifications and concepts in the centralised metadata system according to the instructions of the Classification and Metadata Services unit. The unit also trains

and consults the statistics departments in metadata matters and is in charge of quality control of the metadata.

7. The role of the IT Management unit is to co-ordinate matters related to the general information architecture, of which metadata tasks form one element. The Dissemination Services unit has an active role in developing the metadata connected with the dissemination of statistical data.

8. For several years now there has been an unofficial interest group known as the Metadata Co-ordination Group which has been working on metadata issues. In 2009 the group, comprised of members working on metadata and permanent members from all statistics departments, was given an official status. The group provides an important forum for presenting and discussing developments related to metadata systems. The goal of the group is to widen knowledge about metadata and metadata systems and inform its members about achievements in metadata work in-house, nationally and internationally. A further important objective is intensification of co-operation between the statistics departments and the parties responsible for general metadata work.

9. The distribution of tasks related to metadata has for a long time been fragmented at Statistics Finland and because of this the areas of responsibility regarding different issues have not always been quite clear. The organisation of metadata management at Statistics Finland is at present being restructured. In the near future there will be organisational changes which will also affect the Classification and Metadata Services unit. These changes will hopefully clarify the division of responsibilities and tasks between the different units working on metadata within the statistical office.

III. ICT STRATEGY AND THE COSSI MODEL

10. ICT strategy steers the development of a centralised statistical metadata system: "Statistics Finland intends to develop and implement an xml-based common statistical metadata system in order to rationalise and support the harmonisation of statistical processes. The system will be based on Statistics Finland's CoSSI metadata model."

11. The ICT strategy and its future goals lay emphasis on creating common and integrated application tools for metadata systems. In accordance with the main principle, metadata will be created and maintained in a metadata system and made available and transformable to whatever use they are needed in the statistical process from data collection to data dissemination.

CoSSI data and metadata model

12. The development of metadata systems will be carried out at Statistics Finland by applying the CoSSI data and metadata model developed in-house. CoSSI is a modular, xml-based model for describing statistical tables, classifications, concepts, variables, general information on statistical documents, and quality, etc.

13. CoSSI has been designed in accordance with international standards such as the Dublin Core and CALS. If needed, CoSSI can be expanded; new elements for e.g. data descriptions have already been integrated into it. In its ITC strategy, Statistics Finland has provided guidelines for the use of the CoSSI model. The data models of the classifications and concepts in use have been developed in the 1990s, and the elements they contain are presently part of CoSSI.

14. The CoSSI model steering group is in charge of mastering and developing the model according to user needs in a manner that will not expose its main structure to risk. The challenge in the near future will be the creation of a module for process metadata and using the modules of document, statistical, classification and table metadata in new metadata tools.

IV. DESIGN AND CONSTRUCTION OF TOOLS FOR THE PROCESSING AND EXPLOITATION OF METADATA

Present situation

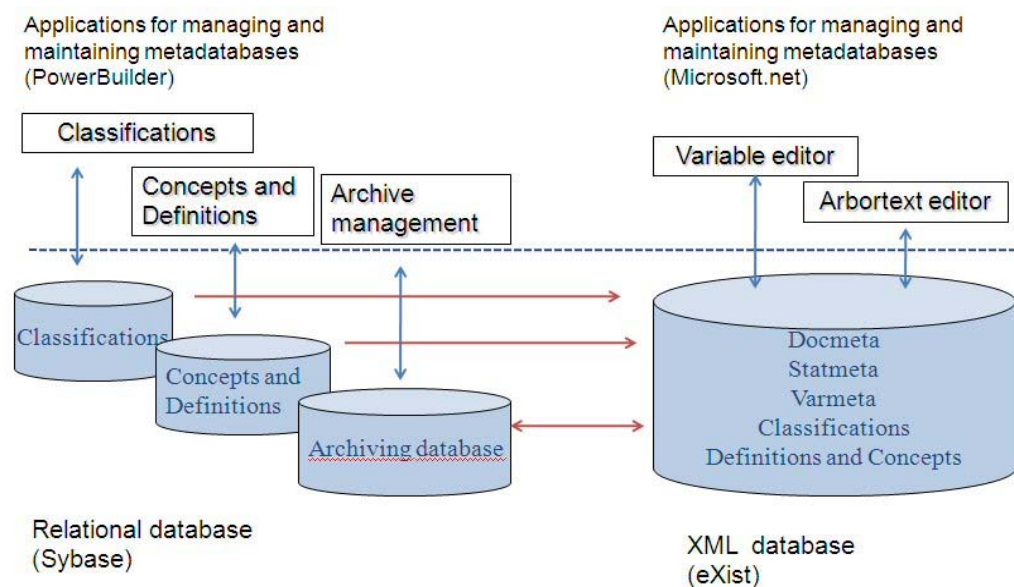
15. At the moment, we have appropriate repositories for storing metadata (classifications, concepts, data descriptions, variables), but we do not yet have all the tools we need for creating new metadata and updating old, and the old tools we have need to be developed further.

16. We also need to think of how we could increasingly benefit from a centralised metadata system so that we could utilise it in all phases of the statistical process. We already have many elements needed for a centralised metadata system but the integration of the system into the statistical process is not sufficiently advanced. For example, the data collecting phase does not yet make use of a centralised system.

17. Our system is in a transitional phase from relational databases to an xml-based environment. The relational databases will be in use for many years to come but at the same time we will be developing the new xml-based system.

18. The metadata repository system consists of an older generation of relational databases and a new eXist xml database. According to Statistics Finland's ICT strategy, in future the xml database will be the primary repository for metadata. However, at the moment the relational databases still feature strongly in the metadata architecture.

Figure 1: Metadata systems at Statistics Finland



19. The relational databases include a classification database, a concepts database and an archiving database which is used for archiving statistical data. The user interfaces for these databases are built with PowerBuilder.

20. The relational databases are used in statistics production but not in all statistical processes or all statistics. The classifications in the relational databases can be used in SAS and Superstar. An archiving database is part of the archiving process. Classifications and concepts are generated from the relational databases to the web pages, but the process is not automated in its entirety.

21. At the moment, the xml database is used mostly in the creation of publications with an Arbortext word processor. Classifications and concepts are copied to the xml database from the relational databases, but the tools for utilising metadata units from the xml database are only just being constructed.

The first metadata tool linked to the xml database is the variable editor which is designed for creating and maintaining the descriptions of statistical data and variables. It is at the testing stage now and its implementation will start in the year 2010. See chapter 4.3.

22. Statistical data are stored in Sybase and MS SQL databases and in an eXist xml database. Programming is made with .NET, PowerBuilder and SAS. The main tabulating tools are SAS and SuperSTAR.

23. It is also worth pointing out that we are making experiments with a MetaAPI interface on an xml database eXist. The purpose of MetaAPI is to ensure that the statistical processing systems and applications could use metadatabases through an application programming interface in which case they would not be dependent on a particular metadatabase. This would make it possible for applications to use metadata in a flexible way regardless of the underlying data models and systems. MetaAPI has also been tested in connection with relational databases but its introduction has not been finalised. The intention now is to re-activate work on this.

Non-centralised systems

24. The above systems are the centralised, common metadata tools at Statistics Finland. In addition to them, there are separate metadata systems for population statistics, income distribution statistics and national accounts. The objective is for these systems to be compatible with the common metadata systems.

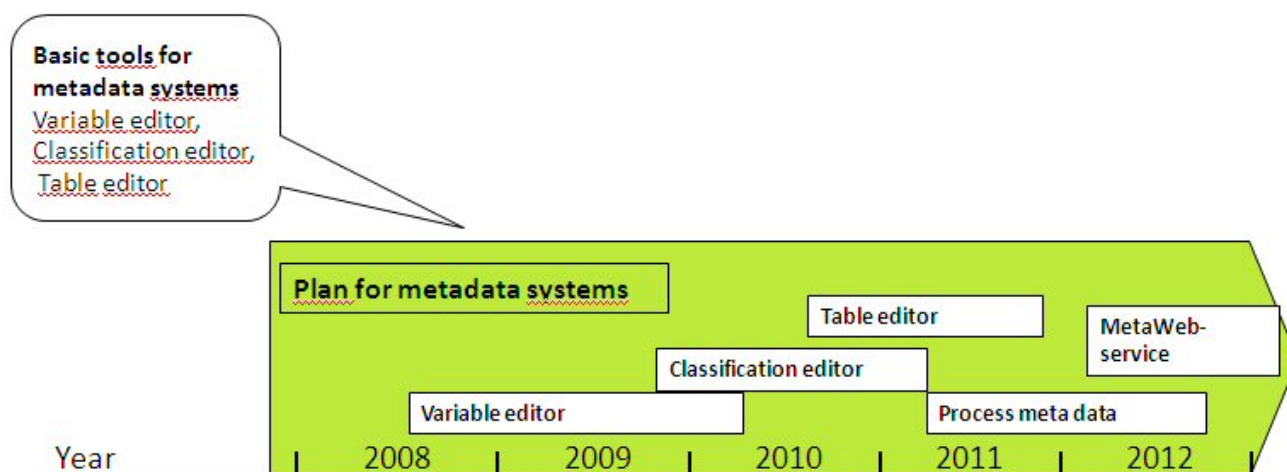
25. Why do some statistics departments still have their own metadata systems? Centralised metadata work progresses too slowly from the perspective of individual statistics. The statistics departments make urgent demands on metadata work, which cannot always be fulfilled in the desired timetable. Quick solutions are requested but the construction of centralised systems must take into consideration the needs of many statistical processes. A way out of this would be to draw up a highly detailed description of the statistical production process. To what extent could the processes of different sets of statistics be harmonised so that they would be able to make efficient use of centralised metadata systems?

26. However, the prevailing attitude still regards the process of an individual set of statistics as unique, and therefore incapable of exploiting systems that are meant for all statistics. A centralised metadata system should support the harmonisation of the production of statistics to a sufficient degree, thus making it more effective, but it should also be flexible enough to a certain extent to serve what statistics specifically call for.

27. Commitment by the Management and their support to the work is crucial for the statistics departments to be able to provide the contribution needed to the development work and for ensuring that the work will be sustained.

Long-term plans

Figure 2. Schedule for building the xml-based metadata system



Projects for the Year 2010

• Testing and Implementing of the Variable Editor

28. The variable editor is a tool for creating and maintaining the descriptions of statistical data and variables. The editor can be used in describing statistical variables from the designing of basic, unit-level data to the tabulation of variables. The data descriptions are saved as xml documents conforming to the CoSSI model in the eXist/xml database. At the moment, the editor is in the testing phase and its implementation is due to start during 2010.

29. The variable editor project which has now concluded covered the construction of the variable editor, compilation of instructions for its use and design of a training package for it. In addition, the project put final touches to the CoSSI metadata model and its documentation and drew up general guidelines for the standardisation of statistical data descriptions at Statistics Finland.

Content and functions of the editor

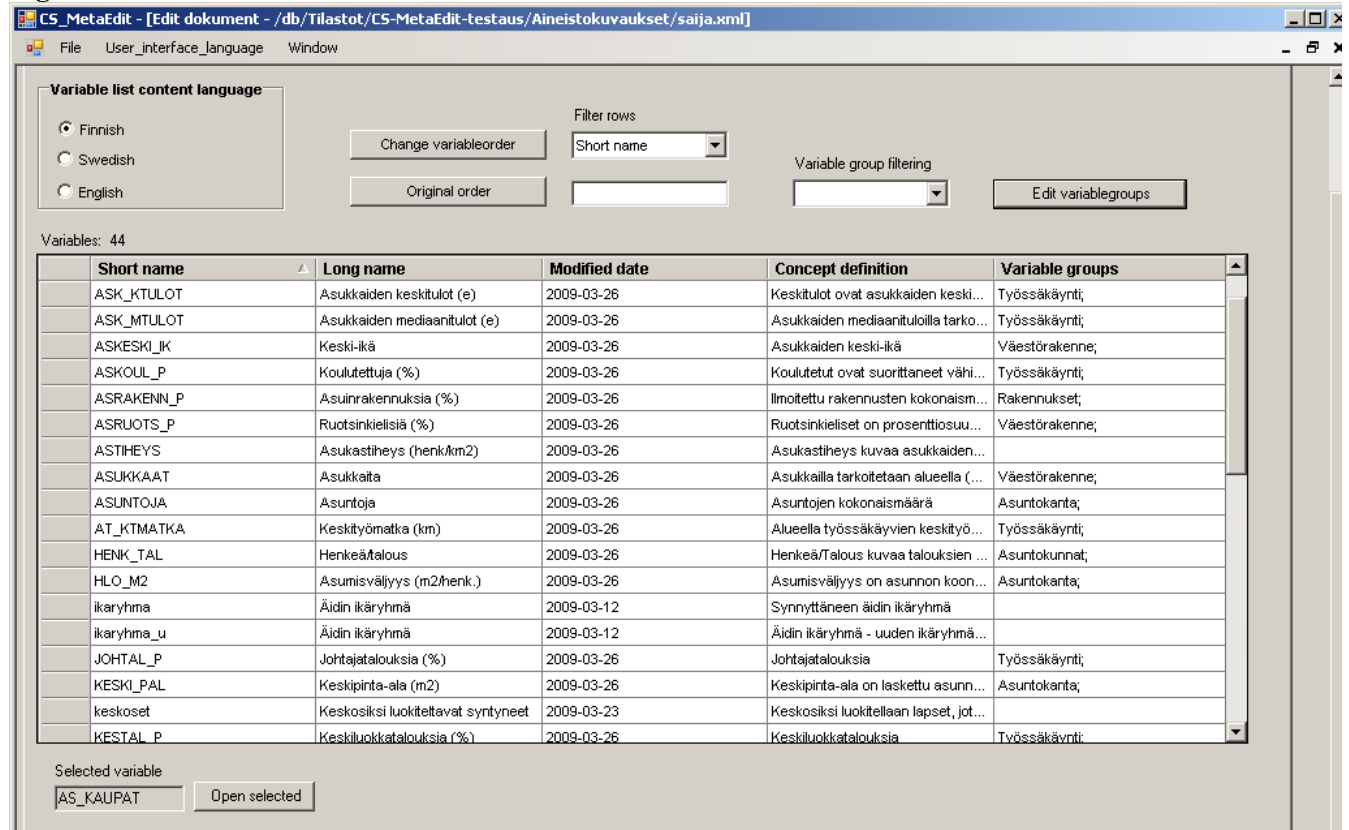
30. The user interface of the variable editor is trilingual. The data descriptions produced with it may be single-language or contain the description in several languages (Finnish, Swedish and English). The data descriptions are comprised of a general description of the data, a list of variables and information about an individual variable.

31. The **Data description** includes descriptive information on the entire data document, such as name and topic of the data, the statistics the data description relates to, information on the creator and content description.

32. The **Variable list** interleaf allows management of the list of variables in the data description and selection of the variable whose description needs editing. Individual variables can be added to the list of variables manually or a number of variables imported to it from a transfer file. Thus, the machine-language metadata need not be entered into the editor manually but data can be imported into it from a fixed-format transfer file (CSV file) into which they have been entered manually or generated automatically from some metadatabase in machine-language. This function makes it easy to import data descriptions in e.g. Word format from the own files of individual statistics into the centralised metadatabase where they can be maintained in an appropriate and safe manner.

33. The user has editing rights only to the data descriptions of his/her own specific set of statistics, but reading rights to all descriptions.

Figure 3. Variable list interleaf in Variable editor



34. The fields available for an individual variable are listed in the following table:

Field name	Description
short name	Short identifying name of variable
long name	Name of variable in natural language
concept definition	Basic conceptual description of variable
operational definition	Verbal description of the formation of the variable
deduction rule	E.g. programming instructions, mathematical formula, etc.
classification ID	Identifier of classification. Refers to a classification in the classification database.
unit of measure	Measurement unit of variable
variable modified	Date of creation or modification of variable (yyyy-mm-dd)
start of validity	Start date of validity of variable (yyyy-mm-dd)
end of validity	End date of validity of variable (yyyy-mm-dd)
status	Stage of editing of variable: draft, ready, validated
variable group	Name of group to which variable belongs. Makes working with long variable lists easier.
work comment	Free text field. Contains information only for the use of the maintainer of a description.

35. Links to the classifications relating to the variable descriptions can be built from the eXist metadatabase into the variable editor. The classifications are still maintained in a relational database but become copied into the xml environment.

Experiences gained during the project

36. Various questions concerning standardisation had to be addressed in connection with the designing of the variable editor although they were not originally expected to fall within the projects' scope of tasks. For instance, a list of units of measurement of research variables complying with the SI system was compiled into the editor. The user does not write the unit of measurement of a variable him/herself but selects the required unit from a drop-down menu, whereby the symbols used for the units become standardised.

37. Because the variable editor project was the first leg in the revision of the metadata system it was subjected to a diversity of expectations. Indeed, adherences to the actual original task and prevention of digression from it typically challenge this kind of a project. Discussions at project meetings tended to drift too far from the core subject which left the project leader with the somewhat unpleasant task of bringing the lively talk back on track. For instance, electronic archiving, lack of guidelines for it and for the division of responsibilities in the organisation caused constant concern and came up frequently in discussions.

38. The data content of the variable editor conforms to the CoSSI model. The data content of the model proved to be exhaustive so that no significant data elements needed to be added for the variable editor.

39. Widespread use of the variable editor and the disciplined way of describing data it requires is beginning to harmonise the structures of databases at all stages of production right from the naming of variables. The variable editor project has created:

- Preconditions for the development of a consistent information architecture which facilitates the harmonisation of applications not only at the releasing and dissemination stage but at all stages of the production process starting right from the processing of basic, unit level data
- Preconditions for the construction of production applications in which metadata need not be separately produced or manually added to data when statistics are published and data are archived
- Preconditions for increasingly efficient information service where excessive time need not be spent on searching for metadata, or on actual reproduction of metadata for special compilation assignments
- A system from which table column and row headings can in future tabulation applications be retrieved in multiple languages for all statistics using the same methods.

• **The planning and building of a classification editor**

40. Year 2010 will be the starting point for a renewal of Statistics Finland's classification tools. The renewal will begin with a definition stage which ascertains the demands the statistical process imposes on classification tools. The construction of a classification editor will be started at the second stage of the project.

41. Our present classification database was constructed in the 1990s. Some changes that were regarded as necessary have been made to it in the course of time but the entire classification system now needs reviewing. The goal of the classification editor project is to develop a new tool for improving the maintenance of classifications in the centralised metadatabase. The project's aim is to gather together the demands imposed on the classification system by the statistical production process and to build a tool to meet these demands.

42. According to its ICT strategy, in the next few years the agency will introduce a common statistical metadata system based on the CoSSI model, which will serve all statistics at all stages of production. Because in principle the new metadata system requires that even all data-specific classifications must be entered into a common metadata store, the maintenance of classifications in the

metadata store must be made flexible and appealing from the perspective of all statistics. This will create the preconditions for a wide introduction of a standardised metadata system even for classification data as required by the ICT strategy.

43. At the moment the original versions of classifications are maintained with the maintenance tool of the old classification database in an SQL environment from which the classifications are copied automatically into the CoSSI metadata system. In order to reduce the diversity of tools and the users' skills requirements the maintenance of all metadata should be centralised into one application environment. The present way of maintaining classifications in the old classification database has been viewed as inflexible by individual statistics. Renewing is also necessary because the technological platform of the old classification database (Sybase/PowerBuilder) will be phased out over the next few years.

44. The classification editor project will:

- (a) Analyse the needs the statistical production process, application development and information service impose on a centralised classification system; design the classification system architecture
- (b) Basing on this analysis determine the functions of the classification editor which will be implemented during the project
- (c) Extend the variable editor application developed in 2009 with a classification editor whose functions include at least the maintenance of the identifiers, headings and definitions of classification categories in an easy-to-use interface and in which the extension needs identified in the analysis of needs relating to a centralised classification system are taken into account.

45. The project will:

- Analyse the service needs required from a centralised classification system: what services should be included in the centralised classification system
- Create maintenance tools for classifications in connection with the CoSSI/eXist metadata store so that the basic maintenance needs of classifications of individual statistics are met in a user-oriented manner which also allows further development of the classification system
- Produce the solutions with which the interoperability of the Sybase classification database and the eXist metadatabase can be ensured
- Compile user instructions for the editor
- Pilot test the editor.

46. The introduction of the editor will be implemented as line work at the Classification and Metadata Services unit.

47. The analysis of the needs imposed on a centralised classification system which the project will make will serve as the foundation for the construction of the system. A classification system which serves well and is intended for the whole agency's use will encourage centralised and structured maintenance of classifications, which will bring down the number of separate classification files in individual statistics and make classifications accessible to everybody. The documentation of classifications will improve, making them easy to find for use in-house and for the provision of information service.

48. The tool for the maintenance of classifications which the project will build will support smooth movement between data descriptions, variable descriptions and maintenance of classifications and thus improve the efficiency of the maintenance and use of classifications in statistics.

49. A centralised classification system eases the workload needed to maintain classifications because classifications are only saved in one place from where they can be retrieved for use at different stages of the statistical production or dissemination processes. It reduces the possibility of errors because classifications are documented in the system consistently so that they are accessible to everybody and easy to find. A centralised repository where the classifications are appropriately documented also

improves the efficiency of time use because working hours need not be spent on looking for classifications and trying to find their background information. At best, a common classification system makes the classifications used in different statistics visible to everybody and thus creates possibilities for their harmonisation. It also improves the efficiency of the management of language versions of the classifications and facilitates automatic attachment of the translations into tables at the publishing and dissemination stages.

- **Other projects**

50. A project has been set up in connection with the renewal of the Business Register for the designing and implementation of a system for the reception of administrative data. The system exploits metadata in the receiving and processing of the data. The system will be designed so that it can also be used in collections of data from other respondents than business enterprises. The project will finish at the end of 2010.

V. CONCLUSION: MATTERS STILL TO BE ADDRESSED

51. The variable editor project emphasised the importance of identification of the sub-processes shared by several statistics, and standardisation of the definitions, concepts and terminology in them. Projects aiming to develop common applications for several statistics should ascertain whether process descriptions have been taken as far as the activity level so that they can be used for analysing common activities of statistics and whether the terminology used of the activities is well-established. If this is not the case, the analysing and describing of processes and matters relating to terminological standardisation must be addressed by the project. This is work that requires analysing, digesting and discussing by several parties, and must be given enough time, for which an IT application project is not the best place.

52. When development is not only aimed at technological change but at a new procedure, standardisation and centralisation of metadata systems, it is beneficial to the project if the line organisation already has appointed the party that will be responsible for the created new activities after the project and that this party has the capability and resources for making a significant contribution already at the project phase. This way it can form an in-depth perception of the new system, its purpose of use and significance as part the metadata system so that correct information is conveyed in the right manner to the users at the introduction stage. Active participation at the development phase also nurtures commitment to the use of the system.

53. The introduction of the metadata system will be a transition project taking place over several years, the successful completion of which requires improvement of the organisation of responsibilities relating to metadata and strengthening of the resources of the organ responsible for metadata matters. Metadata work needs a process owner or programme leader who would take care of the planning of resources so that projects would not be running late because e.g. a programmer is engaged in another project at a crucial moment. From the point of introduction of the metadata system it is also crucially important that the application planners responsible for statistical systems are involved and committed to it so that the new common procedures are not bypassed when statistical systems are being revised.