WP.21
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, 9-11 February 2004)

Topic (i): Functions of metadata in statistical production

RECENT DEVELOPMENT OF SORS METADATA REPOSITORIES FOR FASTER AND MORE TRANSPARENT PRODUCTION PROCESS

Contributed Paper

Submitted by Statistical Office of the Republic of Slovenia[1]

I. INTRODUCTION

1. The paper describes SORS's experience in using metadata within the statistical production process. The recent attention was focused on the questionnaires and variables. Some outputs were developed that benefit from centralized repositories and will be briefly examined. The SWOT analysis, conducted in October 2003, revealed major obstacles in the METIS module with variables. Problems were further analysed within the E-CoRE project (Electronic Collection of Raw data from Enterprises).

2. The need for further development of the classification server (KLASJE) for the SORS production purposes (and within e-government initiative) and better transparency (notifications revealing changes in classifications) is emphasized. Within e-government new functionalities will be developed to assist searching within classifications and concordances. Another result of E-KLASJE is a feasibility study on the possible use of the server within the government institutions authorised to perform statistical surveys (according to the National Statistics Act).

II. RECENT DEVELOPMENT OF SORS METADATA REPOSITORIES FOR FASTER AND MORE TRANSPARENT PRODUCTION PROCESS

3. The main goal of the Statistical Office of the Republic of Slovenia (SORS) in the field of metadata is to develop an efficient and effective, standardized and integrated system for collecting and editing metadata as an important part of the statistical information system. From that system metadata could be quickly and easy exported and used in other applications and programs to support SORS's statistical production.

4. In a broad sense, "production" covers the whole life cycle of a statistical survey or a statistical information system, including design, implementation, operation, monitoring, maintenance and evaluation. Producers of statistical data therefore include designers, input data providers and subject-matter statisticians. All these categories of producers of statistical data have their typical metadata needs[1].

[1] Prepared by Julija Kutin julija.kutin@gov.si, Andreja Arnic andreja.arnic@gov.si

III. BOTH METADATA REPOSITORIES IN THE TARGET DATA FLOW

5. The target data flow is based and enabled with the classification server and metadata repository.

IV. CLASSIFICATION SERVER

6. Within the general metadata concept at SORS a modern and capable classification server called KLASJE was built, which supports the entire life cycle of a classification. Life cycle means: creation, updating, maintenance, creating new versions and deleting the obsolete. Together with METIS repository, KLASJE represents the basic metadata infrastructure at SORS. KLASJE enables various contents and time comparisons between classification versions. Setting up of KLASJE enabled the standardization of processes linked to the use of classifications and other code lists and nomenclatures. With the rules embedded in the tool, the quality of new classifications improved, which directly influenced the quality of statistical data. At the end of November 2003 there were 800 classifications and 75 concordances ("owned" by 16 owner groups) in our classification server. The expected target is 2 900 classifications and concordances. KLASJE has been developed on the basis of the methodological approach of CARS[2] and the in-house development at SORS.

7. The need for further development of the classification server for the SORS production purposes (and within e-government initiative) and better transparency (notifications revealing changes in classifications) was recommended during the recent months. When we started with the production of the classification server, some additional needs for search, navigation and display appeared: notifications, i.e. informing users about changes on the classifications and news about the classification server, concordances via the Internet, the module which enables automatic load of dimensional tables in data warehouse locate items in the index, search in classification using wildcard (further development of search functionality on the Internet), e-government initiative.

8. Within e-government initiative the classifications database is available via the Internet and at the e-government portal (hyperlink). Special access to standard classifications is provided. Additional search facilities and accessing concordances via the Internet is being developed. Within e-government initiative the possibility (of initiation) of common rules and tools will be explored, which would include all institutions authorized for collecting and publishing of statistical data within the frame of the national program of statistical surveys.

9. Notifications are informing users about changes on the classifications and news about the classification server. The application automatically notifies all user groups owning classifications having the same topic as the classification version being released and all user groups which have registered an interest in a classification topic which is the topic of the classification version being released.

10. The application notifies the classification version, the classification title and abbreviation, version attributes such as version id and date of change and the reason for the change. The notification is made on release of a classification version. Error correction changes are allowed pre or post release or even if the version is in use. This is controlled (manual procedure). It is also notified.

V. METIS REPOSITORY

11. The main modules we will talk about are Questionnaires and Methodology (variables). The application for managing metadata has much functionality: initial loading with SQL statements, interactive feeding metadata base, interchanging the metadata between metadata base and CSV files, preparation and export metadata for special use. Sharing information and metadata requires standard solutions both for the technology and the content.

V.1 In production since 2003

12. We have already defined the standardized statistical process and we will use standard tools for each subprocess. If we take a look at the dissemination process, we can say that it is centralized (standard processes). We should have standard tools to disseminate data – central information infrastructure, central dissemination server, dissemination tables and metadata in structured form (dimensional matrix in PC-Axis form). Dissemination of statistical tables and data as well as preparation of publications (Web publishing) will be done with PC-Axis family. The objectives of modernization are integration of contents and dissemination processes, preparation of standard procedures and shortening the preparation of dissemination services for external users, and improvement of quality in individual categories. Internet should be the main dissemination tool, even though other ways of dissemination are not and will not be neglected.

13. Metadata play a major role in dissemination of statistics, including helping users to find, understand and assess statistics in the context of their specific objectives. As standard tools and approaches for creating and managing, the metadata used for dissemination can also be used for survey design and statistical production activities.

- Making tables with query tool (Discoverer, PC-make, PX web) and preparing tables for dissemination. One of the most important needs was to bring data more closely to the statistical users and enable them to make their queries by themselves. We applied user friendly query tools and dimensional modelling technique: Ad. Hoc tables – Oracle discoverer- the result is XLS table; periodical tables: PC-Axis SQL – the result is PX table, PL/SQL – the result is ACSII table, Tabulation with hiding individual data - Tau Argus with interface for exporting micro data from database and metadata from METIS – the result is XLS table.
- Prepare PC-Axis files with PX-Make
- Publish tables on the Internet - Put PC-Axis files on the Internet with PX-Web
- Data shooting (user interface is being prepared)

14. The other functions of metadata in statistical process: the dissemination process will be made out of a series of mutually linked procedures such as: creating/updating the calendar of realization and other metadata - Advance release calendar as in Norway, creating/updating px/info files, creating/updating publications, copying px/info files, sending publications to the Editorial Board for review, editing publications, returning publications from review, creating html and pdf files, publishing, archiving. Then we have quality declarations with PX-files. By making a quality declaration, the producer can specify the properties of a product so that it can be used in a proper way and inform users about what quality in different aspects they can count on. Since user's opinions and preferences change over time, the producer must continuously strive to adapt the product to new needs and expectations. A user oriented quality concept facilitates communication between users and producers. It is always important that production processes are as efficient as possible. A producer therefore wants to be able to judge the relative benefits and costs for different parts of the production process. The quality concept provides one of the instruments for efficiency evaluation and an optimal allocation of production resources. Then we are working on archives of statistical publication as in Sweden (pdf and html), achieves of statistical publication, searching html files, consulting thesaurus, GESMES/CB message converted from PX-file, statistical Database (faster access).

Re – engineering needed in 2004

15. Metadata in METIS are connected among themselves with strict identifications and common classifications. Modules can be added, developed and updated separately. Metadata can be imported interactively or by procedures from CSV templates prepared for this purpose. We have to start thinking about feeding metadata base through the Internet.

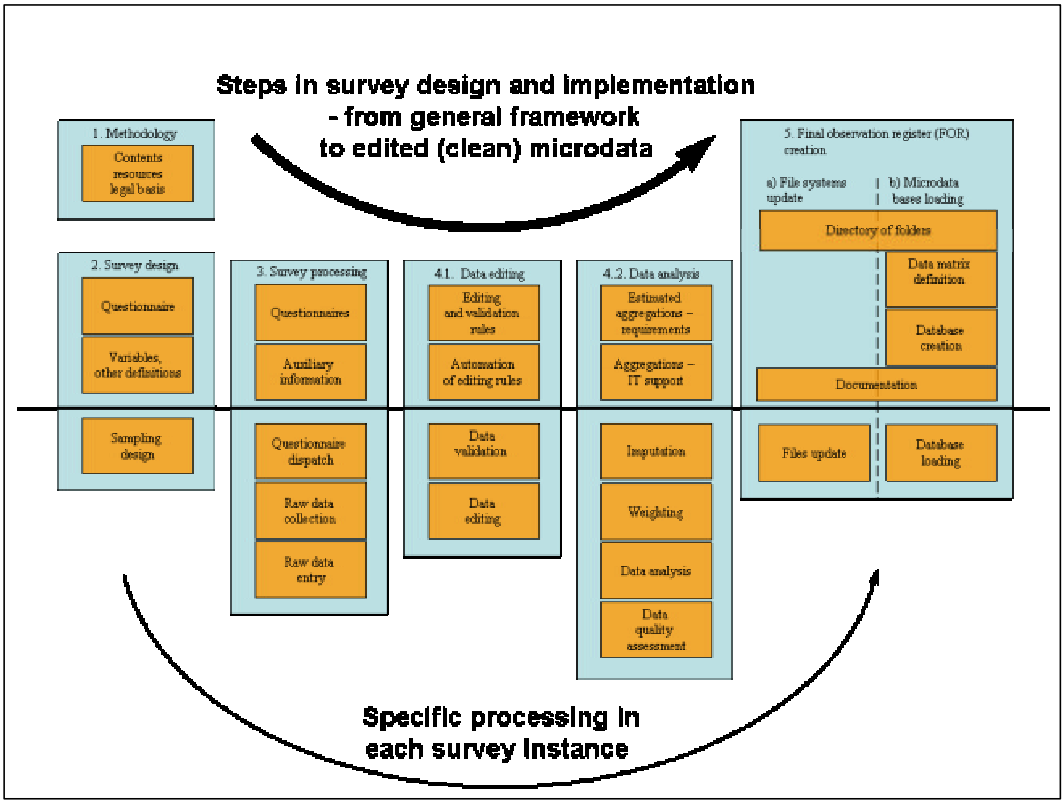
16. The main module of METIS consists of variables, their values and questions. It has to be managed by a methodologist responsible for statistical resource. The module is composed of two parts: Statistical variables with values and Questionnaires with questions. They were developed separately and they can work independently of each other. From the part of statistical variables we can prepare the catalogue of statistical variables through statistical resources and other contents analysis and reports. Because of the uses of the statistical process there are many attributes of the variables needed for modelling paper questionnaires, questionnaires in electronic form, some data for generating form for fast entering data and some data for logical

control of the data. The metadata could be more transparent and then easier and faster analyses of statistical variables could be possible.

17. The questionnaire part was developed within the standardization of modelling questionnaires and does not cover the whole subprocess of collecting data. The philosophy and rules built in the model are not in accordance with those in comparable modules of other statistical offices (IQML[3]). We have a lot of metadata about sections, tables on questionnaire and values they do not have. A lot of them are obligatory, so we have to analyze them again and decide which are really needed for the process in all tools for preparing questionnaires. We should eliminate those that are important for only one tool from METIS and transport them into that tool's meta database from where the tool could export them. On the other hand, there are a lot of metadata we do not have and the others (IQML) have. There are sub-questions and metadata about them, element groups, control element components (calculations), validation (handler), rule (navigation), formula. We know that the same problems others are exploring, namely Austria, Norway, USA, the Netherlands. Because of the in-house use of Blaise, maybe the most important for study is the Nederlands. It would be very good to carry over the knowledge in the some way as we did with KLASJE.

18. We think we could make a revision of this module, so now we started with detailed analysis of data collection subprocess. We analyzed 7 typical SORS's questionnaires from methodology definition to the FOR – micro database, or final observation register (Picture 1.). Some possibilities were offered, which should serve as a starting point for further reflections: electronic form as option next to paper form, reducing periodicity (times a year), reduce frequency (number of companies), improve/introduce explanations/instructions to fill in the form (manuals), simplifications of a form (redesign form and make them more user-friendly, skipping unnecessary questions), integrating forms (by redesigning G2G data flows, skipping overlap (in case of similar questions/answers)), adapting forms to source administrations / common practices of enterprises (data requested similar to data as administrated in enterprises).

Picture 1. Steps in survey design and implementation from general framework to edited (clean) microdata



19. In the first of the instances of a statistical survey, it is mandatory to go through all steps of the process. But in the following instances of the same survey there are only steps under the line mandatory. At the end of this part of the process data have to be updated in the files and stored in the directory system and documented with FOR. From there they can be used for analyses in different tools (Excel, SAS, Access and COBOL) or imported into Oracle database.

20. Then we found out a lot of problems of layout of questionnaires and general data on the questionnaire like different survey codes in use, no public mapping of different codes, lack of documentation available, lack of transparent archiving procedures, poor (or none) naming conventions for surveys; difficult or impossible to map survey data and respondents list for the same reference period without personal contact, lack of (or none) layout standardisation.

21. Statistical survey data processing begins with survey metadata entry in the metadata base. Each new survey should be registered in the system (see also: Zeila 2003)[4]. For each survey it is necessary to create the survey version, which is valid for at least one year with concrete content and layout. If survey content and/or layout do not change, then the current survey version and its description could be operational for the next year.

22. Each statistical survey contains one or more data entry tables or chapters or sections. For each chapter it is necessary to describe table type, which can be a constant table with fixed rows and columns number or a table with variable rows or columns number. For each survey (version) chapter there has to be description of tables and rows with their codes and names. All survey values from questionnaires should be stored and each value should have a relation to cell which describes value meaning. Cells definition should be a combination from two sides – from description of variable and a coordinate within the survey version.

23. This has to be well documented. It is very important to use the same coding system through all the process. First of all, for a statistician who works with one survey every day and is familiar with the survey content it is the easiest way to work with values using table and row language (table and row codes). Also all validation rules for the survey could be described using table and row language. This is necessary for automatic data entry application generation and for later mapping of data in the raw and clean data bases too.

24. With the new survey version, very often for the same cell table and row coordinates are changing. For data analysis such approach is not useful, because value in one year is identified with one table and row coordinates, but in the next year it is identified with another one.

VI. Findings

25. Important findings and measures need to be taken into account in individual phases of work in order to provide the highest possible level of quality of processes that result in the quality of data and services, namely:

- The statistical process needs to be constantly analyzed and modernized in view of developments and changes in the field of IT.
- Adequate standardized IT infrastructure needs to be provided and producers need to be linked to data sources and users.
- The necessary metadata infrastructure needs to be set up and developed, which enables the integration of data and processes and directly influences the quality.
- Regular training and cooperation in international exchange of knowledge and experience needs to be organized. Effective use of expert knowledge (experience, new knowledge and creativity), coordination of the counterparts, process modelling and implementing of renewed internal standards in technical, organizational and layout sense with the use of unified programming tools are the most important issues related to quality.
- Processes and procedures need to be standardized.
- Deadlines need to be set and monitored.
- Users and producers satisfaction needs to be measured.

26. It is essential to be aware that in-house development and the ability to implement new methods, technologies and knowledge are crucial since only in this way can we provide permanent development and progress for employees and the institution and thus also the quality of processes, services and data.

VII. Sources:

[1] UN, UN/ECE, Guidelines for the modelling of statistical data and metadata, United Nations, Geneva, 1995 http://www.unece.org/stats/publications/metadatamodeling.pdf

[2] Dallas Welch, Classifications and Related Standards System (CARS), Output Database Workshop, Stockholm, September 1997

[3] LAMB J., A SOFTWARE SUITE AND EXTENDED MARK-UP LANGUAGE (XML) STANDARD FOR INTELLIGENT QUESTIONNAIRES, IST-1999-10338, FINAL REPORT, DRAFT 2, http://www.epros.ed.ac.uk/iqml/deliverables/D13/IST-1999-10338-D13v4.doc

4 Zeila K.; Metadata Driven Integrated Statistical Data Processing and Dissemination System, AMRADS Final Conference, Rome, 2003
http://amrads.jrc.it/index.asp, Proceedings 1