

Metadata Driven Integrated Statistical Data Management System

**Author: Karlis Zeila, Vice President
Central Statistical Bureau of Latvia**

The aim of this report is to share with participants the experience gained from the development and implementation process of the new generation statistical data processing system, which integrates several subsystems, and is metadata driven.

The new system, in accordance with CSB IT Strategy, was developed as a centralized system, wherein all data are stored in a corporate data warehouse. The new approach is to unite what logically belongs together by using advanced IT tools, to ensure the rationalizing, standardization and integration of the statistical data production processes. This in practice means movement from a classic stovepipe to a process-oriented data processing approach.

The theoretical basis for system architecture is the result of metadata studies carried out and published by Professor B. Sundgren of Statistics Sweden.

The metadata driven data processing system is based on systemisation and storage in the centralised metadata base - all metadata on surveys, indicators and classifications used, thus ensuring preconditions for unification and harmonisation of the statistical indicators avoiding duplication in different surveys. The introduction of a unified system of statistical indicators leads to expansion of the system's functionality and metadata becomes the key element of the entire system. System integration is based on direct use of the Business Register data in the case of data processing of business statistics surveys. Register modules of the system foresee exploitation of several different statistical registers, depending on the statistics to be produced. The main purpose of this type of integration is to ensure an easy access to a wide range of the respondent's data; these data are located in statistical registers. Most of the system software modules are connected with the Register modules. All system software modules are connected with the Core Metadata module. We can refer to this as an integrated system because all data are stored in the corporate data warehouse.

The system is divided into the following business application software modules, which must cover and support all phases of the statistical data processing:

- Core metadata base module;
- Registers module;
- Data entry and validation module;
- Mass data entry module;
- WEB-based data collection module;
- Missing data imputation module;
- Data aggregation module;
- Data analysis module;
- Data dissemination module;
- User administration module.

The WEB-based data collection module functions basically as an extension of the paper-based questionnaire data collection technology. The layout of the electronic form is almost the same as that of the hard copy version; therefore respondents do not encounter any changes moving from paper to the screen. There are even more advantages in submitting data electronically. Data submitted in previous periods can be more easily reviewed and changes can be made to data submitted in previous periods, etc.

The main conclusions of the development and implementation process of the system are as follows:

- Design of the new information system should be based on the results of a thorough analysis of the statistical processes and data flows.
- Clear objectives of achievements have to be set up, discussed and approved by all parties involved: Statisticians, IT Personnel and Administration.
- As the result of feasibility studies, it was clearly understood that all steps of statistical data processing for different surveys allow standardization, while each survey may require complementary functionality (non standard procedures), which is necessary just for this exact survey data processing.
- For solving problems with the non-standard procedures, interfaces for data export/import to/from the system have been developed to ensure the use of standard statistical data processing software packages and other generalized software available on the market. Within the process of the design and implementation of metadata-driven integrated statistical information system, both statisticians and IT specialists should be involved from the very beginning.
- A clear division of the tasks and responsibilities between statisticians and IT personal is the key point to achieve successful implementation.
- Both parties must have a clear understanding of all statistical processes which will be covered by the system, as well as the meaning and role of metadata within the system from both the production and user sides.
- Initiative to move from a classical stovepipe production approach to a process oriented approach must come from the Statisticians side, not from that of IT Personnel or Administration. Motivation of the Statisticians to move from existing to the new data processing environment is essential.
- Improvement of metadata knowledge is one of the most important tasks throughout the entire process of the design and implementation phases of the project.
- It is necessary to establish and train a special group of Statisticians, who will maintain the metadata database and who will be responsible for its accuracy.
- To achieve the best performance of the entire system it is important that the execution of the statistical processes are organized in the proper sequence,
- For the administration and maintenance of the system it is necessary to have well trained IT staff, familiar with the MS SQL Server 2000 administration, MS Analysis Service, other MS tools, PC AXIS family products and system Data Model, system applications.