METADATA DRIVEN INTEGRATED STATISTICAL DATA PROCESSING AND DISSEMINATION SYSTEM

By Karlis Zeila

Central Statistical Bureau of Latvia

Abstract

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

The new system in accordance with Central Statistical Bureau (CSB) IT Strategy is developed as centralized metadata driven data processing system, where all data and metadata is stored in 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 classic stovepipe to process oriented data processing approach.

System integration is based on direct use of the Business Register data in case of Business statistics surveys data processing. Registers module of the system foreseen exploitation of several different statistical registers. Most of the system software modules are connected with the Registers module. All System software modules are connected with the Core Metadata module.

System is modular and consists of 10 software modules, which are covering and to supporting all phases of the statistical data processing.
The main conclusions gained within the system development and implementation process are listed at the end of this report.

1. INTRODUCTION

Production of high quality statistics is impossible without wide use of newest solutions of information technologies. Therefore CSB of Latvia has set the goal to create a modern, effectively working IT infrastructure which provides convenient, fast and safe collection, processing, storage, analysis and dissemination of statistical data taking into account the requirements of information society.

In development of new data processing system, applications or subsystems, make use of metadata, as well as key management elements, ensuring thus a maximally end-user-friendly working environment where there will be no need to reprogram the processes in cases of recurrent changes.

Centralisation of data collection and entry procedures was the starting point of gradual transition towards process-oriented data processing approach at the CSB of Latvia.

The main business and information technology (IT) improvement objectives, that the CSB intended to achieve was as follows:

- Increase efficiency of the main process at CSB, production of statistical information;
- Increase the quality of the statistical information produced;

- Improve processes of statistical data analysis;
- Modernise and increase the quality of data dissemination;
- Avoid hard code programming via standardisation of procedures and use of metadata within the statistical data processing.

2. TECHNICAL PLATFORMS AND USED STANDARD SOFTWARE

The proposed system is in line with the CSB IT strategy, existing computer and network infrastructure utilized.

The Microsoft SQL Server 2000 handles system databases. All applications comply with the client/server technology model, where data processing performed mostly on server side. Client software applications are developed using Microsoft Access 2000. Other components of Microsoft Office 2000 are used as well. For multidimensional statistical data analysis is used Microsoft OLAP technology, which was tested with positive results, for example, in Statistics Netherlands.

As the tool for data dissemination was chosen product PC-AXIS developed by Statistics Sweden, which is widely used in different statistical organizations in different countries.

To date System is operating on the servers connected in asymmetric cluster consisting of:

- ProLiant DL760 with 4 (Pentium III Xeon 900 MHz) processors and 8 Gb RAM;
- ProLiant 8000 with 2 processors (Pentium III Xeon 550 MHz) and 8 Gb RAM;
- Storage Works RAID Array 4000 with 8 disks of 36 Gb, 10 000 rpm.

Cluster system with workstations is connected within LAN Fast Ethernet 1Gbit base 100Mbit per sec.
Types of workstations used are Pentiums II to IV with required RAM not less than 128 Mb equipped with OS MS Windows 95 (better MS W-2000) and MS Office 2000.
To date with System are working ~200 users in CSB central office and more than 60 remote workstations in regional data collection and processing centres connected on-line.

3. SYSTEM ARCHITECTURE

As the result of analysis of the statistical processes and data flows within the feasibility study period it was found, that most of statistical surveys have the same main steps of data processing starting with survey design and ending with statistical data dissemination. And statistics production workflow in CSB of Latvia could be standardised on the first step for the production of Business statistics as it is shown below in the figure 1.

As the theoretical basis for system architecture was taken invited paper from Meeting on the Management of Statistical Information Technology (Geneva, Switzerland, 15-17 February 1999) “An information systems architecture for national and international statistical organizations” prepared by Mr Bo Sundgren, Statistics Sweeden.

The new system is developed as centralized system, where all data is stored in 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.

Important task during design of the system was to foresee ways and to include necessary interfaces for data export/import to/from already developed standard statistical data processing software packages and other generalized software available on the market, which functionality was irrational to recode and include as the system component.

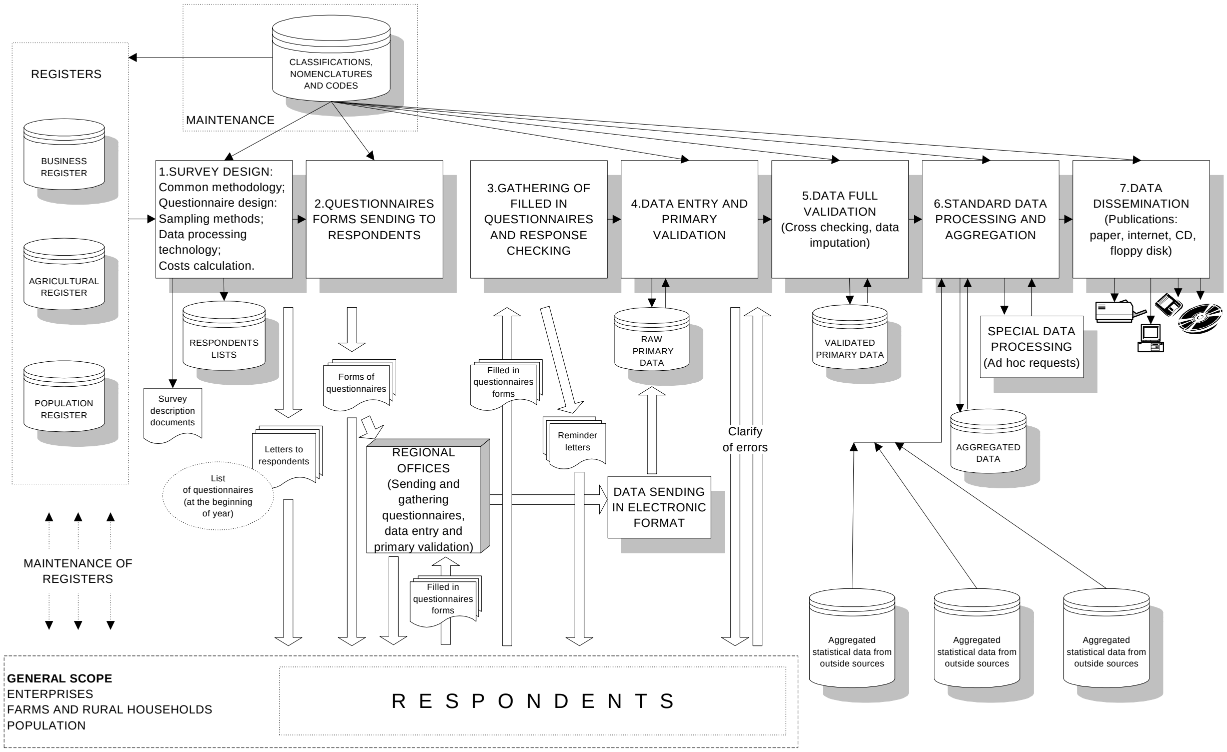

CLASSIFICATIONS, NOMENCLATURES AND CODES
MAINTENANCE
REGISTERS
BUSINESS REGISTER
AGRICULTURAL REGISTER
POPULATION REGISTER
1.SURVEY DESIGN: Common methodology; Questionnaire design: Sampling methods; Data processing technology; Costs calculation.
2.QUESTIONNAIRES FORMS SENDING TO RESPONDENTS
3.GATHERING OF FILLED IN QUESTIONNAIRES AND RESPONSE CHECKING
4.DATA ENTRY AND PRIMARY VALIDATION
5.DATA FULL VALIDATION (Cross checking, data imputation)
6.STANDARD DATA PROCESSING AND AGGREGATION
7.DATA DISSEMINATION (Publications: paper, internet, CD, floppy disk)
RESPONDENTS LISTS
Survey description documents
Forms of questionnaires
Letters to respondents
List of questionnaires (at the beginning of year)
REGIONAL OFFICES (Sending and gathering questionnaires, data entry and primary validation)
Filled in questionnaires forms
Filled in questionnaires forms
Reminder letters
RAW PRIMARY DATA
DATA SENDING IN ELECTRONIC FORMAT
Clarify of errors
VALIDATED PRIMARY DATA
SPECIAL DATA PROCESSING (Ad hoc requests)
AGGREGATED DATA
MAINTENANCE OF REGISTERS
Aggregated statistical data from outside sources
Aggregated statistical data from outside sources
Aggregated statistical data from outside sources
GENERAL SCOPE
ENTERPRISES
FARMS AND RURAL HOUSEHOLDS
POPULATION
RESPONDENTS

Figure 1. Standardised Statistical data flow diagram

System consists of following business application software modules, which have to cover and to support all phases of the statistical data processing:

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

System architecture is represented in figure 2.

4. METADATA STRUCTURE AND KORE METADATA BASE MODULE

The Core metadata base module is one of the main parts of the new system and can be considered as the core of the system. All other modules of the system use Metadata base data handled by this module.

In order to cover all concepts commonly referred to as metadata, one can define statistical metadata as:
“All the information needed for and relevant to collecting, processing, disseminating, accessing, understanding, and using statistical data”.

The data in the metadata base, in essence, is information about micro and macro data i.e. description of the numerical data within the statistical production process and the real world meaning of this numerical data. Also the system metadata base contain description of statistical surveys itself, their content and layout, description of validation, aggregation and reports preparation rules.

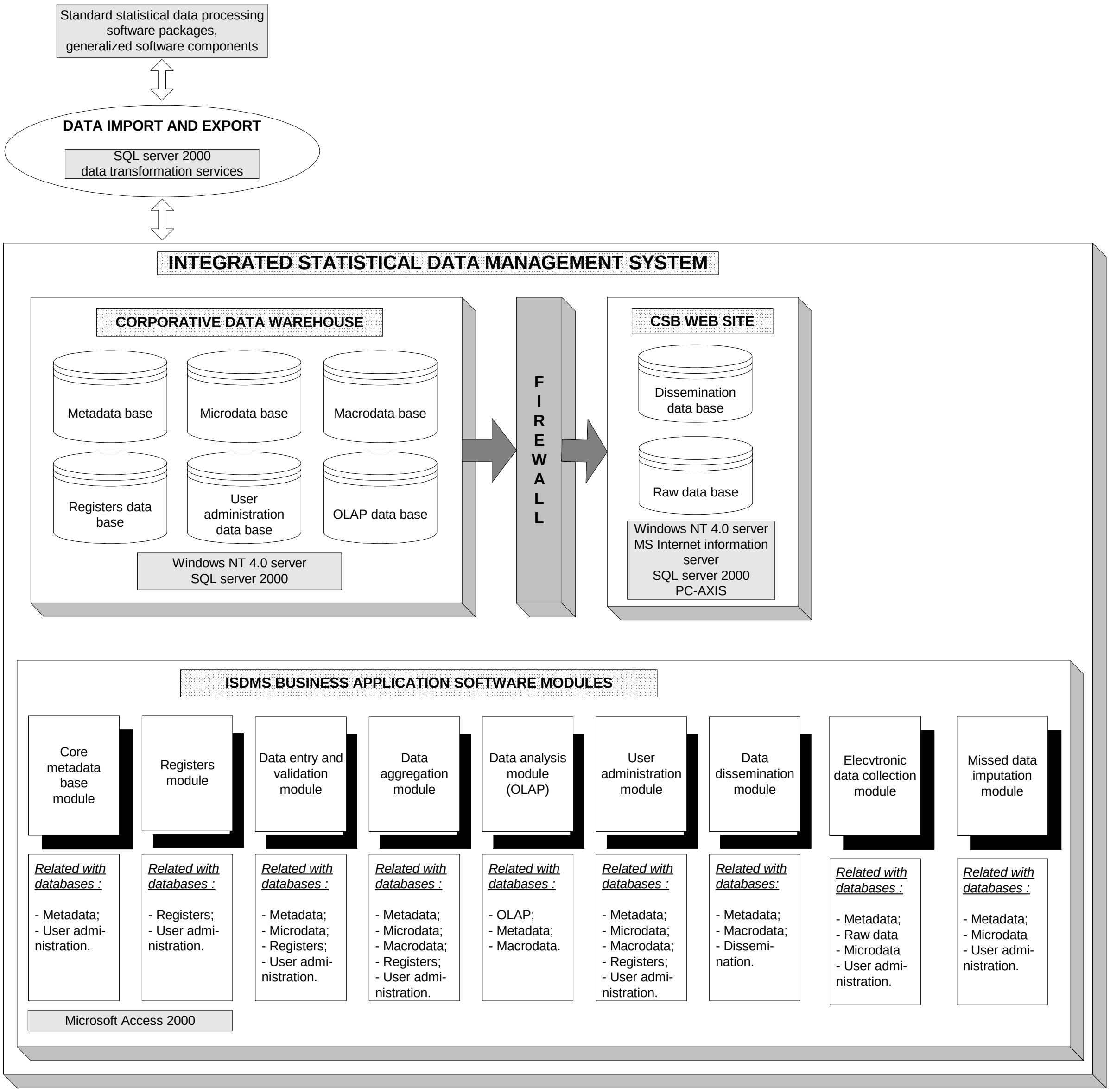


Figure 2. System architecture

The system ensures that Metadata base is used as the key element for the creating universal, common, programming-free approach for different statistical surveys data processing.

4.1. Structure of micro data (observation data) [Bo Sundgren model]

Objects characteristics:

$$C_o = O(t).V(t) \qquad (1)$$

where: **O** - is an object type; **V** - is a variable; **t** - is a time parameter. Every results of observations is a value of variable (data element) – $\mathbf{C_O}$

All variable values have object (respondent) requisites added, which can be called vectors or dimensions. By analysing all the respondents' population, these dimensions are used for creating different groupings and for data aggregation.

In business statistics the following respondents requisites (vectors) for example can be added to each value of variable:

- Main kind of Activities (NACE classification);
- Kind of Ownership and Entrepreneurship (IUFIK classification)
- Regional location (Regional classification - ATVK)
- Employees group classification
- Turnover group classification.

4.2 Structure of macro data (statistics)

Macro data are the result of estimations (aggregations). The estimations are made on the basis of a set of micro data.
Statistical characteristics:

$$\mathbf{C_S = O(t).V(t).f} \qquad (2)$$

where: **O** and **V** - is an object characteristics; **t** - is a time parameter, **f** – is an aggregation function (sum,count,average, etc) summarizing the true values of **V(t)** for the objects in **O(t).**
The structure for macro data is referred in metadata base to as box structure or "**alfa-beta-gamma-tau** " structure.
For data interchange **alfa** refers to the selection property of objects **(O)**, **beta** – summarized values of variables **(V)**, **gamma** – cross classifying variables, **tau** – time parameters **(t)**.

4.3 Structure of Surveys (questionnaires)

New survey should be registered in the System. For each survey a questionnaire version should be created, which is valid for at least one year. If questionnaire content and/or layout do not change, then current version and it description in Metadata base is usable for next year.
Each survey contains one or more data entry tables or chapters (data matrix), which could be constant table - with fixed rows and columns number or table with variable rows or columns number.

For each chapter we have to describe rows and columns with their codes and names in the Metadata base. This information is necessary for automatic data entry application generation, data validation e.t.c.

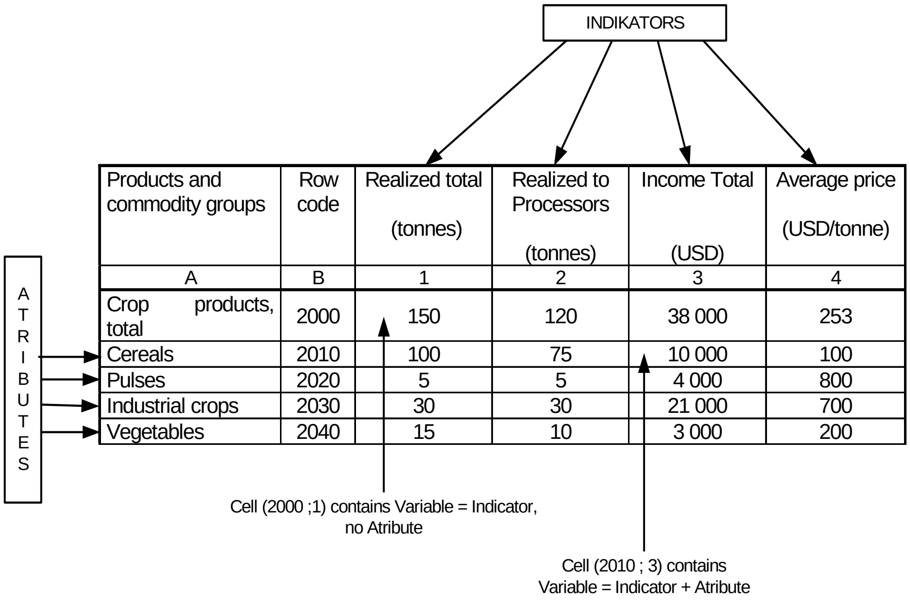
Last step in the questionnaire content and layout description is cells formation. Cells are the smallest data unit in survey data processing. Cells are created as combination of row and column from survey version side and variable from indicators and attributes side.

As an example we could look at Retail Trade Statistics Questionnaire structure from Meta data point of view:

- General information

Name of Questionnaire, index, code(unique), corroboration date, Respondent (object) code, name and address;
Period (year, quarter, month)
Name of chapter

- Data matrix - fixed table



Products and commodity groups	Row code	Realized total (tonnes)	Realized to Processors (tonnes)	Income Total (USD)	Average price (USD/tonne)
A	B	1	2	3	4
Crop products, total	2000	150	120	38 000	253
Cereals	2010	100	75	10 000	100
Pulses	2020	5	5	4 000	800
Industrial crops	2030	30	30	21 000	700
Vegetables	2040	15	10	3 000	200

Cell (2000 ;1) contains Variable = Indicator, no Atribute

Cell (2010 ; 3) contains Variable = Indicator + Atribute

Example where data matrix is with fixed number of rows and variable number of columns. Indicators of the economics activity.

Column heading Row heading	Row code	*Total*	Title1	Title 2	n	Title n-1	Title n
A	B	9999	ISIC 1	ISIC 2		ISIC n-1	ISIC n
Number of employees	1010						
Net turnover	1020						
Other income	1030						

Example of the data matrix with fixed number of columns and variable number of rows. Production of Industry products.

Product title	Product Code (HS or SITC)	Produced in natural measurement	Sailed in natural measurement	Income in USD
A	B	1	2	3
Product 1	1234567			
Product 2	2345678			
………….	…..			
Product n-1	3456789			
Product n	4567890			

4.4 Creating of variables

INDICATOR + ATRIBUTE (Classification) = VARIABLE

ATRIBUTES = dimensions or vectors of INDICATORS
Vectors always are classifications and they could be as follows:

- Kind of activity – NACE
- Ownership and entrepreneurship – IUFIK
- Territory and etc.

Example:

Number of employees + no attribute	= Number of employees total
+ kind of activity (NACE)	= Number of employees in breakdown by kind of activities
+ location (Territory classification)	= Number of employees in breakdown by territories

System users can easy query necessary data form Micro data / Macro data databases navigating via Metadata base. Metadata are widely used for data analysis and dissemination.
Metadata base is linked at database structure model level with Micro database and Macro database (see figure 3).

Statistical survey data processing begins with survey metadata entry in the Metadata base. Each new survey should be registered in the system. For each survey it is necessary to create survey version, which is valid for at least one year with concrete content and layout. If survey content and/or layout does not change, then current survey version and it description in Metadata base is usable for the next year.

Each statistical survey contains one or more data entry tables or chapters. In Metadata base for each chapter it is necessary to describe table type. For each survey version chapter in the Metadata base describes rows and columns with their codes and names. All this information about survey version chapters, rows and columns is necessary for automatic data entry application generation, which layout looks like paper questionnaires.

Information about statistical indicators is saved in the Metadata base. In Metadata base common indicators list is stored. Indicators itself are independent from surveys. This gives a possibility to attach one indicator to several surveys and to get information about one indicator from several surveys as well.

It is possible to define attributes– classifications for each indicator in the system, which gives opportunity to describe and store indicators values in a much-detailed division. Indicators could be without attributes

When indicators and attributes are defined, it is necessary to define variables. Variables are combination of indicators and corresponding attributes. Created variables are connected to survey.

Last step in the survey content and layout description is cells formation. Cells are the smallest data unit in survey data processing. Cells are created as combination of row and column from survey version side and variable from indicators and attributes side.

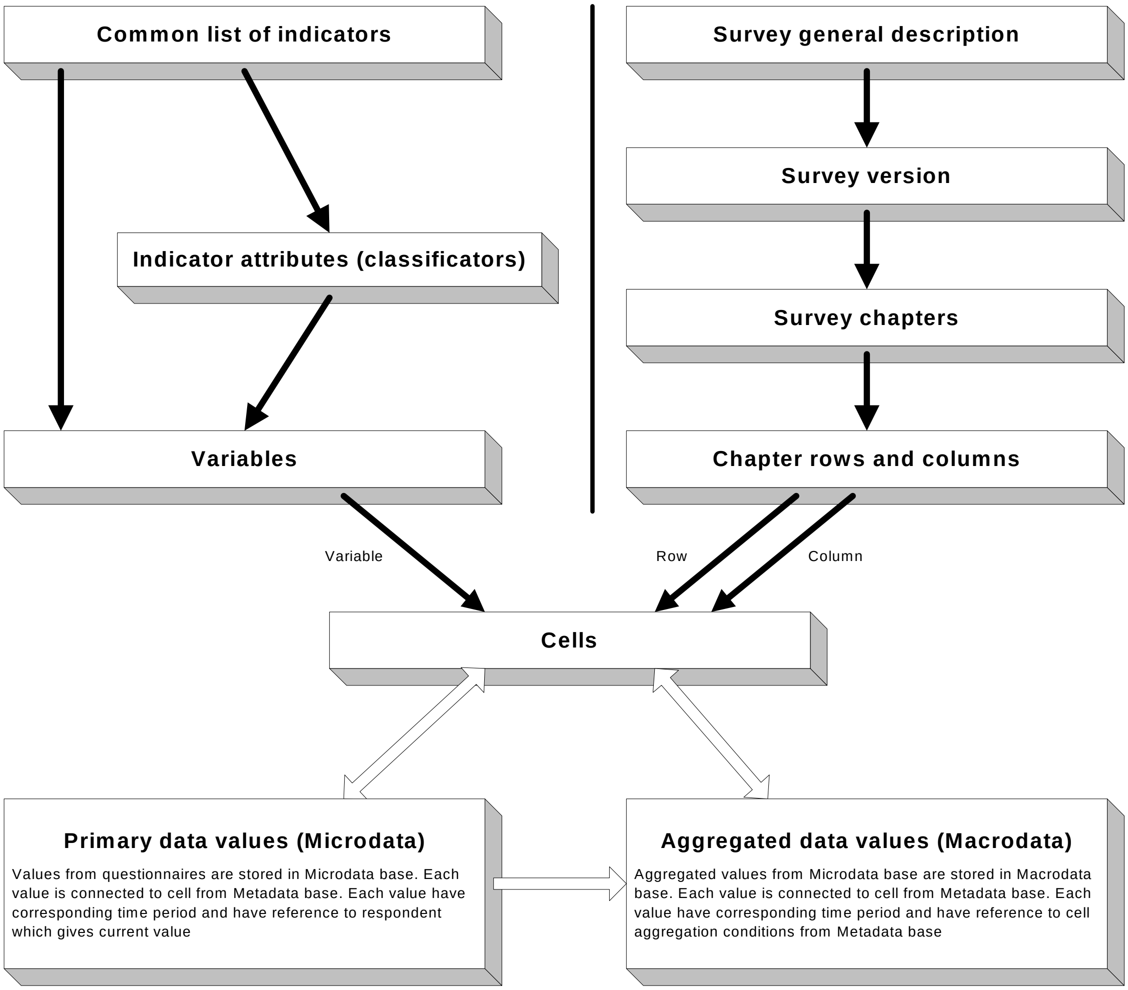

Common list of indicators
Indicator attributes (classificators)
Variables
Survey general description
Survey version
Survey chapters
Chapter rows and columns
Variable
Row
Column
Cells
Primary data values (Microdata)
Values from questionnaires are stored in Microdata base. Each value is connected to cell from Metadata base. Each value have corresponding time period and have reference to respondent which gives current value
Aggregated data values (Macrodata)
Aggregated values from Microdata base are stored in Macrodata base. Each value is connected to cell from Metadata base. Each value have corresponding time period and have reference to cell aggregation conditions from Metadata base

Figure 3. Meta database link with Micro data /Macro databases

All survey values from questionnaires are stored in Micro database and each value has relation to cell (from Metadata base), which describes value meaning. Also each value in Micro database has additional information about respondent, which gives current value and time period. The same situation is in Macro database, where aggregated values are stored. Each aggregated value has reference to cell (from Metadata base), reference to each value aggregation conditions (from Metadata base) and corresponding time period.
The processes chain of the statistical data production with usage of different meta data profiles are shown on Figure 4.

5. CONCLUSIONS

Design of the new information system should be based on the results of deep 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;
- Administration.

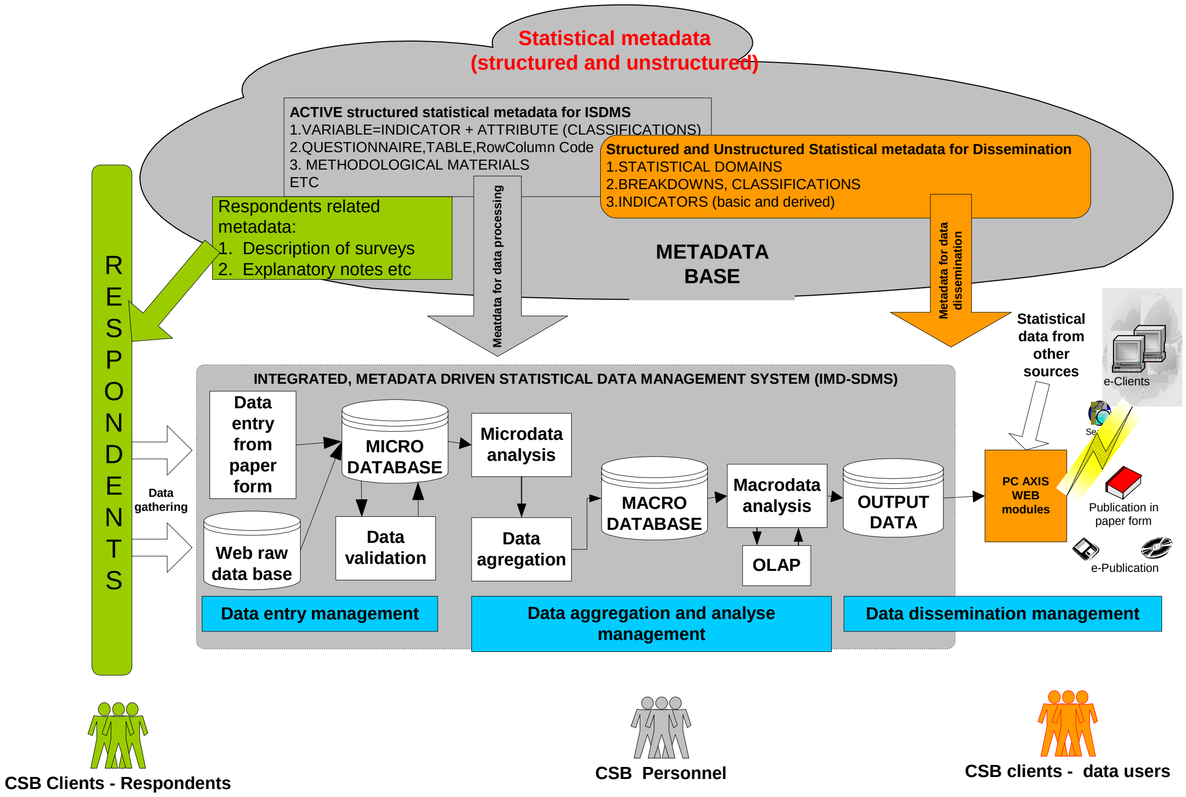

Statistical metadata
(structured and unstructured)
ACTIVE structured statistical metadata for ISDMS
1.VARIABLE=INDICATOR + ATTRIBUTE (CLASSIFICATIONS)
2.QUESTIONNAIRE,TABLE,RowColumn Code
3. METHODOLOGICAL MATERIALS
ETC
Structured and Unstructured Statistical metadata for Dissemination
1.STATISTICAL DOMAINS
2.BREAKDOWNS, CLASSIFICATIONS
3.INDICATORS (basic and derived)
Respondents related metadata:
1. Description of surveys
2. Explanatory notes etc
Meatdata for data processing
METADATA BASE
Metadata for data dissemination
RESPONDENTS
Statistical data from other sources
e-Clients
Se
INTEGRATED, METADATA DRIVEN STATISTICAL DATA MANAGEMENT SYSTEM (IMD-SDMS)
Data entry from paper form
MICRO DATABASE
Microdata analysis
Data gathering
Web raw data base
Data validation
Data agregation
MACRO DATABASE
Macrodata analysis
OLAP
OUTPUT DATA
PC AXIS WEB modules
Publication in paper form
e-Publication
Data entry management
Data aggregation and analyse management
Data dissemination management
CSB Clients - Respondents
CSB Personnel
CSB clients - data users

Figure 4. Statistical data production process chain

Initiative to move from classical stovepipe production approach to process oriented has to come from statistician's side not from IT personnel or administration.

Improvement of knowledge about metadata is one of the most important tasks through out of all process of the design and implementation phases of the project.

Clear division of the tasks and responsibilities between statisticians and IT personnel is the key point to achieve successful implementation.

To achieve the best performance of the entire system it is important to organize the execution of the statistical processes in the right sequence.

The new system is developed as really metadata driven, centralized system, where all data is stored in corporate data warehouse and which allows data processing using unified (standardised) approach for data entry and validation, data aggregation, data analysis and data dissemination for different surveys.

The high level of flexibility of the system has been achieved via using the statistical metadata as the key element of the system. Any changes in the survey content and layout can be done without participation of IT professionals and require just accordant changes in the Metadata base.

As the result of feasibility study we clearly understood, that there are some steps of statistical data processing for different surveys, which defy standardization, some 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 system have been developed to ensure use of the standard statistical data processing software packages and other generalized software available in market.

It is necessary to establish and train special group of statisticians, which will maintain Metadata base and which will be responsible for accurateness of metadata.
For the administration and maintenance of the system it is necessary to have well trained IT staff, which is familiar with the MS SQL Server 2000 administration, MS Analysis Service, other MS tools, PC AXIS family products and system Data Model, system applications.

Motivation of the statisticians to move from existing to the new data processing environment is essential.

Administrative restructuring could be usable to move from stove - pipe data processing to process oriented data processing.

For the proper installation and functioning of the system is necessary to use workstations not lower than Pentium II with RAM not less than 128 Mb equipped with OS MS Windows 95
(better MS W-2000) and MS Office 2000.

Summing up improvement goals and IT strategy realised in the system, there are mainly the following targets achieved by the system implementation:

- Increased quality of data, processes and output;
- Integration instead of fragmentation on organisational and IT level;
- Reduced redundant activities, structures and technical solutions wherever integration can cause more effective results;
- More efficient use and availability of statistical data by using common data warehouse;
- Users provided (statistics users, statistics producers, statistics designers, statistics managers) with adequate, flexible applications at their specific work places;
- Tedious and time consuming tasks replaced by value-added activities through an more effective use of the IT infrastructure;
- Using meta data as the general principle of data processing;
- Use electronic data distribution and dissemination;
- Making extensive use of a flexible database management for providing internal and external users with high performance, confidentially and security.

6. REFERENCES

1 Sundgren Bo (1999) “An information systems architecture for national and international statistical organizations” Invited Report, Meeting on the Management of Statistical Information Technology UN ECE Geneva, Switzerland, 15-17 February,1999

2. “Terminology on Statistical Metadata”,
Conference of European Statisticians, Statistical Standards and Studies No 53.

3. “Guidelines for the Modelling of Statistical Data and Metadata”
Conference of European Statisticians, Methodological Material;

4. Willeboordse Ad “Towards a New Statistics Netherlands”, blueprint for a process oriented organisational structure.