

# METADATA DRIVEN INTEGRATED STATISTICAL DATA PROCESSING AND MANAGEMENT SYSTEM

By Karlis Zeila, Vice president CSB of Latvia,<sup>1</sup>  
<Karlis.Zeila@csb.gov.lv>

## I. INTRODUCTION

The aim of this report is to introduce participants with the Metadata Driven Integrated Statistical Data Processing and Management System (Further in text - system) of Central Statistical Bureau of Latvia (CSB), which is developed and implemented with recent directions, international experience in statistical data processing.

From 1997 – 1999 Central Statistical Bureau of Latvia (CSB) experts in cooperation with experts contracted from PricewaterhouseCoopers were prepared Technical Requirements for the project “Modernisation of CSB – Data Management System”. Technical Specification written on above mentioned bases embody all technical and functional requirements for the new system where statistical Metadata are used as the key element in statistical data processing.

The main business and information technology (IT) improvement objectives that the CSB intends to achieve as the result of project have been identified and are as follows:

Using modern IT solutions:

- 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.

The development of the first version of the system was outsourced and started in 2000.

## II. TECHNICAL PLATFORMS AND USED STANDARD SOFTWARE

In view of complexity of the system and necessity to process large amounts of data, question about system performance became as one of the main implementation success factors.

Microsoft products originate with mostly the best price-performance criteria, which is very important indicator especially in case of development of complicated software systems, therefore, it was decided that CSB should standardize on Microsoft products. The first version of the system was developed on the software platform as described below:

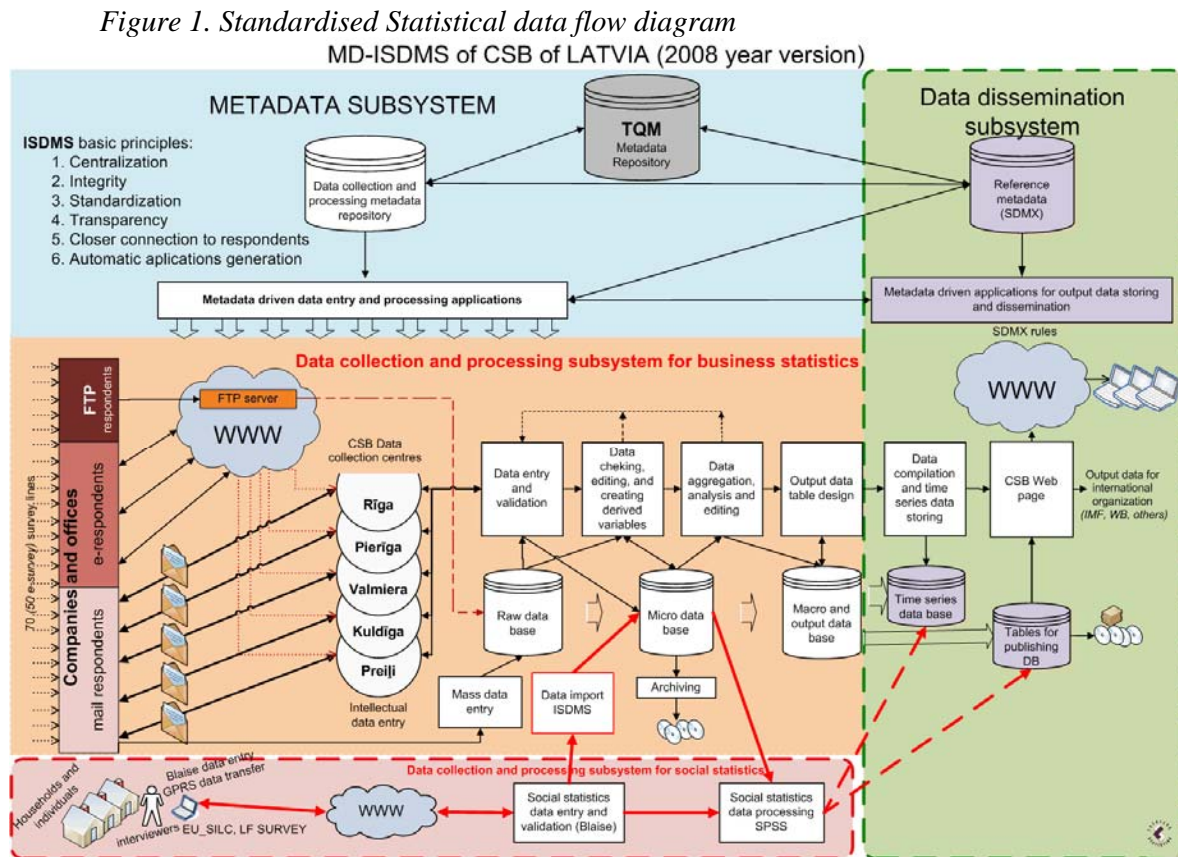
- The Microsoft SQL Server 2000 handles system databases (2002 until 2007).
- All applications comply with the client/server technology model, where data processing performed mostly on server side.
- Client software applications are developed using Microsoft Access 2000.
- Other components of Microsoft 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 tool for data dissemination was chosen product PC-AXIS developed by Statistics Sweden, which is widely used in different statistical organizations in different countries.

- At the time being (2008) we are moving from the data base platform MS SQL SERVER 2000- to version 2005 at the server side and from MS ACCESS to .net on the client side, thus getting independent from MS Office version changes.

## II. SYSTEM ARCHITECTURE

Before implementation of the system most of the statistical data processing took place on personal computers using logically isolated data bases and business applications (mostly FoxPro & MS Access) with limited access to information by other CSB staff, complicate data processing, analysis, dissemination and kept high costs for application development and maintenance. Classic Stove Pipe data processing approach with all appropriate technical incompatibilities existed as a consequence of the wide range of technology solutions used.

As the result of deep analysis of existing processes and data flows 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, which afterwards by Conference of European Statisticians has been approved as the guidelines for the development of Statistical Information Systems.

The system is developed as centralized system, where all data are 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.

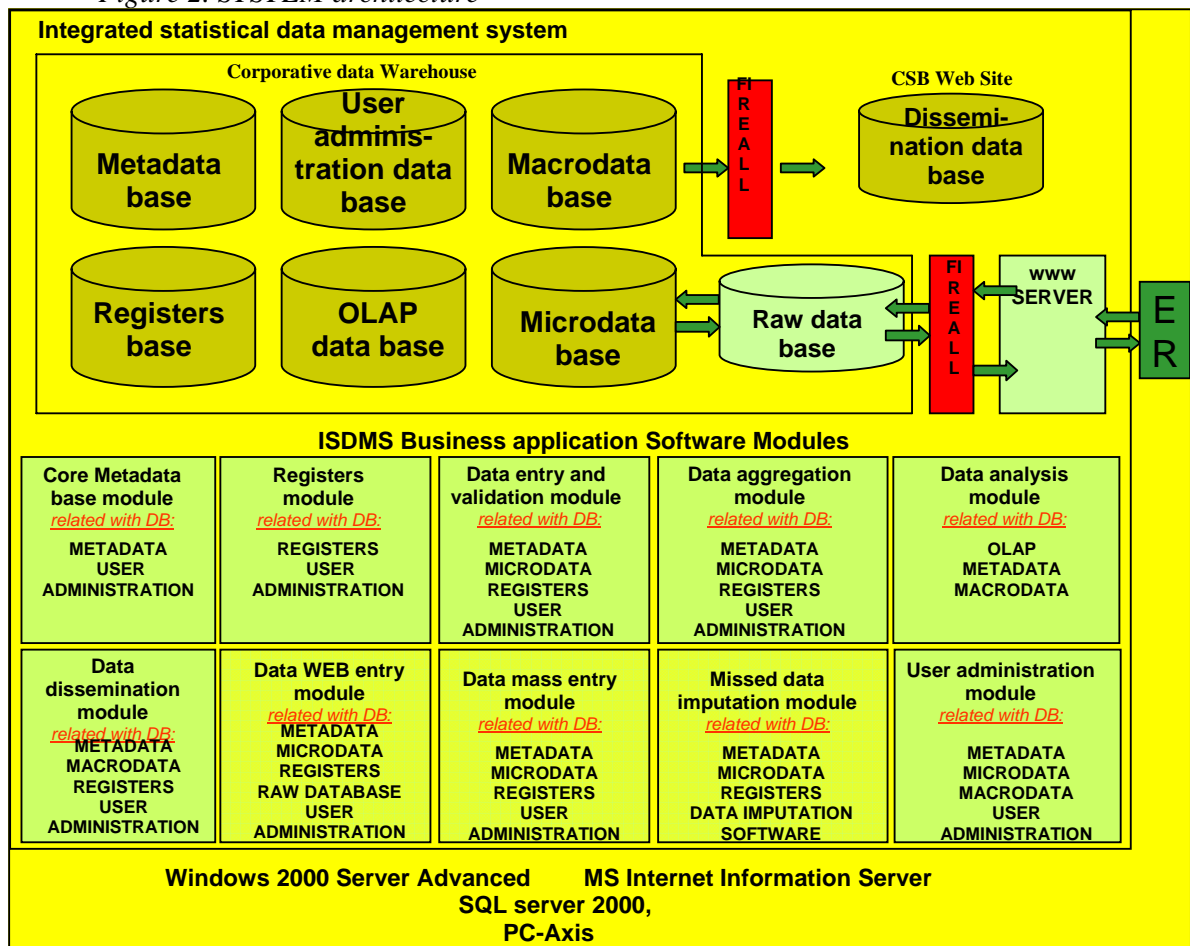
Important task during design of the system was to foresee ways and to include necessary interfaces for data export/import to/from commercial statistical data processing software packages and other commercial software like Excel, SPSS and etc.

System is divided into 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.

Figure 2. SYSTEM architecture



### **A. Core 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. Metadata base data handled by this module are used by all other modules of the system.

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”.

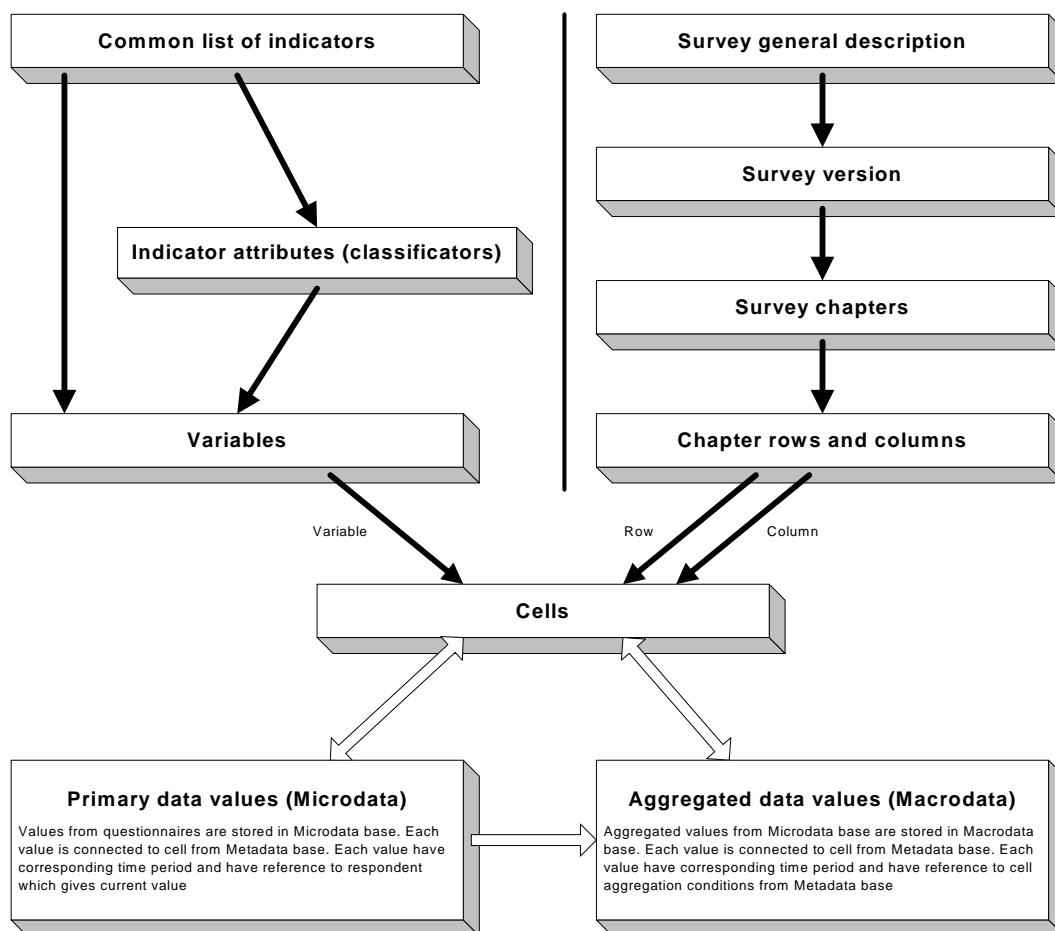
The data in the Metadata base, in essence, is information about micro and macro data i.e. description of the numerical data within the statistical production process and the real world meaning of this numerical data. Also the system Metadata base contain description of statistical surveys itself, their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

The system ensure that Metadata base is used not only as knowledge base for statisticians and users, but also as the key element for the creating universal, common, programming-free approach for different statistical surveys data processing instead of development of software specially for certain survey, where any change of the survey will require a corresponding adaptation of the programs source code, and where it will be also necessary to develop new software application for each new survey.

System users can easy query necessary data form Microdata / Macrodata bases navigating via Metadata base. Metadata are widely used for data analysis and dissemination.

The creation of the core Metadata base structure model and Microdata/Macrodata base structure model was the primary task of the system development. Metadata base is linked at database structure model level with Microdata base and Macrodata base (see figure 3). Correct and carefully deliberative databases structure model design is the basis for successful further system development and implementation.

Figure 3. Metadata base link with Microdata/Macrodata bases



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 is necessary to create survey version, which is valid for at least one year. If survey content and/or layout do not change, then current survey version and its description in Metadata base is usable for next year.

Each statistical survey contains one or more data entry tables we shall call them chapters. In Metadata base for each chapter it is necessary to describe table type. System recognizes three types of tables

- constant tables with fixed number of rows and columns,
- tables with variable rows number and constant number of columns,
- tables with variable columns number and constant number of rows.

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.

During survey version content description, in the Metadata base we are using information about statistical indicators, which are stored in Metadata base in the common list of indicators. In cases where in the survey version we are foreseeing usage of newly created indicator we have to add its description to the common list of indicators. Indicators themselves 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.

For each indicator in the system it is possible to define attributes. The role of attributes in the system plays classifications, which gives opportunity to describe and store values of indicators in breakdown by classification items. Indicators could be presented without attributes.

The meaning of the Variable used in the entire system is a value of an indicator for the exact item of classification. Therefore the value of the variable is equal to the value of the indicator only in cases when attributes are not in use and represent the total value of the indicator for a measuring object.

When indicators and attributes are defined, it is necessary to define variables. As we have described above 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 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.

All survey values from questionnaires are stored in Microdata base and each value has relation to cell (from Metadata base), which describes value meaning. Also each value in Microdata base has additional information about respondent, which gives current value and reporting period.

In the latest version of the system for identification of variables has implemented several additional information fields like:

- number,
- text,
- classification code,
- date.

The same situation is in Macrodata base, where are stored aggregated values. Each aggregated value has reference to cell (from Metadata base), reference to each value aggregation conditions (from Metadata base) and correspondent reporting period.

Such cells definition – combination from two sides – has several advantages. First of all, for statisticians, who works with a survey every day and are familiar with survey content, it is the easiest way to work with values using row and column language (row and column codes). Also all validation rules for survey is described in Metadata base using row and column coordinates as well.

Described approach is useful inside one survey version, which is valid for at least one year. If we are introducing any changes in the survey, then we have to describe a new survey version in the Metadata base in which very often for the same cell row and column coordinates differs from the previous version. Therefore for data analysis purpose users can use another approach, where cells values it is possible to search from variables side (list of variables stored in the Metadata base). For the definition of variable cell location is not important.

Metadata base module contains following main applications:

- General description of statistical survey;

For every new statistical survey starting Metadata filling in it is necessary to enter first of all basic information about this survey, e.g., name, index (code), periodicity. Application allows attaching to survey general description methodological documents.

- Description of survey version;

Statistical surveys content and layout could change each year or stay without changes several years. Each statistical survey content and layout, called survey version, have to be approved at the end of every year and is in force at least during next year. Application purpose is to register for each survey version information about survey version name, code, periodicity, information in which years survey version was in force, date of approval of survey version, tieback with State statistical programme chapters and survey division in chapters. Methodological documents files, which belongs to survey version it is possible to attach as well.

- Description of indicators and attributes of statistical survey;

Indicators are sorted by thematic groups and sectors.

For each survey version statisticians select corresponding indicators. Indicators are characterized by code, name, periodicity, measurement unit. It is possible to describe new indicator or select existing indicator from common list of indicators.

Next task is to attach attributes (classifications) to indicators. Attributes There are cases, where indicators have not attributes. Attributes description is characterised by item name and attached corresponding classification name. When indicators and their corresponding attributes are defined, application allow automatically generate variables, which are combination of indicators and attributes. By default variable name format is following –“indicator name / attribute name”. Statisticians can manually edit this default variable name if it is necessary. For each variable it is necessary to indicate whether variable will be used for aggregation or no.

- Description of content of statistical survey chapters;

In this application statisticians for each survey version chapter describe rows and columns, i.e., allocate codes, descriptions. In application main window are shown also all previous created corresponding variables of survey. When rows and columns for chapter are described, each row and column combination should be connected with variable to create cell. For cell value it is necessary to define format.

- Description and maintenance of validation rules of statistical survey;

For the description of the validation rules a special notation language with very simple syntax has been created.

With help of that validation rules can be described or previously described rules could be changed any for survey version at any time, when it is necessary. To make it possible for system users, without involving in this process IT personal (programmers), validation rules maintenance application was developed. It was created special, easy understandable syntax for description of validation rules using surveys version chapters rows and column codes. Each validation rule description contains validation rule code, error message text, description of validation rule, validation rule conditions. When all necessary validation rules are described and stored in the Metadata base, there is possible to generate validation procedure automatically. When statisticians need to make any changes in validation rules, they simply change validation rules description in Metadata base using developed application.

Validation procedure executes from Data entry and validation module applications.

- Derived indicators calculation description

Description of the algorithms for the derived indicators and corresponding variables calculation in accordance with the attributes used has to be stored in the Metadata base and automatically calculated values stored in Microdata base.

- Description of aggregation conditions of statistical survey;

Aggregation conditions are defined for each survey version cell or for group of cells, which have the same aggregation conditions. Aggregation conditions describe how data from surveys are summarized. This means, that user for each survey cell (or group of cells) must describe aggregation type used for data, which should be aggregated. In Metadata base it is possible to define three types of aggregation conditions. First type is data aggregation using respondent parameters. Each respondent have such parameters as territory, NACE branch and so on. So, if we describe that data values located in the cells, which must be aggregated by respondent parameter “territory”, then respondents are grouped by territories and corresponding respondent data values are summarized. System supports the following aggregation functions like sum, count, min, max and mean.

Second type of aggregation conditions is data aggregation by indicator attributes. Such data aggregation type is used when cells are connected with indicators which attributes are international classifications like PRODCOM. In this case cells values for each PRODCOM code is summarized from all questionnaires where data about this concrete PRODCOM code are.

Third type of aggregation conditions is data aggregation for variable (1) by another variable (2) data value in the same questionnaire. Variable\_(2) data values must be possible to classified. Therefore variable (2) data values can be used as classification for questionnaires variable (1) data values grouping and summarizing.

Using aggregation conditions description it is possible automatically generate aggregation procedure, which is executed from Data entry and validation module applications:

- Grouping of classifications records;

Purpose of such application development was to ensure facility for the system users to re-group, reorganize classification records in necessary hierarchic groups. These classification groupings statisticians can create and store in the system as much as they need. Classification groupings are used for preparation of different reports. Such approach lightens production of reports. During classification grouping process no one classification record can be removed. Also already defined classification grouping it is possible to remove from the system only if no one report is not produced based on this classification grouping.

- Description of reports;

One of the last steps in statistical surveys data processing before dissemination is reports creations. For reports creation are used data values aggregated by classifications groupings described above and reports preparation description from Metadata base.

Reports preparation description in Metadata base is possible to create for each survey version. Each reports preparation description consists of two steps:

- The first step is the general description of the report - name, periodicity, which is used for data summarizing.
- Second step is Row definition. It is necessary to define report grouping conditions – which classifications names and/or users created classification groupings names will be used in report. These grouping conditions specify how data in report rows will be grouped and represented. For each report it is possible to describe up to six grouping conditions, one of which can be grouping by time periods. If in report preparation description is specified data grouping by time periods, then in each report row data will be represented for an exact time period, starting from current time period and continuing with previous time periods data in descending order.
- Third step for report preparation description is columns definition. Each column have column name and for each column we must define, which survey cell data will be calculated in current column. If it is necessary, for each column is possible to set up additional grouping conditions, but these grouping conditions have limitations. If user for column have defined data grouping by classification, then for such column is possible to calculate and represent cell data only for one classification record. Also, if for column is defined data grouping by time periods, then for such column is possible to calculate and represent cell data only for one fixed periodicity unit, for example, if we have survey with monthly periodicity, then we can create report preparation description with year periodicity where we include 12 columns (January, ..., December), which represent one survey cell value in each month of the year. Each report column has such parameter as aggregation function, which is used for column data values calculating. It is possible for each column choose one of standard aggregation functions like sum, count, minimal, maximal, average (by default it is sum).

When users create report preparation description and define grouping conditions for report and report columns, they can use only those classifications, which are used in survey cells aggregation conditions.

Calculated reports results are stored in Macrodata base.

Using reports preparation description it is possible automatically generate report data values calculation procedure, which is executed from Data entry and validation module applications.

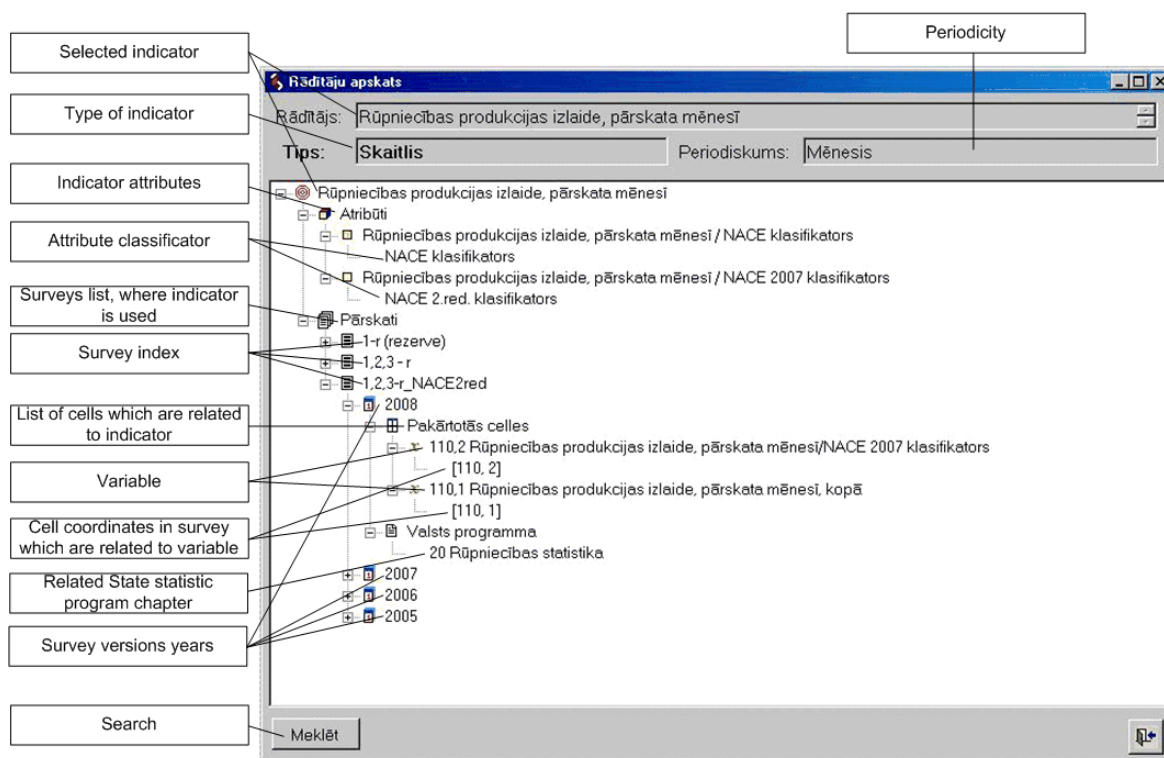
- Common Metadata base data browsing;

Using above-mentioned Metadata base module applications it is possible to retrieve and view from Metadata base information about one concrete survey version (exception is common list of indicators, which is independent from surveys) after that for another and so on. To ensure possibility to analyse Metadata information from Metadata base for different surveys in one application, common Metadata base data browsing applications were developed.

One of the Metadata base browsing applications gives possibility to see how one chosen indicator relates with other Metadata base objects. Results are represented in tree form (see figure 4). This application represents following information about one concrete statistical indicator:

- Indicator hierarchy. It is possible to see child and parent indicators, which have relationship with current indicator;
- Indicator attributes. Application shows all indicator attributes and for each attribute users can see connected classifications records;
- Surveys list where indicator is used;
- For each survey application shows in which years current indicator is used in survey;
- About each survey version year users can see list of cells, which are related to indicator;
- For each cell application shows information about variable, which is connected to current cell and indicator;
- For each cell application shows cell coordinates in survey – what row and column code current cell have.

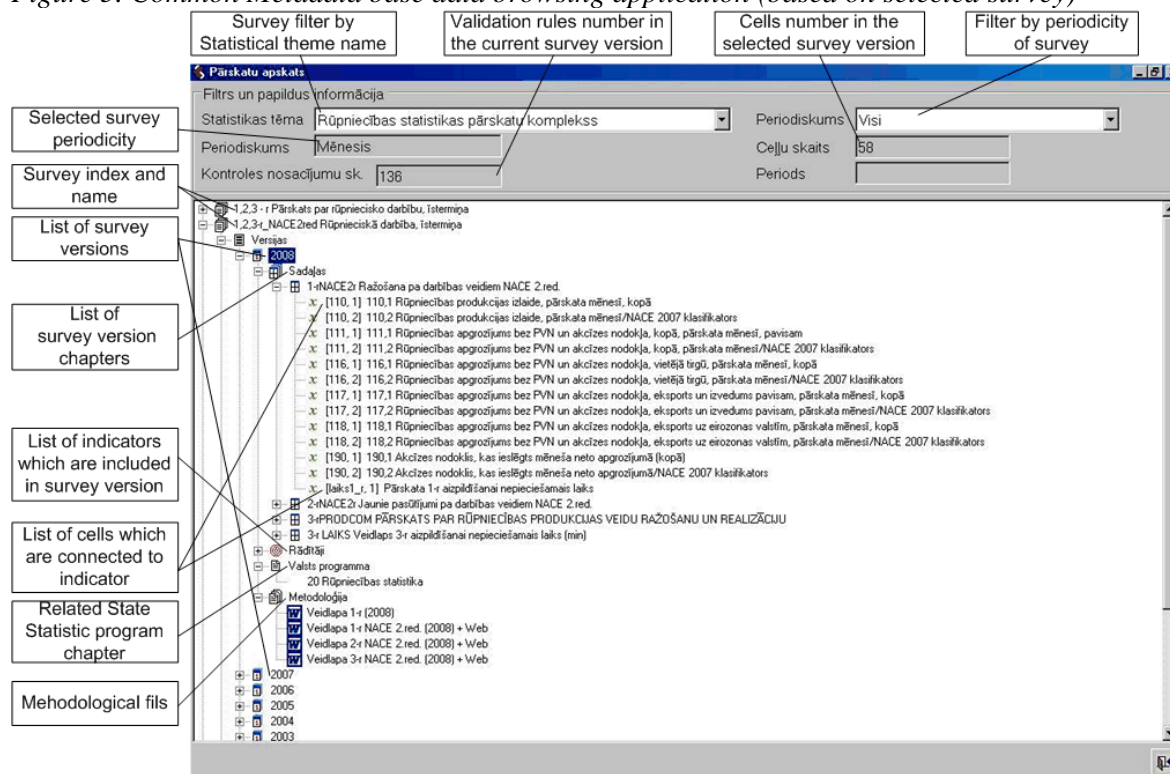
*Figure 4. Common Metadata base data browsing application (based on selected indicator)*



Another Metadata browsing application gives possibility for users see survey content and relationships with another Metadata base objects (see figure 5). Results are represented in the same way as in previous described Metadata browsing application for indicators. Application shows to users following information:

- Survey Metadata browsing begins with survey index and name;
- Application shows each survey version, which is related to selected year. For selected survey version users can see how many chapters are defined;
- Under each survey chapter application shows list of cells, which are created. Cell coordinates (row and column) it possible to see on each cell in selected chapter as well as variable name, which is connected to each cell;
- All indicators which are used in survey version we can see in another branch of the tree;
- If indicators have attributes, application shows them as well as classifications, which are connected to attributes;
- For each indicator is possible to see with which survey version cells current indicator is related. And for each cell users can see cell coordinates (row, column) and connected variable name;
- For each survey version application shows State statistical program chapters to witch current survey version belongs;
- For each survey version it is possible to see methodological documentation files, which are connected to survey version;
- Selected survey version periodicity, number of cells and number of validation rules application shows in three fields under tree form.

Figure 5. Common Metadata base data browsing application (based on selected survey)



Some remarks about implementation of other survey version if we have defined survey version in the system:

- Variables (combination of indicator and attribute) are related to an exact survey. When statisticians implement new survey version, they can use in previous survey version already defined variables or to create new ones from new indicators. Variables, which values have been stored, never can be deleted. If in comparison with previous survey version new survey version do not contain determined variable, during this survey version description statisticians this variable do not connect to survey version chapter row and column combination, i.e., do not define cell with such variable;
- During implementation of other survey version system Metadata base module ensure facility for chapter's rows and columns as well as cells description information copying from previous survey version instead of rewriting this information for every new survey version. Also it is possible to copy validation rules description from previous survey versions.

Specially established and trained people group called "Metadata group" operates the Metadata base module. They have rights to perform Metadata entry, actualisation, introduce changes and are responsible about accurateness of Metadata. It is very important, that Metadata entered into Metadata base are carefully checked and correct, because these Metadata are used for automatic generation of data entry applications, validation, aggregation, reports preparation procedures as well as during data conversion for OLAP and PC-AXIS needs.

## B. Registers module

Register is stored in separate databases and managed by a specific module – Registers module.

Register module contains Statistical Business register and Intrastat register. Possibility to expand register module with other registers has been foreseen.

Main tasks of Statistical Business register are:

- Maintenance of the database of economically active enterprises;
- Preparation of the sample frame of economically active enterprises for annual and quarterly statistical surveys;
- Maintenance of lists of respondents for almost all statistical surveys;
- Preparation of information according to requirements of local and international organizations, institutions and other data users;
- Preparation of information for publications.

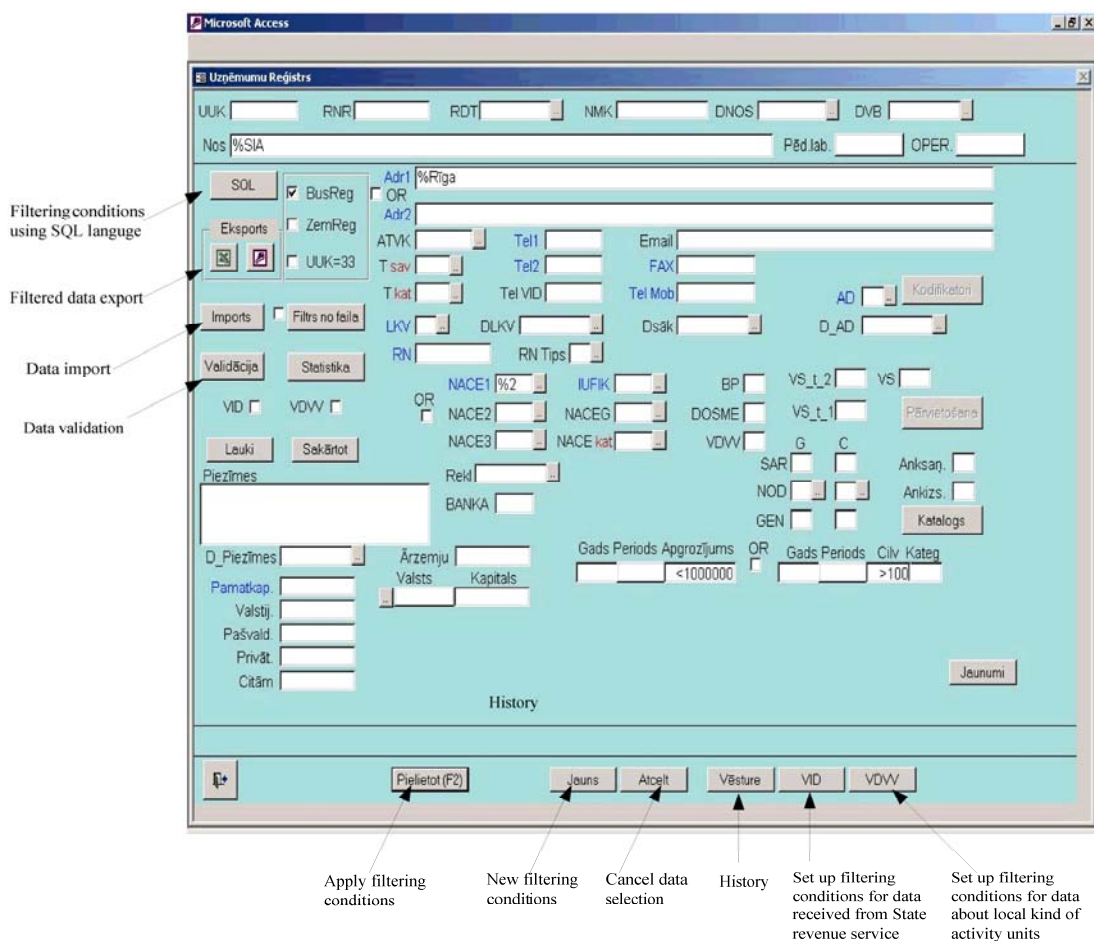
There users of the Register module are divided in three categories:

- Administrators (full functionality is allowed);
- Operators (data import is not allowed);
- Users (data import, entry, update is not allowed).

Statistical Business register provide following functionality:

- Data import:
  - monthly data import from administrative registers – State Enterprise Register and Taxpayers register of State Revenue Service;
  - State Enterprise Register monthly sends fixed format files with information about creations and closures, restructuring of existing enterprises, legal changes of names and addresses of enterprises. After data import all records should be manually checked and approved;
  - Quarterly CSB receive full copy of State Business register and automatically compare the data with Statistical Business register;
  - State Revenue Service monthly sends fixed format files with information about payment of taxes and persons declared;
  - Monthly information from Latvian Central bank (Balance of Payment register);
  - Data import from other external data sources and statistical surveys.
- Data export:
  - Monthly data export to State Enterprise Register, Taxpayers register of State Revenue Service and Latvian Central Bank in a fixed format files;
  - Lists of respondents (catalogues) for statistical surveys;
  - Information for data users according to special requests;
  - Possible Export formats – Microsoft Access, Microsoft Excel, dBase, Paradox data base format.
- Data filter (see figure 6):
  - Filter from file allows to extract list of enterprises or local kind activity units according to ID codes in the files;
  - Filter according to different criteria (e.g. only one region; only NACE 2010; NACE 1 2010 or 2036, etc);
  - For complex filter, SQL language scripts or Microsoft Access Query tools should be used.

Figure 6. Statistical Business register data searching application



- Data entry and update (figure 7):
  - The application for update of selected group allows to perform equal changes in the same field in Business register for the whole selected group;
- Data entry and update for single enterprise and it's local kind of activity units. History of enterprise (figure 8):
  - Only the most important events are foreseen to be saved automatically for the history (not all fields of BR);
  - If by mistake unnecessary records are saved, only administrator may change or delete records from the history.
- Data validation (figure 9):
  - The validation may be performed for all enterprises of Statistical Business register, for the selected group of enterprise, or for single enterprise;
  - All possible validation rules at once, or selected rules may be checked;
- There are some cases, when some validation errors are necessary to ignore. The administrator may put a mark to ensure, that in next time, when validation procedure will be executed, these errors will be not displayed in the errors list. If necessary these marked errors may be Local kind of activity units (figure 10):
  - All units comprising an enterprise are included in this dataset. Each enterprise may have at least one or several local kind of activity units. The units operating in different addresses or operating in the same address, but

in different economical activities are registered separately as a local kind of activity units;

- Data import from external files is possible to perform in this application;
- Data export may be performed for different data requests;
- Validation rules between main enterprise and its units are foreseen.

*Figure 7. Statistical Business register data entry and update application*

Uzņēmumu Reģistrs

UUK [99767391] RNR [000320707] RDT [22.07.1994] NMK [40003207074] DNOS [ ] DVB [20.05.2002]

Nos [RANLET PRODUKCIJA SIA] Pēd lab [APukite] OPER [APukite]

UUK [99767391] Nos [RANLET PRODUKCIJA SIA] Adr1 [Maskavas iela 257, Rīga, LV-1019]

Adr2 [Maskavas iela 240, Rīga, LV-1063]

ATVK [010093] Tel1 [7187072] Email [ ]

T sav [01] Tel2 [7187451] FAX [7112742] Avots [ ]

T kat [01] Tel VID [7245039] Tel Mob [9221699] AD [ ]

LKV [ ] DLKV [ ] Dsk [ ] D\_AD [ ]

RN [ ] RN Tips [ ]

NACE1 [1822] IUFIK [0151] BP [ ] VS\_t\_2 [1B] VS [1B]

NACE2 [ ] NACEG [1822] DOSME [ ] VS\_t\_1 [1B]

NACE3 [ ] NACE kat [1822] VDVV [ ] G [ ] C [ ]

Rekl [26.07.1993] Gads Periods PVN SAR [1] [1] Anksan. [ ]

BANKA [ ] NOD [B] [B] Anksz. [ ]

GEN [1] [1] Katalogs

D\_Piezīmes [ ] Ārzemju [48630] Gads Periods Aprozijums Gads Periods Cilv Kateg

2002	Gads	892186	↑	2002	Gads	186	3	↑
2003	1.cet.	172990		2003	1.cet.	176	3	
2003	1.nusa			2003	1.nusa			
2003	3.mēn.			2003	3.mēn.			
2003	Gads			2003	Gads			

Piezīmes

Pamatkap [97260] Valsts [48630] Valstj. [ ] Pašvald. [ ] Privāt. [48630] Citām [ ]

Filtrs (F7) Grupa Saglabāt(F2) Pievienot Dzēst Atcelt Vēsture VID VDVV Validēt (F8)

To filtering conditions Data changing for multiple records (group) Save record New record Delete record Cancel changes History for selected enterprise Information for selected enterprise from state revenue service Information about local kind of activity units for selected enterprise Validation

Figure 8. History of enterprise

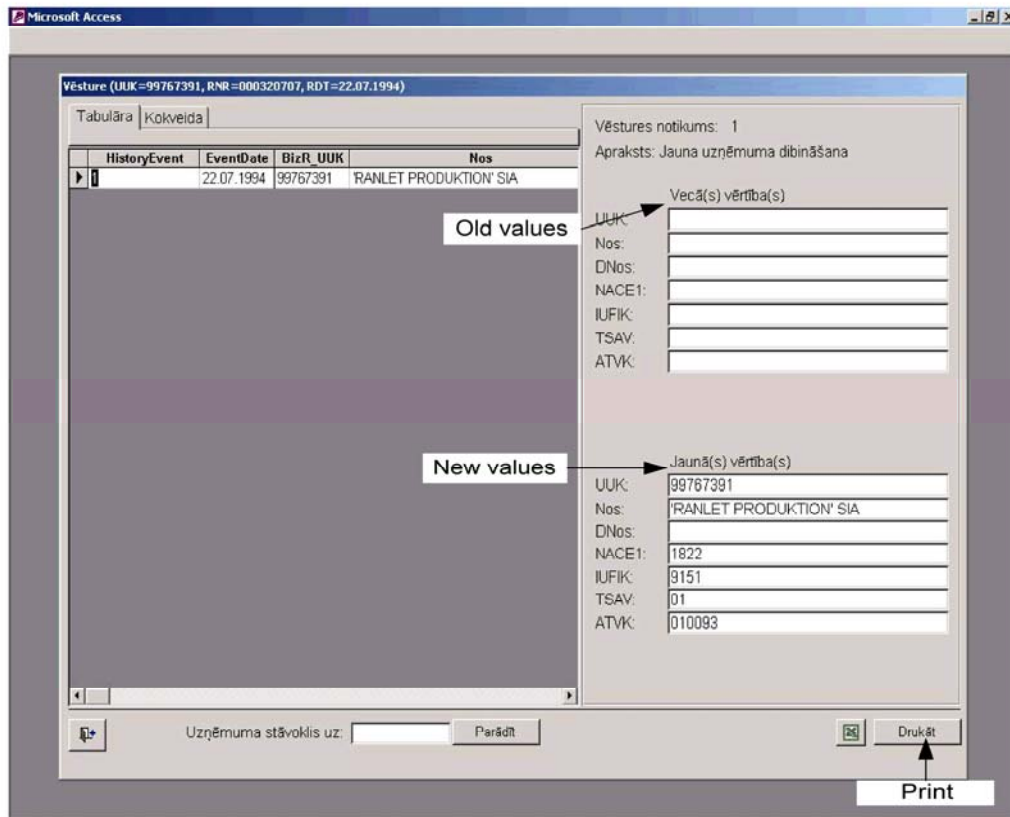


Figure 9. Data validation output

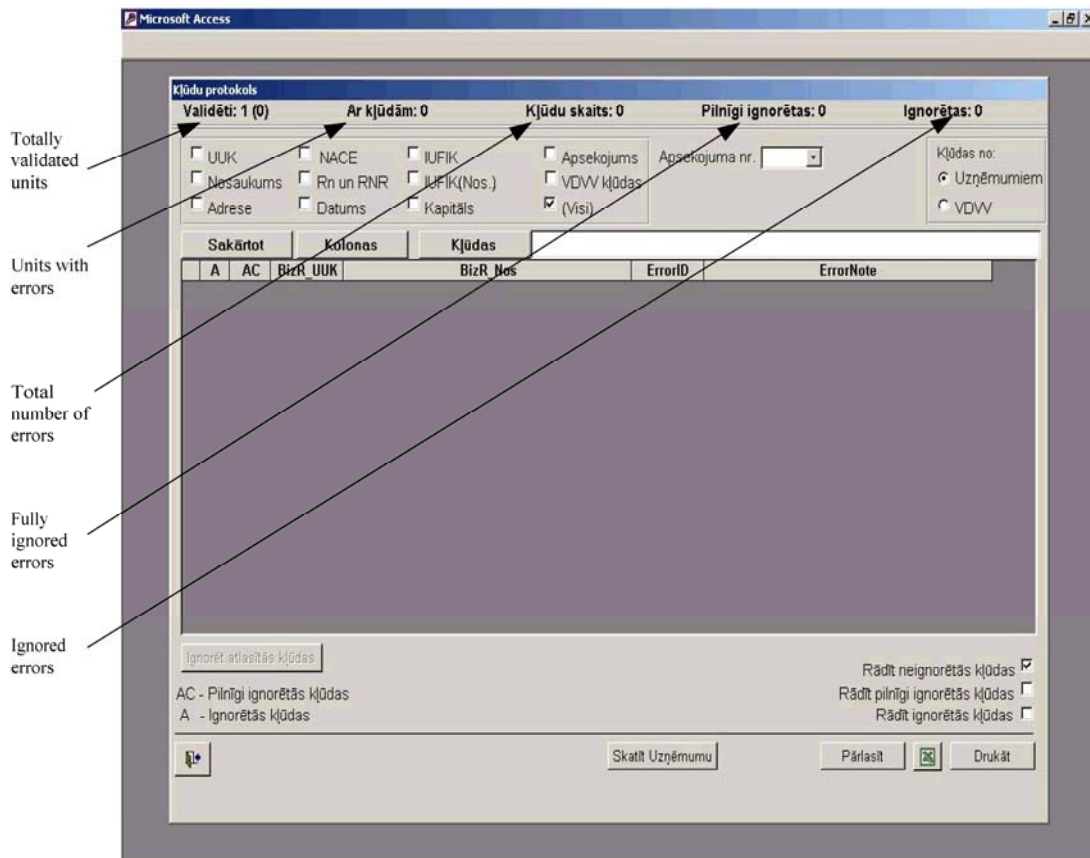


Figure 10. Local kind of activity units

Enterprise name

Local kind of activity units (LKAU's)

Name of LKAU

Address of LKAU

Exit LKAU

Save changes

Add new entry

Delete entry

Cancel changes

### C. Data entry and validation module

Module provides standardised approach to different statistical surveys data processing. For surveys data processing is implemented automated data entry forms generation, automated generation of data validation, aggregation and reports creation procedures by using entered Metadata base information.

If it is necessary to make any changes for survey content and/or layout, then it is necessary only to change survey description in Metadata base.

Such approach make much more easier system maintenance and reduce costs. It is not necessary in case of changes re-write program codes, which can be done only by IT professionals.

For each selected survey for selected period following main functionality is available respondent's list maintenance:

- Data entry and validation;
- Data aggregation;
- Reports creation;
- Data export/import.

The system ensures linking of the statistical survey to a particular list of respondents obtained from Business or Intrastat register. Each survey version for each period has its own list of respondents. Respondent's maintenance application is shown in figure 11.

Figure 11. Respondent's maintenance application.

Respondents list

Data input

Validation

Information about Respondents creation

Information about data entering

Filter

Add

Cancel

Update

Export

Import

Save

Delete

History of imputations

Data entry and validation application is displayed in figure 12.

Data entry screen shows selected respondent survey chapters using tab control and allows select, enter, update and delete data. There are three types of survey chapters:

- With fixed rows and columns number;
- With fixed rows and variable columns number;
- With fixed columns and variable rows number.

For example, if we have survey with chapter where is several indicators in chapter rows and as indicators attributes are used values from NACE classification in chapter columns, then we have chapter with fixed rows and variable columns number, because each enterprise can have different number of economic activities, which are classified in NACE classification.

Application manually allows adding or deleting rows or columns, when it is necessary.

History data (previous period's data) can be displayed in two ways:

- History data of all previous survey periods for currently selected row in data grid;
- History data for all data rows at the same time for one of selected previous periods.

It is possible to hide historical data section to increase screen space for data entry and later again to show it.

Figure 12. Data entry and validation application

The screenshot shows a data entry and validation application window. The main area contains two tables. The first table has columns 'Rnr', '1', and '3212'. The second table, titled 'Periods', has columns 'Periods', '1', and '3212'. A toolbar at the bottom includes buttons for 'Delete row', 'Delete cell', 'Delete survey', 'Show/hide history', 'Export to Excel', 'Validation', 'Save', and 'Cancel'. Callout boxes identify various UI elements and data points.

Rnr	1	3212
110	46.8	46.8
111	13.6	13.6
116	13.6	13.6
117		
118		
190		
laiks1_r	15	

Periods	1	3212
2008M01	43.3	43.3
2007M12	47.3	47.3
2007M11	35.4	35.4
2007M10	44.4	44.4
2007M09	30.2	30.2
2007M08	29.2	29.2
2007M07	33.5	33.5
2007M06	29.9	29.9
2007M05	28.7	28.7
2007M04	37.3	37.3
2007M03	33.4	33.4
2007M02	26.9	26.9
2007M01	25.4	25.4
2006M12	59.6	59.6
2006M11	36.9	36.9
2006M10	22.8	22.8

After executing of validation procedure of respondent survey data (we can run this procedure from data entry and validation application) opens form with list of validation errors. In this form for each error is following information (see figure 13):

- Error number;
- Error description;
- Error type:
  - Acceptable – error can be marked with “A”;
  - Unacceptable – error is marked with “X” and its type cannot be changed (defined in Metadata base);
  - Not marked.
- Error type reason – for some errors it is necessary to describe the acceptability reason.

Errors marked with “X” are critical and there is no way to pass them during validation process. All other errors in the list of validation errors can be marked with letter “A”. Errors marked with “A” will not appear in error list next time (to see in errors list also acceptable errors checkbox “show all errors” must be marked). These are less important errors in surveys, which can be passed as not significant for further survey data processing.

Also it is possible to run validation procedure for all respondents of one survey. Then we receive list of respondents, which surveys data have validation errors.

To define exceptions in validation rules defined in Metadata base two additional applications were created, which interacts with data validation process. First application allows to “switch off” one or more validation rule defined in Metadata base for survey. Then during validation process such validations rules will be not checked. Second application allows to

“switch off” one or more survey validation rule defined in Metadata base for the respondent and again during validation process such validations rules for this respondent will be not checked.

Figure 13. Validation errors list

Respondent's code: [ ] Respondent's name: [ ] Error priority: [ ] Error type: A - acceptable, X - unacceptable

Pārskata datu ievade [1.2.3-r. NACE2red par: 008M03]

UUK: xxx ... Parādīt Pacija: 440 NMIK: xxx RDATE: 10.12.1991 LKV: LDAT: SAR: 1 NOD: B

Nos: xxx NACEkat: 3622 NACE1: 3622 NACE7kat: 3212 NACE71: 3212

Kļūdu protokols

Respondenta kods: 00227301 Prioritāte: 2 (Vaiaks: 2) Kolonnas V decimālās daļas garums: 3

Nr	Kļūdas apraksts	V1	V2	V3	Kat.	A	Pakaidrojums
1-r.14	Uzrādīt iemeslu [110.1] izmaiņas ar iepriekšējo mēn. lielākas ...	-45.940	25.300	46.800	1	X	piebērkoti saražoti produkcijas krājumi
1-r.15	Pārbaudīt [111.1] izmaiņas ar iepriekšējo mēn. lielākas par 5...	116.176	29.400	13.600	1	X	pieprasījuma palielinājums
Q-3-r.11	Natūrā uzrādītā prod vērtība (3a.tūkst.Ls.) < 80% no 111r.ce...	0.000	86.300	0.000	1		
Q-3-r.13	Nav ievadīts pārskats 3-r. 111r.cet. summa=	86.300			1	X	

Rinda kopēšanai uz clipboard: V1=-45.940; V2=25.300; V3=46.800; Rādīt visas kļūdas Pazīme, ka kļūdu nevar atslēgt X

Error number      Error description      Show all errors/ show unacceptable and not marked errors      Calculated validation values for analysis

This row can be copied to clipboard for analysis

Survey data for specified period can be exported to fixed format Microsoft Access database file for external use. Also fixed format Microsoft Access database file can be imported into the system.

Other features, which are included in functionality:

- to ensure possibility for each statistician using module applications to work only with their set of respondents, in application is included facility to set filter, which select necessary set of respondents for department;
- for safety needs was included functionality for data locking. When it is necessary, responsible statistician can lock survey data, which forbid further data changes. Before data can be changed it is necessary to unlock survey data;
- statisticians can create lists, which helps to follow to surveys data input process. There are there types of lists:
  - List of respondents who have sent in survey;
  - List of respondents who have not sent in survey;
  - List of respondents, which survey has one or more errors after validation.

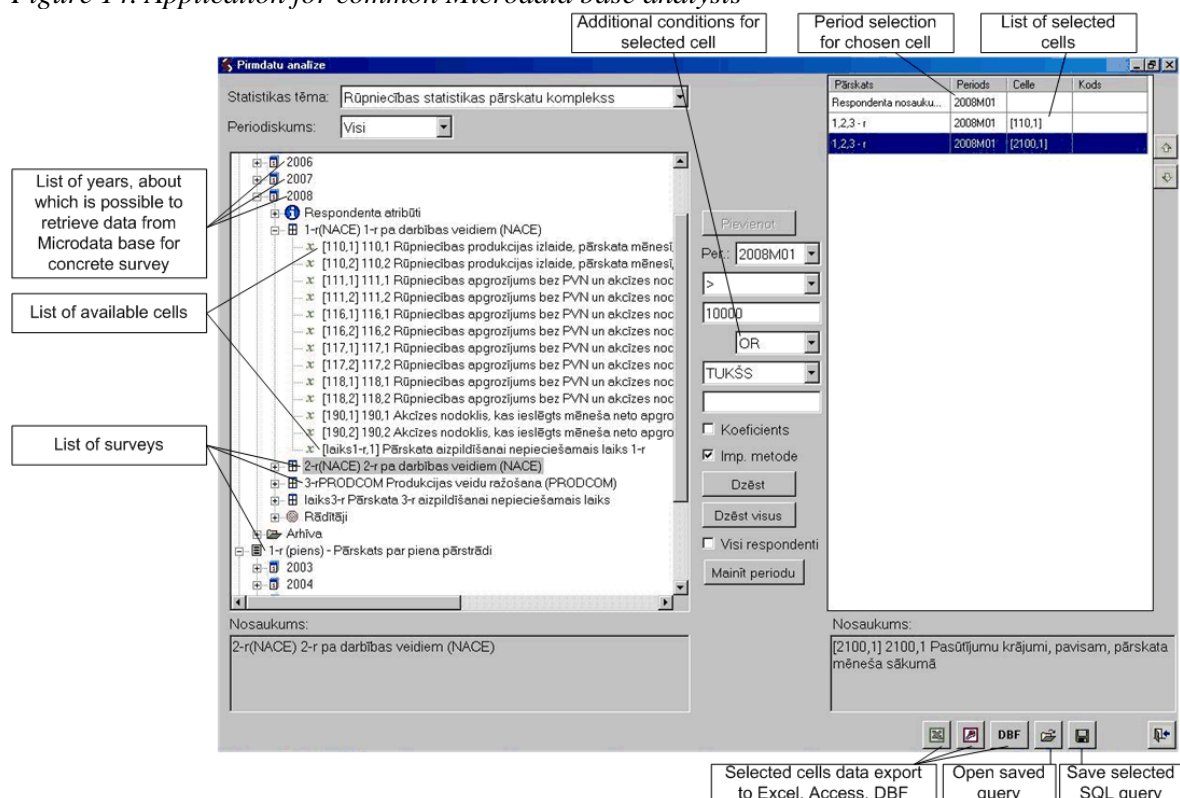
To ensure high and stable performance of the system data from the surveys elder than three years has been stored in a separate tables so called archive of the Microdata base. This data could be retrieved, but it can't be changed and new aggregations can't be made. But in a special case when it is necessary to make new aggregations data from archive tables have to be moved to active part of the system.

From Data entry and validation module users can create reports using report preparation description from Metadata base. Also from Data entry and validation module users can run application, which is used to maintain aggregated data and another application for common Macrodata analysis (see description of these applications in the next section of report – “Data aggregation module”).

Additionally in Data entry and validation module is included application for common Microdata base analysis navigating via Metadata (see figure 14). For end-users it is not necessary to know where and how Microdata are physically stored in Microdata base.

Using this application, users can select and combine data from different surveys for different time periods and to define different conditions for data, which should be queried. All selected data will be exported to Microsoft Excel, Microsoft Access or .dbf file format. Prepared selection is possible to store for further use

Figure 14. Application for common Microdata base analysis



To ensure security of the Microdata user rights management subsystem has been developed.

#### D. Electronic Data Collection Module

Electronic survey data collection module is based on WEB data entry applications, including survey design and preparation, special data validation algorithms, automatic request sending, response check and management.

The core element for Electronic Data Collection Module is WEB based data entry and validation forms that is used for different surveys. These features are available to respondents and can replace ordinary paper questionnaires. Responses (completed forms) are transferred to the CSB through the Internet. Certain features of electronic surveys contribute to increasing of data quality data can be checked immediately.

The business and design goals for Electronic survey data collection module were:

- Design and preparation of electronic surveys in automated mode;
- Collecting questionnaire data in electronic form from respondents;
- To improve the quality of collected data by using on-line validation rules that were missing in paper questionnaires;
- Automatic request sending to respondents and automatic response control, implementation of reminders system to respondents;

- The system will remove unnecessary stage of retyping of information by CSB personal.

The Electronic data collection software module key features are:

- Statisticians use the same design tool for the design of paper questionnaires as well as for WEB forms
- Use of the Metadata provides universal approach for generation of the WEB based data entry forms, which can be done without participation of IT professionals (Figure 15);

Figure 15. Preparation of the WEB form – linking with Metadata base

The screenshot shows a Microsoft Word document with a context menu open. The menu items are: Respondenta nosaukums, Respondenta pasta adrese, Respondenta biroja adrese, Respondenta tālrunis, Respondenta fakss, Respondenta e-pasts, Respondenta NMK kods, Veidlapas aizpildītāja vārds, Veidlapas aizpildītāja uzvārds, Veidlapas aizpildītāja tālrunis, Veidlapas aizpildītāja e-pasts, Veidlapas aizpildīšanas periods, and Statistika e-pasts. The document content includes the title 'PĀRSKATS PAR RŪPNĀ 2008. G. mēneš' and contact information for the Central Statistical Bureau of Latvia. A table at the bottom lists the metadata fields and their corresponding values.

Respondents	
Nosaukums:	<Respondenta nosaukums>
Pasta adrese:	<Respondenta pasta adrese>
Biroja vai pamatdarbības vienības adrese:	<Respondenta biroja adrese>
Tālrunis:	<Respondenta tālrunis>
Fakss:	<Respondenta fakss>
E-pasts:	<Respondenta e-pasts>
Komercreģistra/Nodokļu maksātāju reģistra kods:	<Respondenta NMK kods>
Veidlapas aizpildītājs	
Vārds:	<Aizpildītāja vārds>

- Management system for WEB forms is created, including version control;
- Response control application allows defining automatic reminders/requests sending timetable and checking response;
- Software module ensures registration of respondents and defines detailed access rights for them;
- WEB forms offer following new features for the respondents:
  - o Pre-loaded data: respondent or survey specific data (e.g. respondent's name and address);
  - o Feedback data: historical data is available;
  - o Automatic validation: WEB forms could include validation rules;
  - o For periodical surveys in WEB based applications respondent have a possibility to see and prove previous periods data previous periods data;
  - o During data entry process in-form validation could be provided;

- Where necessary respondents are able to search and use classifications such as NACE, PRODCOM, etc;
- To see the list of surveys to be completed both electronically and on the paper (Figure 16) as well as information about the submission deadline ;

Figure 16. List of Respondents Surveys

The screenshot shows the 'e-Pārskats' web application interface. At the top, there is a navigation bar with buttons for 'Change password', 'Period', 'Help', 'Name of survey', 'Filter by year', 'Filter by periodicity', 'Status of questionnaire', and 'Deadline for submitting'. Below this is a header section with 'PĀRSKATI' and a filter section for 'Gads' (Year) with options from 2002 to 2008. The main content is a table of surveys with the following columns: 'Nr.p.k.', 'Kods', 'Periods', 'Nosaukums', 'Parskata iesniedzējs', 'Statuss', and 'Termiņš'. The table is divided into sections: 'Gada pārskati', 'Pusgada pārskati', and 'Ceturksņa pārskati'. Each row represents a survey with its details, including the start date and submission deadline.

Nr.p.k.	Kods	Periods	Nosaukums	Parskata iesniedzējs	Statuss	Termiņš
--- Gada pārskati ---						
1	2-EK	2008.	Pārskats par enerģētisko resursu iegādi un izlietošanu	'ALSVIKI' paju sabiedrība	sākts aizpildīt	
2	!PĀRSKATI	2008.	!IESNIEDZAMIE PĀRSKATI	'ALSVIKI' paju sabiedrība	sākts aizpildīt	25.12.2008
3	1-ls	2008.	Pārskats par lauksaimniecības produktu ražošanu un izlietojumu	'ALSVIKI' paju sabiedrība	sākts aizpildīt	12.07.2008
4	3-ls	2008.	Pārskats par lopkopību	'ALSVIKI' paju sabiedrība	sākts aizpildīt	
5	3-EK	2008.	Kurināmā atlikumi 1.oktobrī	'ALSVIKI' paju sabiedrība	sākts aizpildīt	31.12.2008
--- Pusgada pārskati ---						
6	1-irkj	2008.1.pusg.	Investīciju konjunktūras novērtējums rūpniecībā	'ALSVIKI' paju sabiedrība	tukšs	30.06.2008
7	1 - EK	2008.1.pusg.	Pārskats par enerģētisko resursu iegādi un izlietošanu	'ALSVIKI' paju sabiedrība	sākts aizpildīt	15.07.2008
8	1-tūrisms	2008.1.pusg.	Pārskats par tūrisma operatoru, aģentūru un komplekso tūrisma uzņēmumu darbu	'ALSVIKI' paju sabiedrība	sākts aizpildīt	15.07.2008
9	2-ibkj	2008.1.pusg.	Investīciju konjunktūras novērtējums būvniecībā	'ALSVIKI' paju sabiedrība	tukšs	30.06.2008
10	1 - EK	2008.2.pusg.	Pārskats par enerģētisko resursu iegādi un izlietošanu	'ALSVIKI' paju sabiedrība	sākts aizpildīt	15.01.2009
11	1-tūrisms	2008.2.pusg.	Pārskats par tūrisma operatoru, aģentūru un komplekso tūrisma uzņēmumu darbu	'ALSVIKI' paju sabiedrība	sākts aizpildīt	15.01.2009
12	2-ibkj	2008.2.pusg.	Investīciju konjunktūras novērtējums būvniecībā	'ALSVIKI' paju sabiedrība	tukšs	31.12.2008
13	1-irkj	2008.2.pusg.	Investīciju konjunktūras novērtējums rūpniecībā	'ALSVIKI' paju sabiedrība	tukšs	31.12.2008
--- Ceturksņa pārskati ---						
14	1-BA	2008.1.cet.	Pārskats par izdotajām būvatļaujām un objektu pieņemšanu ekspluatācijā	'ALSVIKI' paju sabiedrība	sākts aizpildīt	15.04.2008
15	2-investīcijas	2008.1.cet.	Pārskats par nefinansu investīcijām un būvniecību	'ALSVIKI' paju sabiedrība	sākts aizpildīt	15.04.2008
16	2l - ls	2008.1.cet.	Pārskats par lauksaimniecības produktu realizāciju	'ALSVIKI' paju sabiedrība	sākts aizpildīt	12.04.2008
17	1-IR	2008.1.cet.	Pārskats par ātrās ēdināšanas mēģņu pieņemšanu ekspluatācijā	'ALSVIKI' paju sabiedrība	sākts aizpildīt	15.04.2008

- The respondent is able to print questionnaires, for internal use;
  - Filling in of the web questionnaire is possible in both ways, manually and importing data from file
  - Help facilities.
- System provides means of security by using user access rights control and data transfer via HTTPS protocol;
  - CSB collects all survey data in the raw data base, ensure transfer of the data to the Micro data base to continue processing in the system.

## E. Software module for missed data imputation

Non-response in statistical surveys can never be totally prevented, but it can be considerably reduced. The optimisation of the survey design is one of the main activities for reducing the number of missing data. Nevertheless, more modern strategies to cope with missing data are based on use of modern and user-friendly software for missing data imputation.

In case of our system approach is as follows:

- Extraction of the data files with missing data;

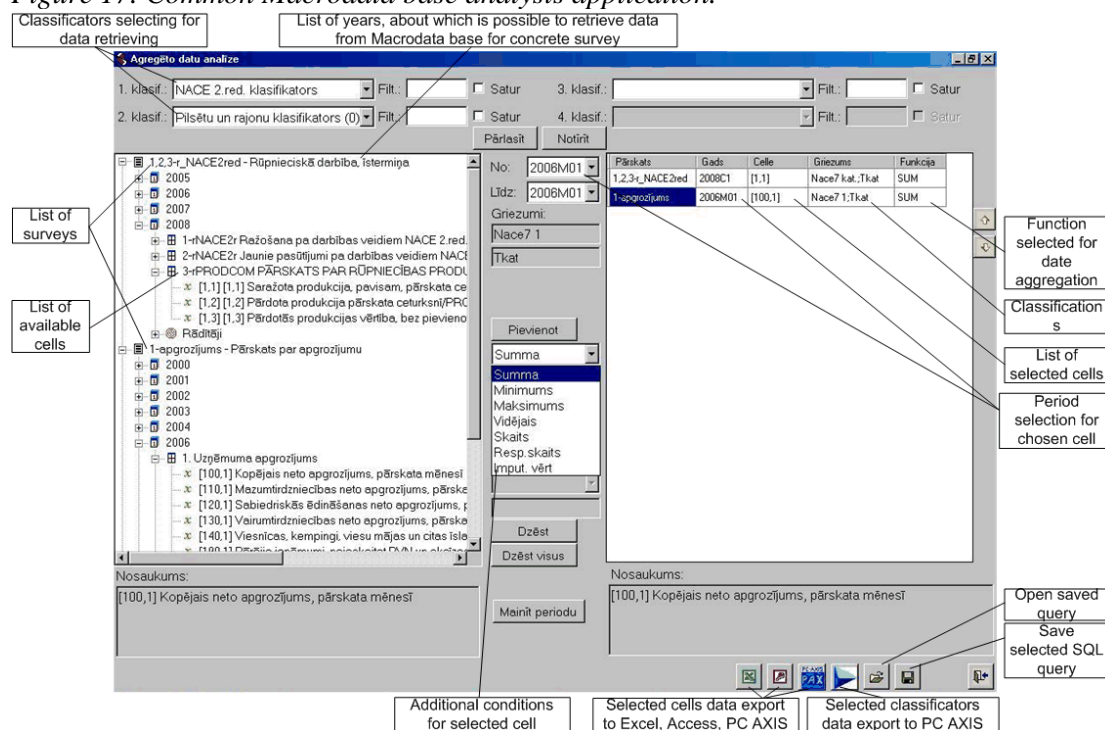
- Using of standard software packages for missing data imputation like WIDE, SOLAS or SPSS ensure imputation;
- Import of the data files/supporting Meta information back into Microdata base/Metadata base.

## F. Data aggregation module

As it was mentioned before from Data entry and validation module users can run application, which is used to maintain aggregated data. For each survey version for selected period is possible to store several aggregated data versions in the system. Using mentioned application users maintain aggregated data versions, i.e., create new versions or delete unnecessary versions of aggregated data. When users create new version of aggregated data, application analyze Metadata information about current survey version aggregation conditions and create data aggregation procedure, which after running store aggregated data in Macrodata base. If an exact survey for selected time period has several data aggregation versions, then one of them must be selected as active. Active aggregated data version is used in all applications, which works with aggregated data (reports creation, aggregated data analysis).

From Data entry and validation module users can also run application for common Macrodata base analysis (see figure 17). Using this application, users can extract from Macrodata base any data, which they need. Data selection starts from classifications selection. Then application display whole surveys list, which have aggregated data for selected classifications. For each survey it is possible to see years when aggregation has been done as well as to see cells list for current survey version. Using this application, users can select and combine aggregated data from different surveys for different time periods and to add additional selection conditions. All selected data will be exported to Microsoft Excel or Microsoft Access. Selected data also it is possible to convert to PC AXIS file format for dissemination purposes.

Figure 17. Common Macrodata base analysis application.



## G. Data analysis module

Data analysis module is based on On-Line Analytical Processing (OLAP) techniques. OLAP is a set of technologies that takes data in a data warehouse and transforms that data into multidimensional structures, called, cubes, to allow for better response to complex queries.

Data analyses module is realized using Microsoft SQL 2000 component – Analyses Service for multidimensional data cube formation and storing and include facility for multidimensional statistical data analysis (see figure 18).

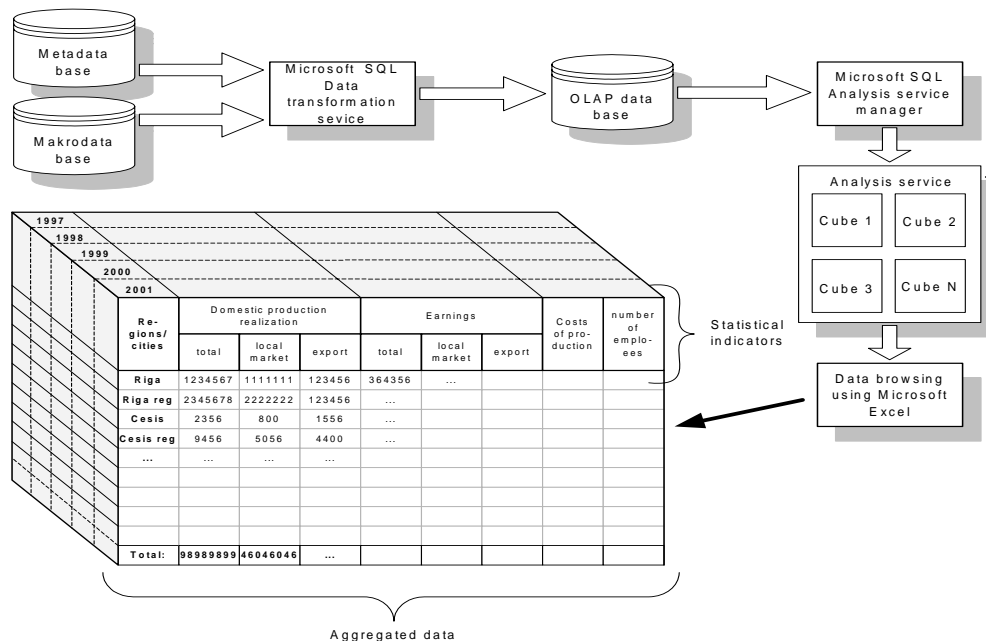
Work with Data analysis module we can divide in two main parts:

- Statistical surveys data preparation for analysis;
- Browsing and analysis of surveys data.

The first part, “statistical surveys data preparation for analysis”, is performed by data administrators, who are familiar with existing surveys and well introduced in Metadata base, Macrodata base data structures. This part is divided in two steps:

- Data transformation form Metadata base and Macrodata base and Microdata base in some special occasions using Microsoft Data transformation service to special data structure format, which is required for Microsoft SQL Analyses service;
- Formations of multidimensional cubes, which administrators are create with Microsoft Analyses service manager.

Figure 18. Data analysis module



The second part, “browsing and analysis of surveys data“, will be done by the system end-users using Microsoft Excel 2000 (or 2003) component – PivotTable, which allow easy to view, rearrange, regroup data in different ways. It is possible to use also other Microsoft Excel component – PivotChart, which allow creating diagrams from PivotTable data.

## H. I. User administration module

To provide easier way for the system users administration was designed and developed Users administration module. Module using consists of two parts:

- During the system development were created a lot of different applications with corresponding functions (functionality) inside them. For each application function system developers were created database roles using Microsoft SQL server 2000 administration tool - Enterprise manager which ensure that each application function can access to all necessary database objects (databases, tables, views, procedures, functions and so on) and can operate. In User administration module is included application containing function, which is foreseen for information registration of all system applications and corresponding functions. For each application function it is possible to describe, which database roles are necessary for this function correct operation. For each application is possible to specify one or more users, which will be current application administrators;
- When above described part is done, system application administrators can start to grant rights to users. Application administrators create new users in the system. They can create user groups and add users to these groups. The next applications administrators' duty is to prescribe for each user group application functions list with which concrete user group can work. Additionally for several application functions it is possible to set rights for concrete function in surveys level.
- In the latest version of the system there has been made significant improvements in the Administration module, which makes the system administration process much easier. For example, it is possible to assign the roles to user groups and to see system journal from applications and to get information about the latest survey data changes, who did them and when it was done.
- All others system users (which have not administrator rights) using administration applications can see each application functions list with which they have rights to work. They can change their own user password.

## VI. 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 personal, Administration,
- As the result of feasibility study we clearly understood, that all steps of statistical data processing for different surveys allows 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 system has been developed to ensure use of the standard statistical data processing software packages and other generalized software available in market,
- Within the process of the design and implementation of Metadata driven integrated statistical information system both parties statisticians and IT specialists should be involved from the very beginning,
- Clear division of the tasks and responsibilities between statisticians and IT personal is the key point to achieve successful implementation,
- Both parties have to have clear understanding of all statistical processes, which will be covered by the system, as well as Metadata meaning and role within the system from production and user sides,
- Initiative to move from classical stove-pipe production approach to process oriented have to come from statisticians side not from IT personal or administration,

- Motivation of the statisticians to move from existing to the new data processing environment is essential,
- Improvement of knowledge about Metadata is one of the most important tasks through out of the all process of the design and implementation phases of the project,
- 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,
- To achieve the best performance of the entire system it is important to organize the execution of the statistical processes in the right sequence,
- 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 administration, MS Analysis Service, other MS tools, PC AXIS family products and system Data Model, system applications.
- For the proper installation and functioning of the system is necessary to use workstations not lower than Pentium III 1GHz with RAM not less than 512 Mb equipped with OS MS Windows 2000 and MS Office not lower than 2000.
- At the time being (2008) we are moving from the data base platform MS SQL SERVER 2000- to version 2005 at the server side and from MS ACCESS to .net on the client side, thus getting independent from MS Office version changes.

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, confidentiality and security.

## References

- “An information systems architecture for national and international statistical organizations” prepared by Mr Bo Sundgren; Meeting on the Management of Statistical Information Technology (Geneva, Switzerland, 15-17 February 1999);
- “Terminology on Statistical Metadata”, Conference of European Statisticians, Statistical Standards and Studies No 53.
- “Guidelines for the Modelling of Statistical Data and Metadata” Conference of European Statisticians, Methodological Material;
- “Towards a New Statistics Netherlands”, blueprint for a process oriented organisational structure, prepared by Ad Willeboordse.