

Distr.
GENERAL

CES/SEM.47/17
13 March 2002

ENGLISH ONLY

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE
EUROPEAN COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint UNECE/Eurostat Seminar on Integrated Statistical
Information Systems and Related Matters (ISIS 2002)**

(17-19 April 2002, Geneva, Switzerland)

Topic III: Object-oriented technologies, component architecture

**DEVELOPMENT OF AN INTEGRATED STATISTICAL DATA
MANAGEMENT SYSTEM – THE LATVIAN EXPERIENCE**

Invited paper

Submitted by the Central Statistical Bureau of Latvia¹

I. INTRODUCTION

1. The aim of this report is to introduce participants to the new Integrated Statistical Data Management System (ISDMS) of the Central Statistical Bureau of Latvia (CSB), which is being developed as a prototype according to recent directions and international experience in statistical data processing.

2. From 1997 – 1999 experts of the Central Statistical Bureau of Latvia, in cooperation with PHARE experts contracted from PricewaterhouseCoopers, prepared a technical specification for the project “Modernization of the CSB – Data Management System”, which described all technical and functional requirements for the new system that uses statistical metadata as the key element in statistical data processing.

3. The main business and information technology improvement objectives that the CSB intends to achieve as a result of the project are the following:

- to increase the efficiency of the main process at CSB in the production of statistical information through the use of information technology;
- to increase the quality of the statistical information produced;
- to improve processes of statistical data analysis; and
- to modernize and increase the quality of data dissemination.

4. Considering the complexity of the project, it was decided to outsource the development and implementation of ISDMS to a company with a serious, long-term experience in the implementation of complex, large-scale and large-budget development projects. The information technology (IT) company FORTECH Ltd., which is one of the largest systems integrator in Latvia, was successful in the international public tender for this project.

¹ Prepared by Karlis Zeila, Vice President of the Central statistical bureau of Latvia, e-mail: kzeila@csb.lv

II. TECHNICAL PLATFORMS AND USED STANDARD SOFTWARE

5. The proposed system is in line with the CSB IT strategy. The existing computer and network infrastructure was utilized for system implementation.

6. In view of the complexity of the system and the necessity to process large amounts of data in short periods of time, the question of system performance became one of the main implementation success factors. To meet system performance and security requirements stated in the technical specifications of the project, CSB was equipped with all the necessary complementary IT infrastructure components for ISDMS – new high capacity high speed servers (Windows NT 4.0 operating system), upgrade of existing local area network at the CSB central office, which significantly increased the network throughput, wide area network including IP telephony option, firewall system and standard software licences.

7. CSB staff have already much experience in the use of the Microsoft family of products, which were purchased over the past years. Employees were trained how to use these products in their daily work. Microsoft products have one of the best price performance, which allow to reach high levels of integration and standardization. Considering the afore-mentioned arguments, it was decided that CSB should standardize on Microsoft products. This choice will help during ISDMS implementation and also to maintain and administrate the system by CSB staff after the warranty period, which should be provided by FORTECH Ltd for 3 years after system implementation.

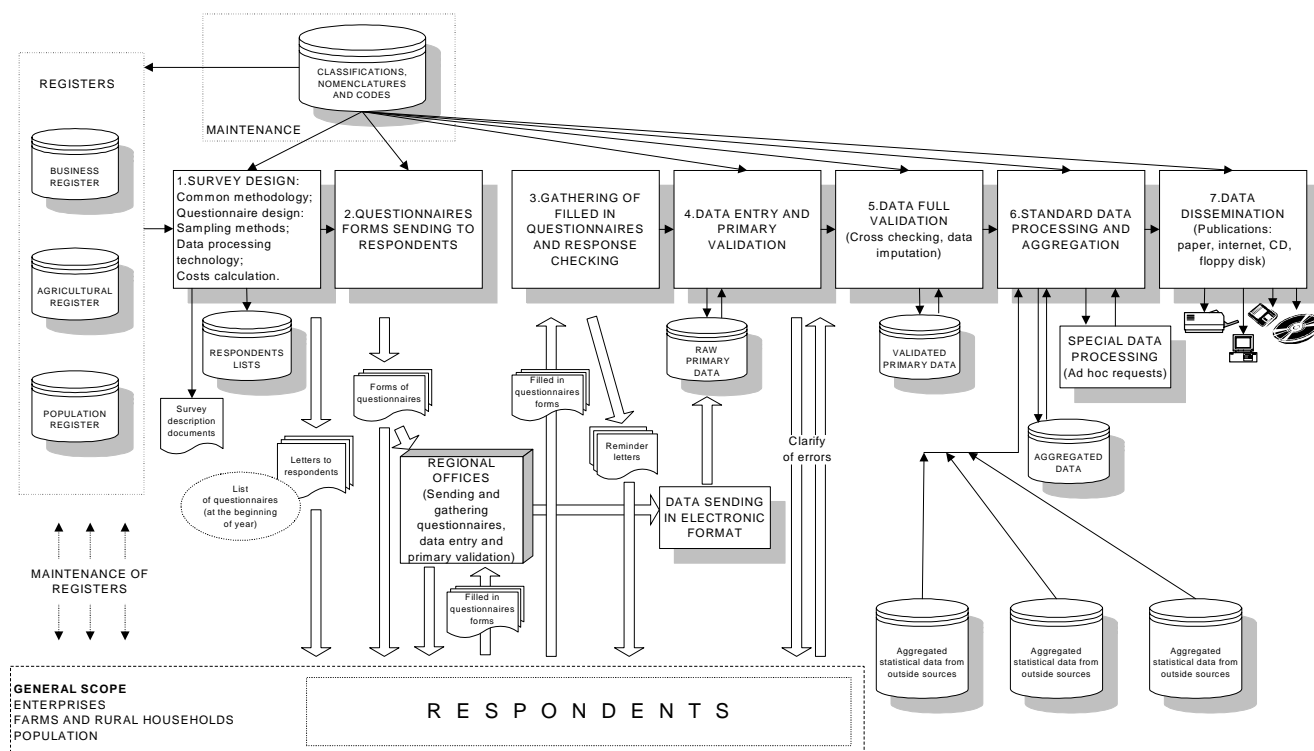
8. ISDMS databases are handled by the Microsoft SQL Server 2000. All applications comply with the client/server technology model, where data processing is performed mostly on the server side. Client software applications are developed using Microsoft Access 2000. Other components of Microsoft Office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

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

III. ISDMS ARCHITECTURE

10. Currently most of the statistical data processing takes place on personal computers using logically isolated business applications, which limit access to information by other CSB staff, complicate data processing, analysis, dissemination and increase the system development and maintenance costs. Technical incompatibilities exist as a consequence of the wide range of technology solutions used. A statistical data flow diagram of CSB is shown in figure 1.

Figure 1. Statistical data flow diagram



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

12. According to the European Union (EU) legal acts elaborated by EUROSTAT and related to statistics and the Statistical Requirements Compendium, the CSB of Latvia should regularly provide EUROSTAT with statistical data in such domains as macro-economic statistics, structural business statistics, short-term statistics, labour statistics and others.

13. The new ISDMS system contributes to the harmonization and standardization of all statistical indicators from different statistical surveys. This helps to meet and perform EUROSTAT’s requirements in the field of statistical data preparation and reporting to EU.

14. The new ISDMS is developed as a centralized system, where all data are stored in corporate data warehouse. The new ISDMS approach is to unite logically coherent items by using advanced IT tools to ensure the rationalization, standardization and integration of the statistical data processing processes.

15. An important task during the design of ISDMS was to foresee ways of including necessary interfaces for data exchange between already developed standard statistical data processing software packages and other generalized software available on the market. Such software functionality was irrational to recode and include as an ISDMS component, for example, statistical data processing software for missed data imputation, confidentiality checking, software for statistical data analysis and so on.

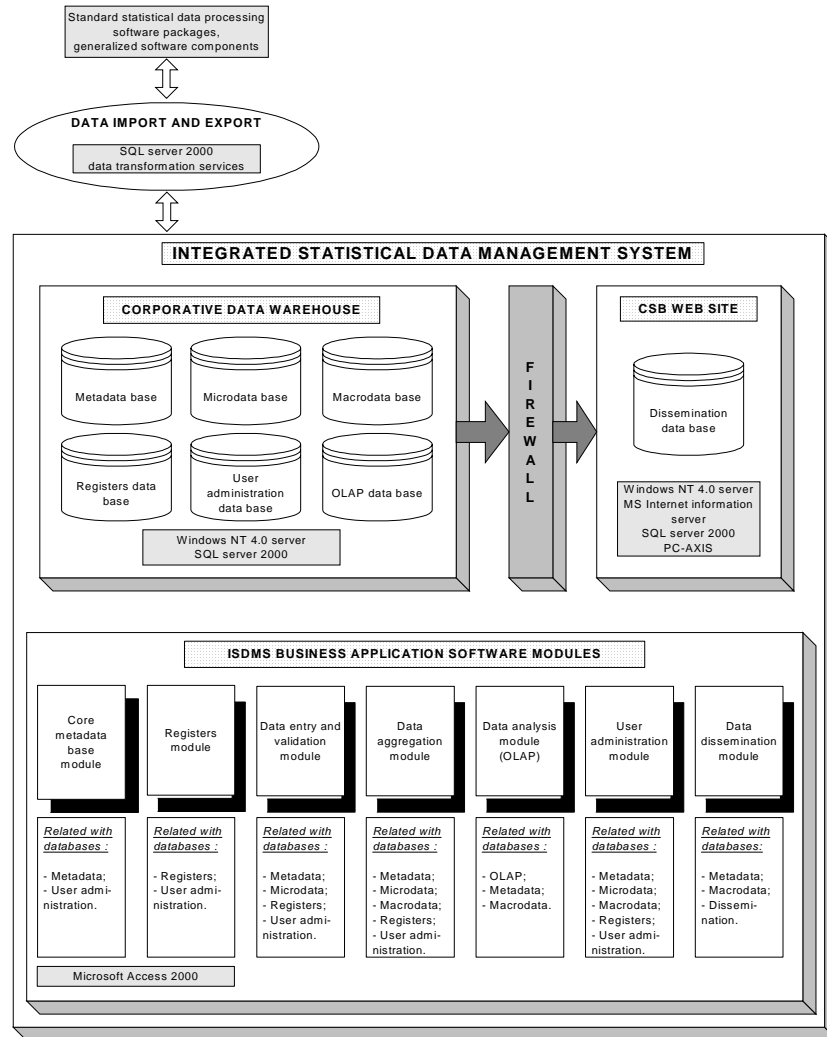
16. ISDMS is divided into the following business application software modules, which have to cover and to support all phases of the statistical data processing:

- Core metadata base module;
- Registers module;
- Data entry and validation module;
- Data aggregation module;

- Data analysis module;
- Data dissemination module;
- User administration module.

ISDMS architecture is represented in figure 2.

Figure 2. ISDMS architecture



17. In summing up the improvement goals and IT strategy to be realised in the ISDMS, there are mainly the following targets to be achieved by the ISDMS implementation:

- to increase the quality of data, processes and output;
- to achieve further integration instead of fragmentation on organisational and IT level;
- to reduce redundant activities, structures and technical solutions wherever integration can cause more effective results;
- to use statistical data more efficiently by using a common data warehouse;
- to provide the users (statistics users, statistics producers, statistics designers, statistics managers) with adequate, flexible applications at their specific work places;
- to replace tedious and time consuming tasks by value-added activities through the more effective use of the IT infrastructure;
- to use metadata as the general principle of data processing;
- to use electronic data distribution and dissemination;
- to make extensive use of a flexible database management to provide internal and external users with high performance, confidentiality and security.

A. Core metadata base module

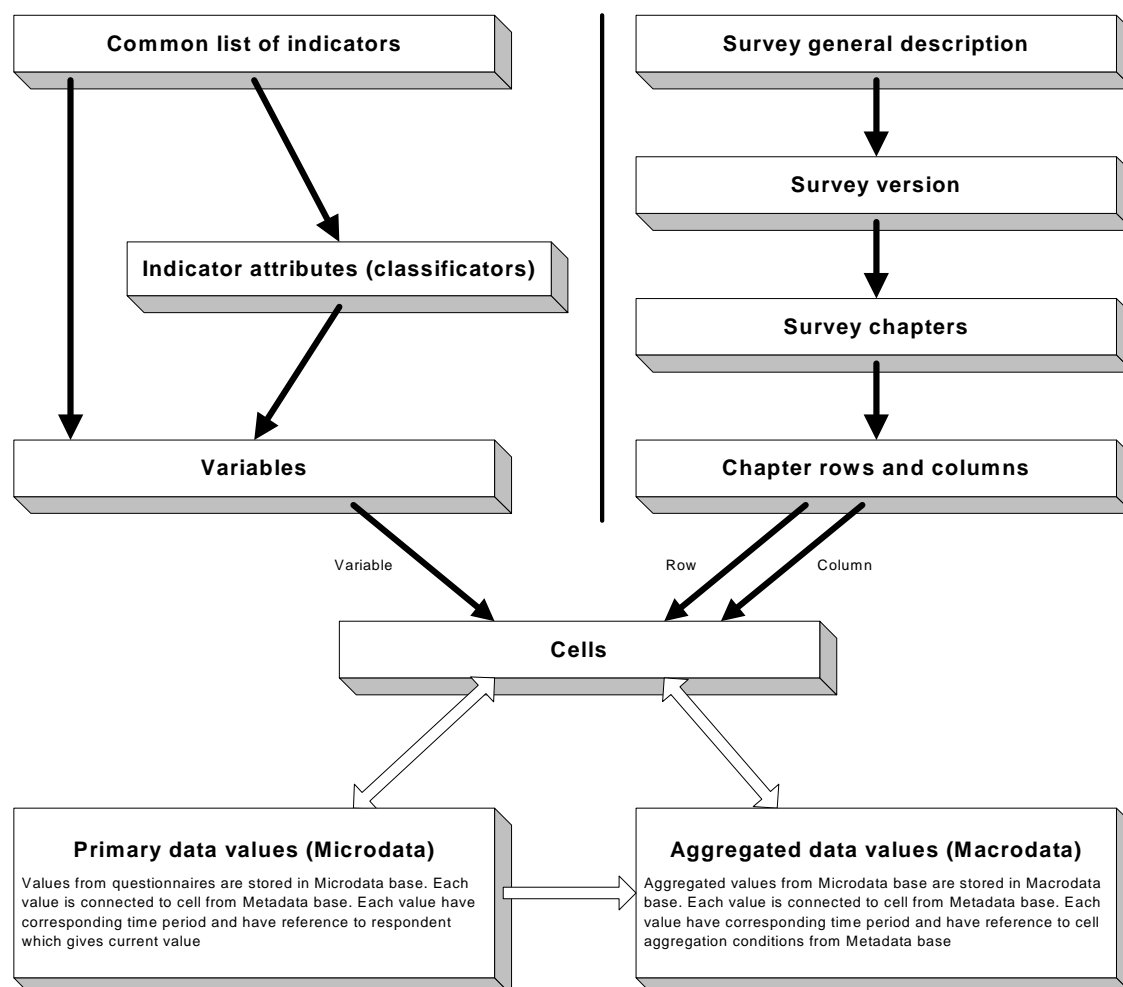
18. The Core metadata base module is one of the main parts of the new ISDMS and can be considered as the core of the system. The basic metadata handled by this module are used by all other modules of the ISDMS (except for the Registers module).

19. 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 metadatabase, 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 ISDMS metadatabase contain descriptions of statistical surveys themselves, their content and layout, description of validation, aggregation and reports and preparation rules.

20. The ISDMS ensures that the Metadatabase is used not only as a knowledge base for statisticians, but also as the key element for the creation of a universal, common, programming-free approach to processing data from different statistical surveys.

21. System users can easily query necessary data from Microdata / Macrodata bases navigating via the metadatabase. 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 ISDMS development. The metadatabase is linked at the level of database structure model with Microdata base and Macrodata base (see figure 3). A correct and carefully deliberate database structure model design is the basis for further successful system development and implementation.

Figure 3. Metadata base link with Microdata/Macrodata bases



22. Statistical survey data processing begins with survey metadata entry in the Metadata base. Each new survey should be registered in the ISDMS. For each survey, it is necessary to create a survey version, which is valid for at least one year, with concrete content and layout. If the survey content and/or layout does not change, then the current survey version and its description in the Metadata base is usable for next year.
23. Each statistical survey contains one or more data entry tables or chapters. In the Metadata base, it is necessary to describe for each chapter the type of the table, which can be a constant table with fixed number of rows and columns or a table with variable number of rows and columns.
24. For each survey version chapter the rows and columns are described 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.
25. In the survey version content description, information is saved about statistical indicators that are used in the current survey version. A list of common indicators is stored. 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.
26. For each indicator in the ISDMS, it is possible to define attributes – classifiers, which gives an opportunity to describe and store indicator values in more detailed division. Indicators do not always have attributes.
27. When the indicators and attributes have been defined, it is necessary to define the variables. Variables are combinations of indicators and corresponding attributes. Created variables are connected to the survey.
28. The last step in the survey content and layout description is the formation of cells. Cells are the smallest data units in survey data processing. Cells are created as combinations of rows and columns from the survey version side and variables from the indicators and attributes side.
29. All survey values from the questionnaires are stored in the Microdata base and each value has a relation to a cell (from the Metadata base), which describes value meaning. Also each value in the Microdata base has additional information about the respondent, which gives a current value and time period. The same situation occurs in the Macrodata base, where aggregated values are stored. Each aggregated value has a reference to a cell (from Metadata base), reference to each value aggregation condition (from Metadata base) and correspondent time period.
30. Such definitions of cells – combination from two sides – has several advantages. First of all, for statisticians who work with one survey every day and are familiar with the 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 the survey are described in the Metadata base using row and column language as well.
31. But, such an approach is useful only inside one survey version, which is valid for at least one year. Next year with the new survey version the same cell row and column coordinates are often changing. For data analysis such an approach is not useful because the value in one year is identified with one row and column coordinates, but in the next year with another one. In such situations users can use another approach, where it is possible to search cells values from the variables side. For variables the cell location (coordinates) inside questionnaire is not important. It is possible that the same cell, which is included in one survey version, is described with certain row and column coordinates and in other survey version has different row and column coordinates while the variable definition remains the same. For statisticians it is very useful to navigate through common indicators lists and find values without knowledge about surveys from where the data comes.
32. The metadatabase module contains following main applications:

- **general description of statistical survey:** for each new statistical survey requiring metadata input it is necessary to enter first of all basic information about this survey, e.g., name, index (code), periodicity. Application allows attaching general description methodological documents to survey.
- **description of survey version:** content and layout of statistical surveys are changing every year or once every few years. Each statistical survey with concrete content and layout, called survey version, has to be approved at the end of each year and is valid at least during the next year. The application purpose is to register for each survey version information about the survey version name, code, periodicity, information in which years survey version was valid, date of approval of survey version, tieback with State statistical programme chapters and survey division in chapters. It is possible to attach as well methodological document files, which belong to the survey version.
- **description of indicators and attributes of statistical survey:** for each survey version, statisticians select corresponding indicators. Indicators are characterized by name, periodicity, unit of measure, hierarchy i.e. indicator names, which are dependent on the attached indicator and indicator name. It is possible to describe a new indicator or select an existing indicator from a common list of indicators. The next task is to attach attributes to indicators. Attributes in ISDMS are always classifiers. There are cases, where indicators have not attributes. Attributes description is characterised by the name and attached corresponding classifier name. When indicators and their corresponding attributes are defined, application allows to generate variables automatically. The default variable name format is following –“indicator name / attribute name”. Statisticians can manually edit this default variable name if necessary. For each variable it is necessary to indicate whether variable will be used for aggregation or not.
- **description of content of statistical survey chapters:** here application statisticians describe rows and columns for each survey version chapter, i.e., allocate codes and descriptions. In the application main window also all previously created corresponding variables of survey are shown. When rows and columns for a chapter are described, each row and column combination should be connected with a variable to create a cell. For the cell value it is necessary to define the format.
- **maintenance of validation rules of statistical survey:** validation rules can be changed or supplemented for the survey version at any time. To make it possible for ISDMS users without involving in this process EDP personal (programmers), the validation rules maintenance application was developed. A special, easily understandable syntax for a description of validation rules using surveys version chapters rows and column codes was created. Each validation rule description contains validation rule code, error message text, description of the validation rule and validation rule conditions. When all necessary validation rules are described, it is possible to generate automatically the validation procedure, which is stored on the Microsoft SQL server. When statisticians need to make any changes in the validation rules, they simply change validation rules description in Metadata base using developed application. It is possible to execute validation procedure from Data entry and validation module applications.
- **description of aggregation conditions of statistical survey:** aggregation conditions are defined for each survey version cell or for a group of cells, which have the same aggregation conditions. Aggregation conditions describe how data from surveys are summarized. This means, that users must describe for each survey cell (or group of cells) which aggregation type data should be aggregated. In the metadata base it is possible to define three types of aggregation conditions. The first type is data aggregation using respondent parameters. Each respondent has such parameters as territory, NACE branch, and so on. So, if we describe that cell data values must be aggregated by respondent parameter “territory”, then respondents are grouped by territories and these respondents’ questionnaire cells data values are summarized. The second aggregation conditions type is data aggregation by indicator attributes. Such data aggregation type is used when cells are connected with indicators which attributes are international classifiers such as PRODCOM. In this case, cell values for each PRODCOM code are summarized from all questionnaires containing data about this concrete PRODCOM code. The third aggregation conditions type is data aggregation for cell (1) by cell (2) data value in the same questionnaire. It should be possible to classify cell (2) data values. Therefore, cell (2) data values can be used as classificatory for questionnaires cell (1) data values grouping and summarizing. Using aggregation conditions description it

is automatically possible to generate aggregation procedure, which is executed from Data entry and validation module applications.

- **grouping of classifier records:** the purpose of such application development was to ensure facility for ISDMS users to re-group, reorganize classifiers records into necessary hierarchical groups. Statisticians can create and store these classifier groupings in ISDMS as much as they want. Classifier groupings are used for the preparation of different reports. Such an approach facilitates production of reports. During the classifier grouping process none of the classifier records can be removed. It is also possible to remove an already defined classifier grouping from ISDMS only if no reports are produced based on this classifier grouping. For example, by using this application it is possible to make NACE classifier codes grouping in sections and subsections level. Another example – it is possible to make classifier grouping, where we include section C with sub-sections CA, CB NACE codes and section named “others”, where are included all other sections NACE codes (excluding section C NACE codes).

- **description of reports:** one of the last steps in statistical survey data processing before dissemination is the creation of reports for survey data summarization. Aggregated data values are used for the creation of reports and for their preparation, a description from the metadatabase is used.

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

- The first step is the description of the general part of the report – name and periodicity, which is used for data summarizing. In addition, it is necessary to define report grouping conditions – which classifiers and/or which user created classifier groupings will be used in the report. These grouping conditions specify how data will be grouped and represented in report rows. For each report it is possible to describe up to six grouping conditions, one of which can be grouping by time periods. If data grouping by time periods is specified in the report preparation description, then in each report row, data will be represented for a concrete time period, starting from the current time period data and continuing with the previous time period data in descending order. A time interval for data grouping by time periods depends on summary periodicity. For example, if the report periodicity is monthly, then the first report data row represents data for the current month, the next row represents the previous month's data and so on.
- The second step for the report preparation description is columns definition. Each column has a name and for each column we must define which survey cell data will be calculated in the current column. If necessary, it is possible to set up additional grouping conditions for each column, but these grouping conditions have limitations. If the column user has a defined data grouping by classifier, then it is possible to calculate and represent cell data only for one concrete classifier record for such column. Also, if data grouping by time period is defined for each column, then it is possible to calculate and represent cell data only for one fixed periodicity unit for such a column. For example, if we have a survey with a monthly periodicity, then we can create the report preparation description with a yearly 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 a parameter as an aggregation function, which is used for the calculation of column data values. It is possible to choose one of the standard aggregation functions like sum, count, minimal, maximal or average (by default it is sum) for each column.

34. When users create the report preparation description and define grouping conditions for the report and report columns, they can use only those classifiers, which are used in survey cells aggregation conditions. Calculated reports are stored in the macrodatabase. Using the report preparation description, it is possible to generate automatically the report data values calculation procedure, which is executed from the data entry and validation module applications.

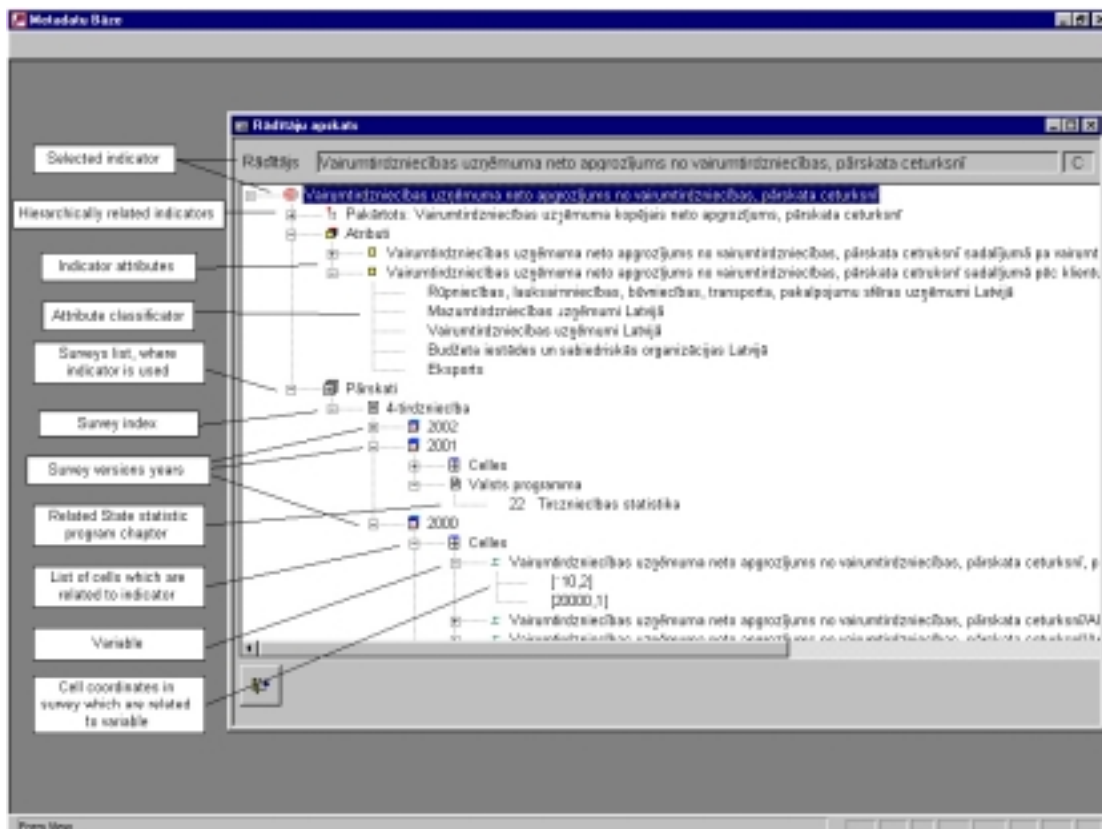
35. Using the above-mentioned metadatabase module applications it is possible to retrieve and view from the metadatabase information about one concrete survey version (an exception is the common list of indicators which is independent from surveys) at a time. To ensure the possibility of analysing metadata

information from the metadatabase for different surveys in one application, common metadatabase data browsing applications were developed.

36. One of the metadatabase browsing applications makes it possible to see how one chosen indicator relates with other metadatabase objects. The results are represented in tree form (see figure 4). This application represents the following information about one concrete statistical indicator:

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

Figure 4. Common basic metadata browsing application (based on selected indicator)

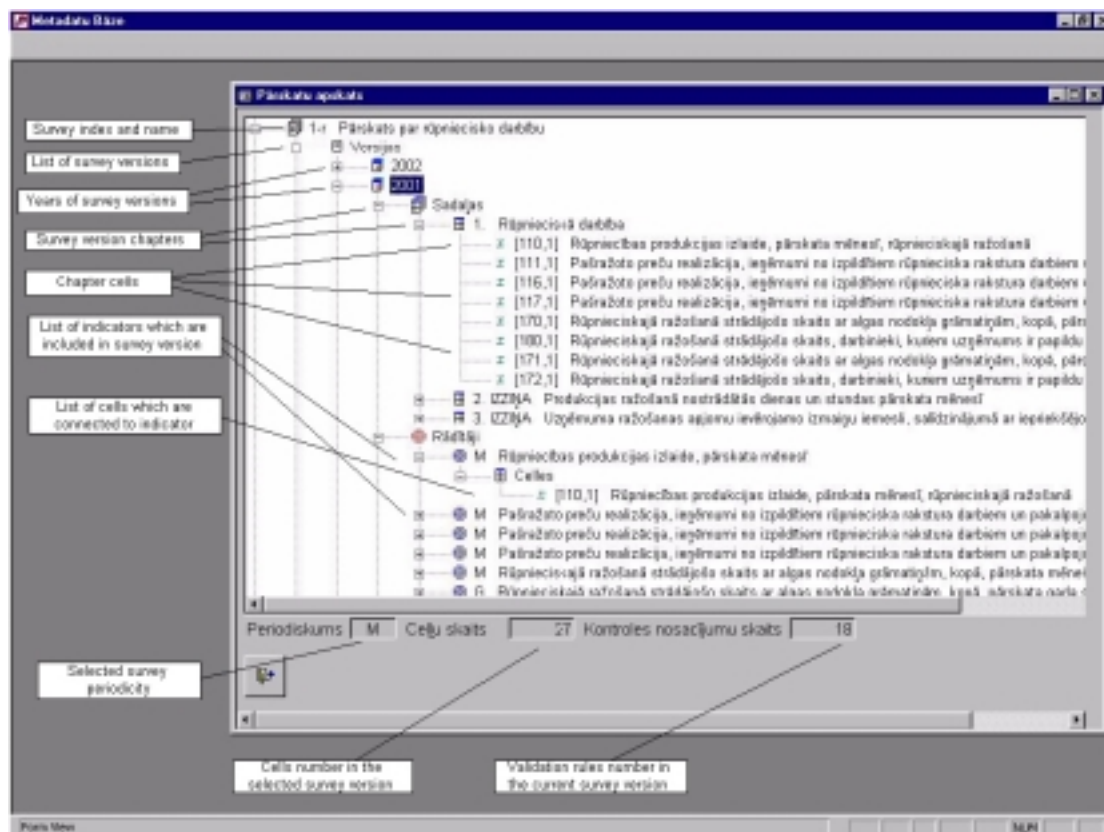


37. Another metadata browsing application makes it possible for users to see the survey content and the relationships with other metadatabase objects (see figure 5). Results are represented in the same way as in the previously described Metadata browsing application for indicators. The application shows users the following information:

- survey metadata browsing begins with the survey index and name;
- for each year application the survey version is shown which is related to the selected year. For the selected survey version, users can see how many chapters are defined;

- the application shows the list of cells which are created under each survey chapter. For each cell it possible to see cell coordinates (row and column) in the selected chapter as well as variable name which is connected to each cell;
- another tree branch shows all indicators which are used in the survey version;
- if indicators have attributes, the application shows them as well as classifiers, which are connected to the attributes;
- for each indicator, it is possible to see with which survey version cells the current indicator is related. For each cell, users can see cell coordinates (row, column) and connected variable name;
- for each survey version application the State statistical programme chapters to which current survey version belongs is shown;
- for each survey version it is possible to see methodological documentation files, which are connected to survey version;
- the application shows in three fields under the form selected survey version periodicity, number of cells and number of validation rules.

Figure 5. Common basic metadata browsing application (based on selected survey)



38. Some remarks about the implementation of other survey versions if we already have a defined survey version in ISDMS:

- variables (combination of indicator and attribute) are related to a concrete survey, not to a survey version. When statisticians implement other survey versions, they can use already defined variables from a previous survey version or create new ones for the new indicators. Variables, which store values can never be deleted. If in comparison with the previous survey version the new one does not contain the determined variable, then statisticians do not connect this variable to survey version chapter row and column combination, i.e., do not define the cell with such a variable;
- during implementation of other survey versions the ISDMS metadatabase module facilitates copying information on the chapter's rows and columns as well as the cell's description from the

previous survey version instead of rewriting this information for every new survey version. It is also possible to copy the description of validation rules from the previous survey versions.

39. The metadatabase module is operated by a specially established and trained group of people called the “Metadata group”, who has the right to perform metadata entry, actualization, changing and who is responsible for correctness of the metadata. It is very important that entered metadata is 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

40. Owing to the organizational structure of CSB, registers are stored in a separate database and managed by a specific business applications software module – the Registers module. The Registers module contains the following registers: Statistical Business Register, Statistical Agriculture Register and Statistical Population Register.

41. Due to the fact, that for the acceptance of ISDMS business applications software modules according to Technical specification requirements were selected 25 statistical surveys, which should be fully implemented. These surveys relate to business statistics. In this report we provide a short description of one component of the Registers module – the Statistical Business register.

42. Main tasks of the Statistical Business register are:

- preparation of respondents' yearly and quarterly samplings (catalogues), which are used as basis for creation of respondents lists for different business statistics surveys;
- preparation of different required information for CSB internal needs as well as for external users;
- to provide EU with required information according to EU methodology.

43. The Statistical Business register provide the following functions:

- **data import**; the main Statistical Business register data source is the State Business register. Approximately 95% of new enterprises in the Statistical Business register are coming from this data source. Once per month CSB receives fixed format files where information about new enterprises, liquidated enterprises and changes of existing enterprises is included. After data import all records are manually checked and approved. Once per quarter CSB receives full copy of the State Business Register and automatically compares these data with the Statistical Business Register. There are two more data sources – the State Revenue Service and Bank of Latvia, which are used for some information updating in the Statistical Business Register database and for extra control purposes as well. Data can also be imported from other external data sources.

- **data query and update**; there are several ways how we can query data form the Statistical Business Register database. First of all, it is possible to search information using any enterprise data field or combination of these fields, i.e., setting up filtering conditions (see figure 6). If we have complex data query conditions, users can write and execute SQL language scripts or to use Microsoft Access Query tool, which is integrated in Statistical Business register applications. Data entry and update application is shown in figure 7. Only some privileged users can delete enterprise information . If it is necessary to make the same changes for several enterprises, then users can select the group of enterprises and make these changes together for all selected enterprises.

Figure 6. Statistical Business register data searching application

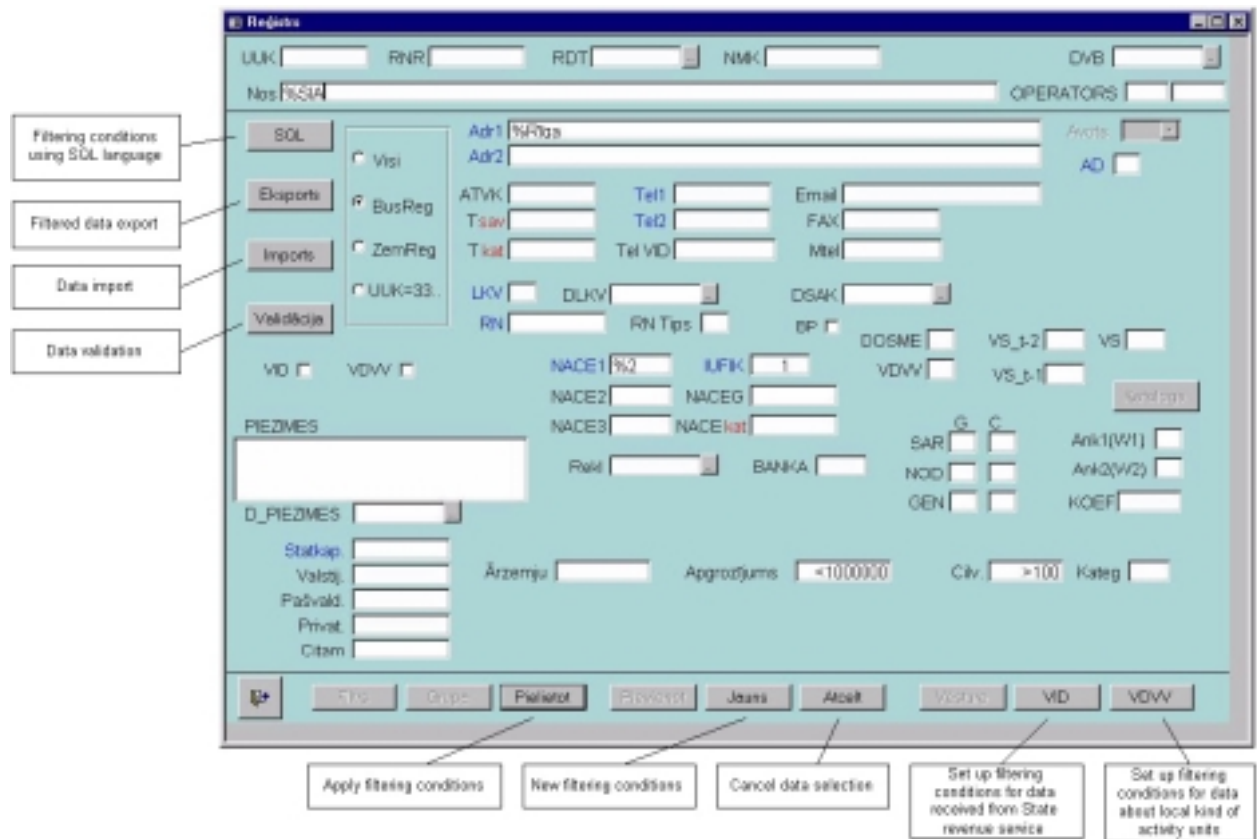
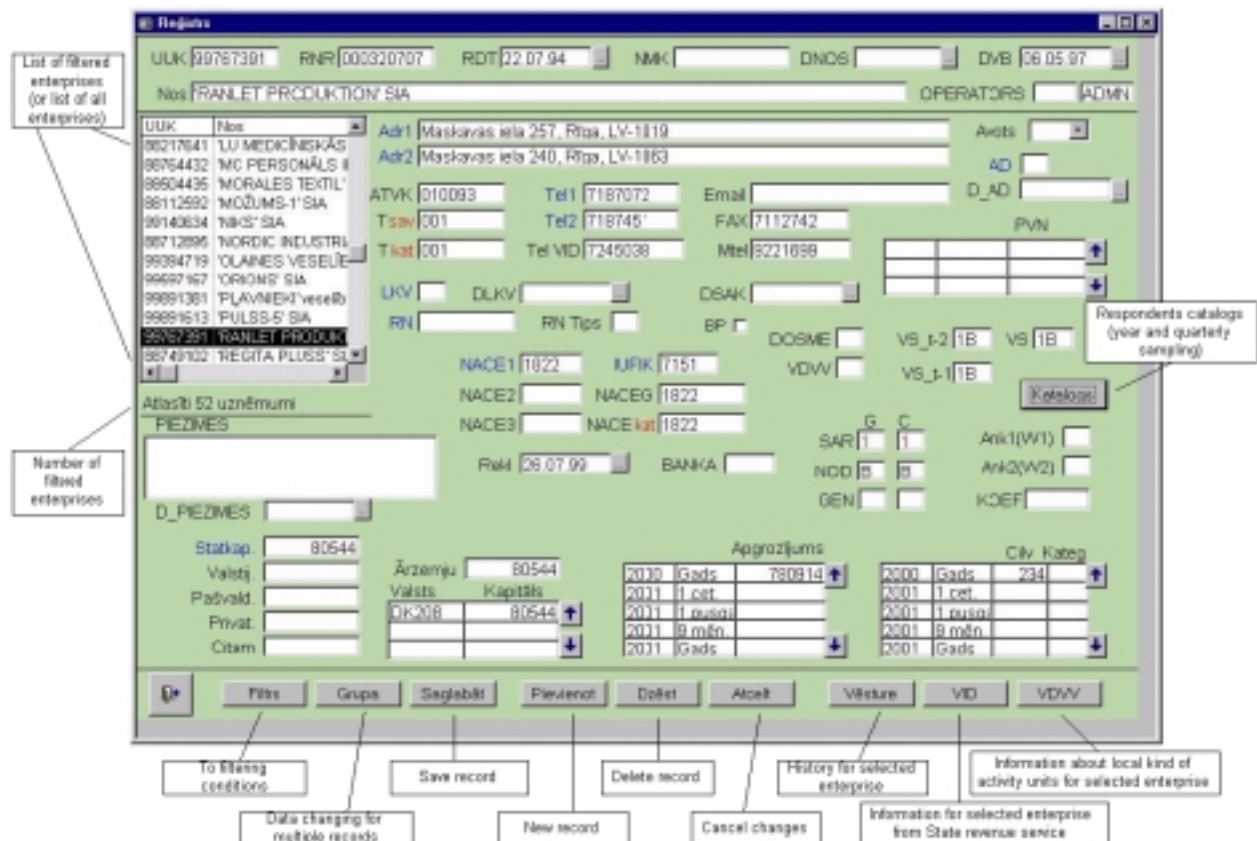


Figure 7. Statistical Business register data entry and update application



- **history of changes;** in the event of changes to enterprise data in the Statistical Business Register, historical information is saved automatically to ensure a possibility of tracking the history of changes for the enterprise. The main register fields (not all fields) are saved in history. When users save changes after updating enterprise information, the system asks if it is necessary to save these changes in history or not. This feature was included to avoid the situation, when unnecessary records are saved in history . For example, if the data input operator made a mistake during data entry and later corrected this mistake, it is not necessary to save such a record in history. If by mistake unnecessary records are saved in history, then privileged users can directly change or delete records in history.
- **data validation;** data validation can be performed for all Statistical Business Register data or just for filtered data. During the validation process the main data register fields (not all fields) are validated. Users can validate all these fields at the same time or can select one or more fields for validation. There are cases, when some validation errors should be temporally ignored. In such cases users mark such errors to ensure, that next time, when the validation procedure will be executed, these errors will be not displayed in the errors list.
- **local kind of activity units;** the enterprise can have structural units, which are not located in the enterprise address or are located in the same address, but perform a different economic activity. To resolve such situations, an appropriate application was developed. All changes in the local kind of activity units data are performed manually.
- **data export;** it is possible to export the Statistical Business Register data to several file formats – Microsoft Access database format (mdb), Microsoft Excel, dBase database format (dbf) and Paradox database format (db).
- **yearly and quarterly samplings.** The main output from the Statistical Business Register is respondents yearly and quarterly samplings (catalogues), which are used as the basis for the creation of respondents' lists for different statistical surveys. Catalogues are created using register query facilities, but it is not possible to write complex SQL query or compose this SQL query using Microsoft Access Query tool and receive as result respondents catalogue in one step. Usually it is an iterative process, where it is necessary to use various techniques and methods.

C. Data entry and validation module

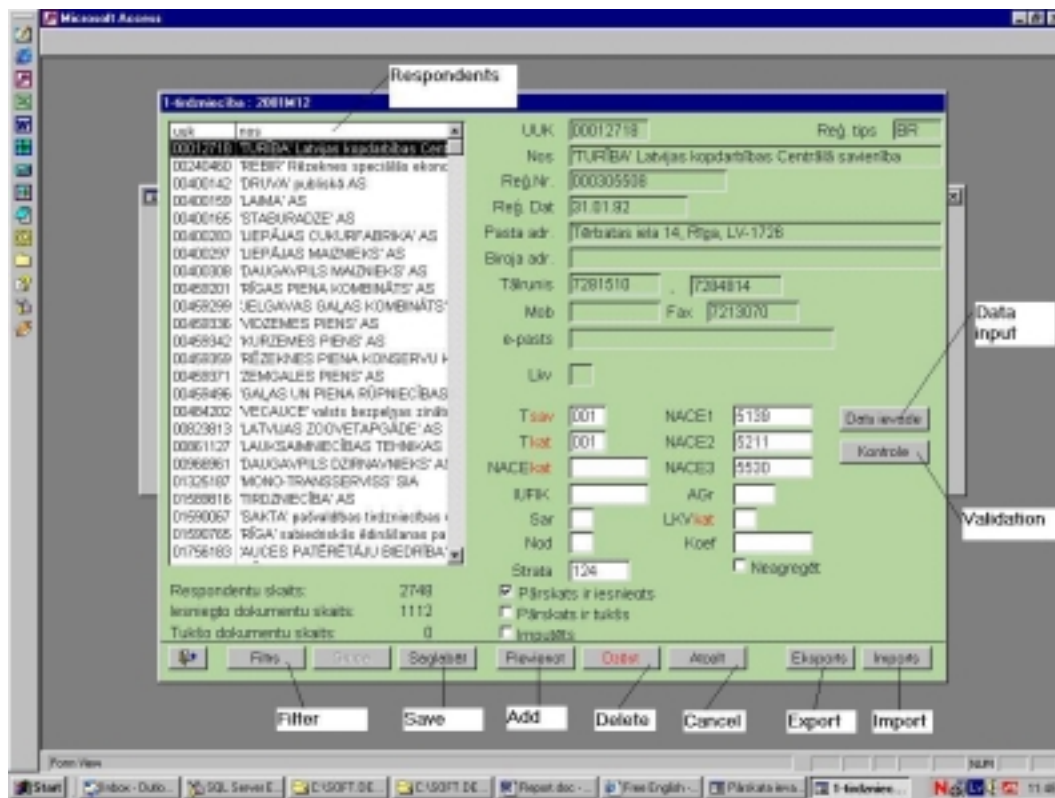
44. This module provides a standardized approach to processing data from different statistical surveys. Automated generation of data entry forms and of data validation, aggregation and reports creation procedures are implemented for data processing by using the metadatabase information. If it is necessary to make any changes to the survey content and/or layout, then it is necessary to change the survey description in the metadatabase. Such an approach makes system maintenance much easier and reduces costs. In case of changes it is not necessary to re-write programme codes, which can be done only by EDP professionals.

45. The following main functions are available for each selected survey and selected period (see figure 6):

- respondents list maintenance;
- data entry and validation;
- data aggregation;
- reports creation;
- data export/import.

46. ISDAVS ensure linking of the statistical survey to a particular list of respondents obtained from the Business Register. Each survey version for each period has its own list of respondents. The respondents maintenance application is shown in figure 8.

Figure 8. Respondents maintenance application.



47. The data entry and validation application is displayed in figure 9. The data entry screen shows selected respondent survey chapters using tab control and allows to select, enter, update and delete data. There are three types of survey chapters:

- with fixed number of rows and columns;
- with fixed number of rows and variable number of columns
- with fixed number of columns and variable number of rows.

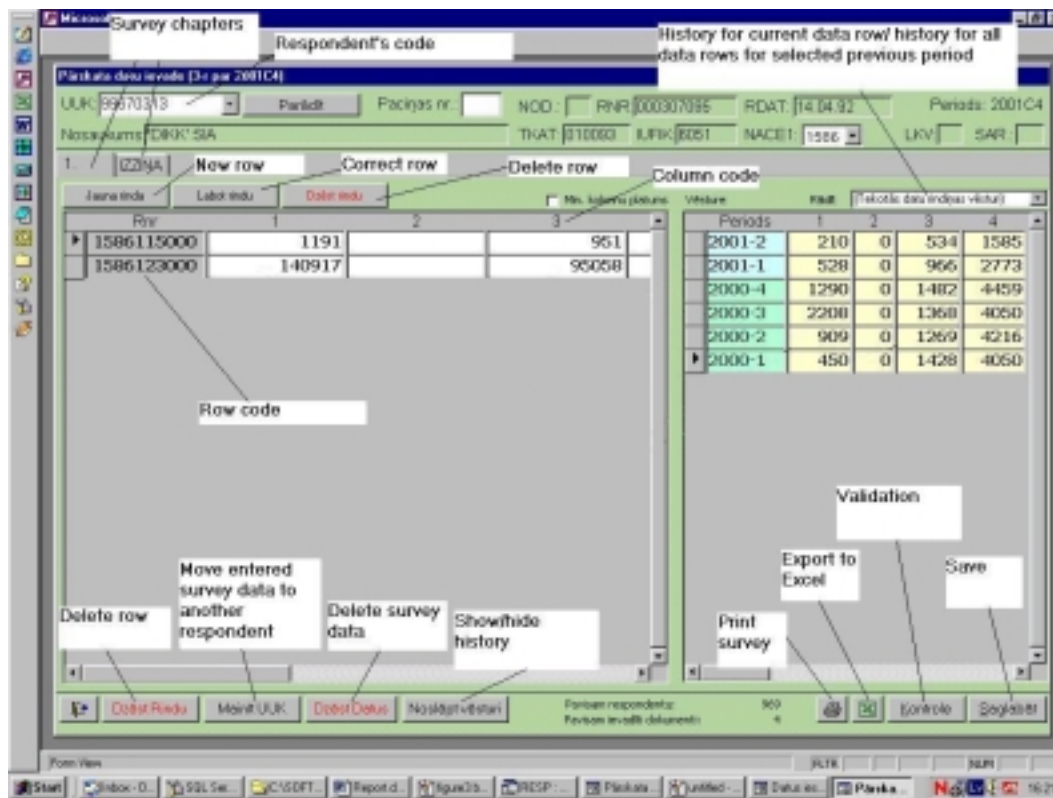
48. For example, if we have a survey with a chapter where there are several indicators in chapter rows and where indicator attributes are values from the NACE classifier in chapter columns, then we have a chapter with fixed number of rows and variable number of columns, because each enterprise can have a different number of economic activities, which are classified in the NACE classifier. The application allows to add or delete rows or columns manually.

49. Historical data (previous periods data) can be displayed in two ways:

- Historical data of all previous survey periods for currently selected row in data grid;
- Historical 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.

Figure 9. Data entry and validation application



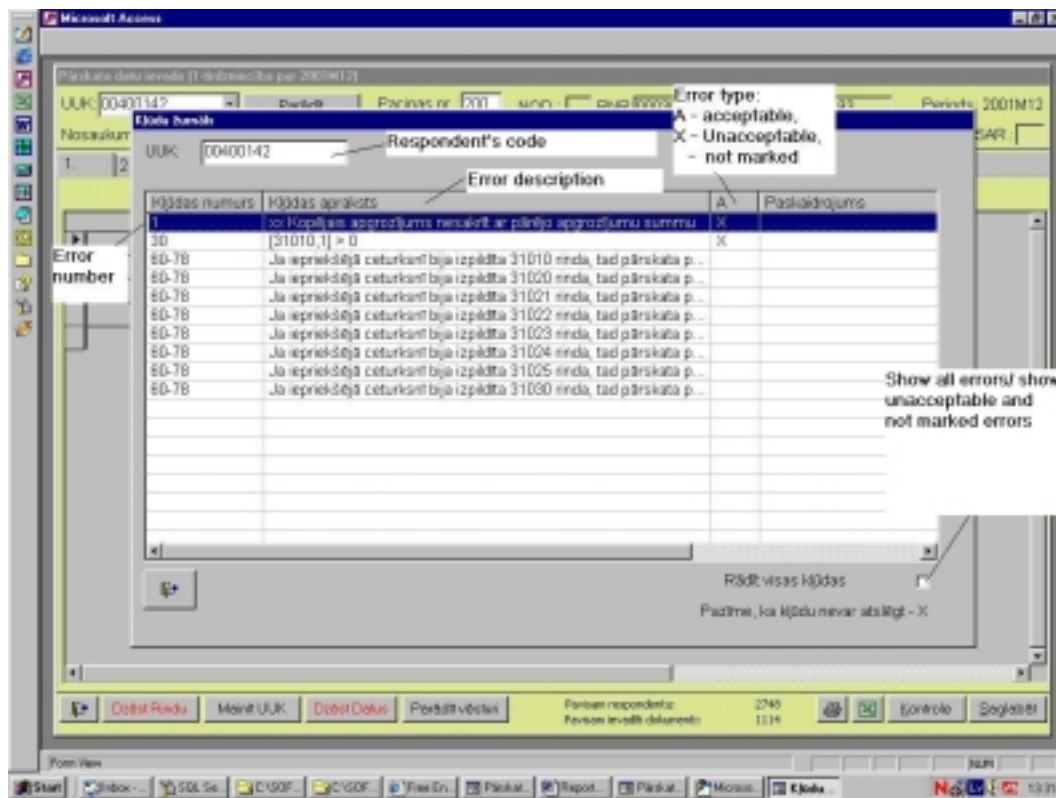
50. After executing the validation procedure of the respondent survey data (we can run this procedure from the data entry and validation application) a form with a list of validation errors opens. In this form, the following information is displayed for each error (see figure 10):

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

51. Errors marked with an “X” are critical and there is no way to pass them during the validation process. All other errors in the list of validation errors can be marked with the letter “A”. Errors marked with an “A” will not appear in the error list next time (to view in the errors list also the acceptable errors, the 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 the validation procedure for all respondents of one survey. Then we receive the list of respondents, which have validation errors in the survey’s data.

52. To define exceptions in the validation rules defined in the metadatabase, two additional applications were created, which interact with the data validation process. The first application allows to “switch off” one or more validation rules defined in the metadatabase for the survey. Then during the validation process such validations rules will be not checked. The second application allows to “switch off” one or more survey validation rules defined in the metadatabase for a concrete respondent and again during the validation process such validation rules for this respondent will not be checked.

Figure 10. Validation errors list



53. Survey data for a specified period can be exported to a fixed format Microsoft Access database file for external use. Also, a fixed format Microsoft Access database file can be imported into ISDMS. Other features which are included in the functions:

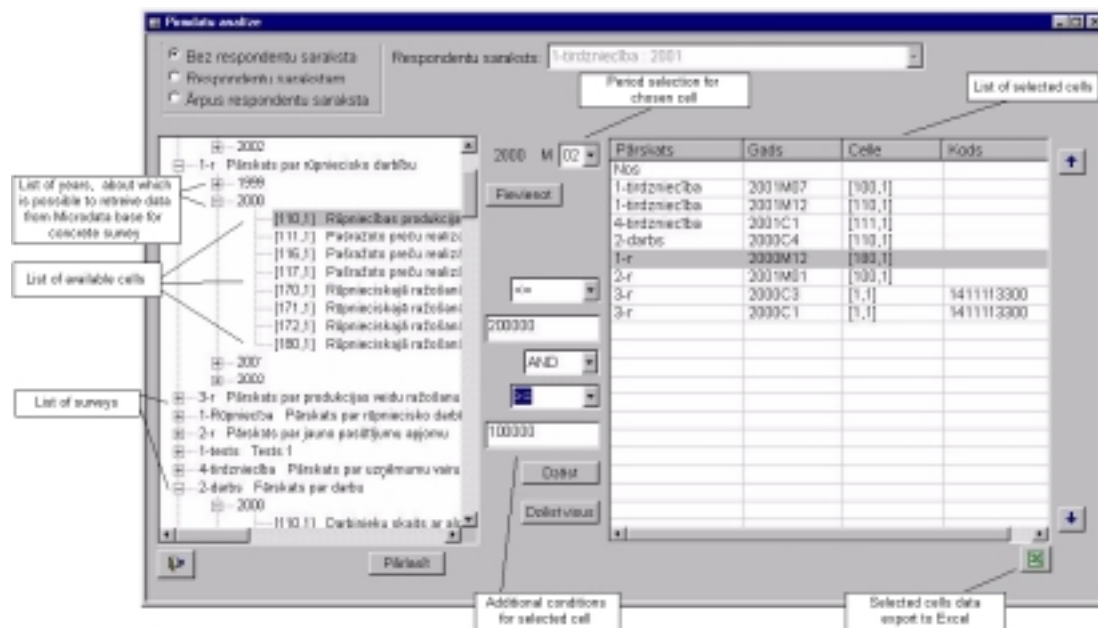
- ensuring the possibility for each department using module applications to work only with their set of respondents, in the application a facility to set the filter is included, which selects the necessary set of respondents for the department;
- for ensuring safety needs, a function for data locking was included. When it is necessary, the responsible statistician can lock the survey data, which prohibits further data changes. Before data can be changed, it is necessary to unlock the survey data;
- statisticians can create lists, which help to follow the data input process. The following are types of lists:
 - o list of respondents who have sent in the survey;
 - o list of respondents who have not sent in the survey;
 - o list of respondents, whose survey has one or more errors after validation.

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

55. Additionally in the data entry and validation module, an application for a common microdatabase analysis navigating via metadata is included (see figure 11). For end-users it is not necessary to know where and how the microdata are physically stored in the microdatabase.

56. Using this application, users can select and combine data from different surveys for different time periods and define different conditions for data, which should be queried. All selected data will be exported to Microsoft Excel.

Figure 11. Application for common Microdata base analysis

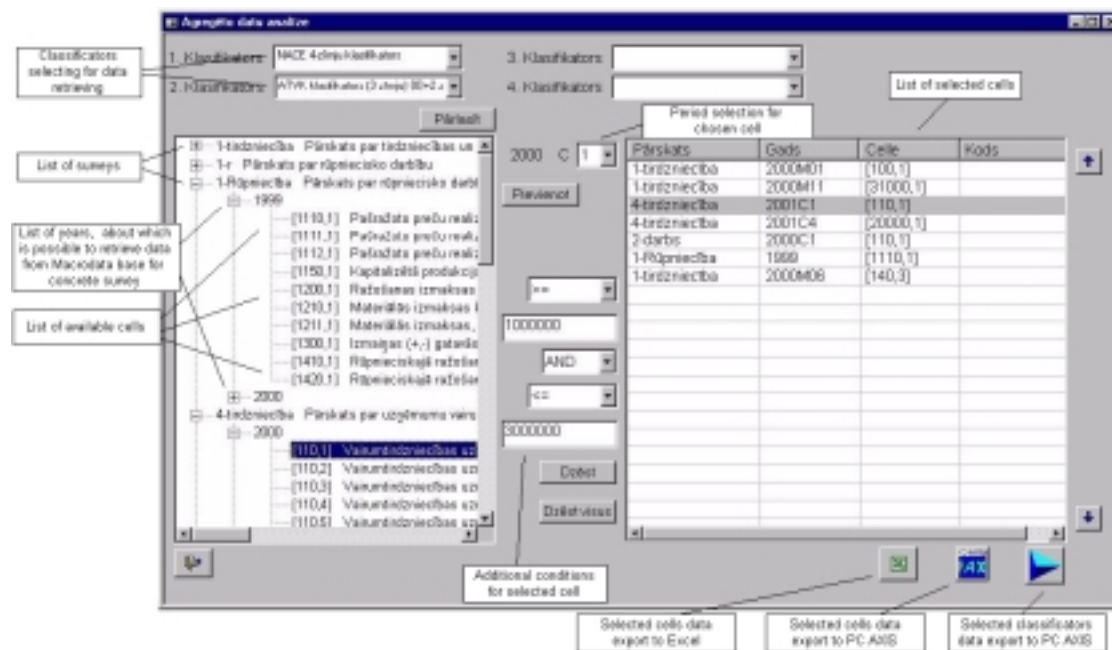


D. Data aggregation module

57. As was mentioned before, users can run the application from the data entry and validation module, which is used to maintain aggregated data. It is possible to store several aggregated data versions in ISDMS for each survey version for the selected period. In using this application, users can maintain aggregated data versions, i.e., create new versions or delete unnecessary versions of aggregated data. When users create a new version of aggregated data, the application analyzes the metadata information about the current survey version aggregation conditions and creates data aggregation procedure, which after running calculates data from Microdata base and stores calculated data in the macrodatabase. If the current survey for a selected time period has several data aggregation versions, then one of them must be selected as active. An active aggregated data version is used in all applications, which works with aggregated data (reports creation, aggregated data analysis).

58. From the data entry and validation module, users can run the application for the common macrodatabase analysis (see figure 12). Using this application, users can extract from the macrodatabase any data, which they need. Data selection starts by selecting the classifiers. Then the application displays the entire survey list, which has aggregated data for selected classifiers. For each survey it is possible to see the years for which data are calculated in the macrodatabase. For each year it is possible to see the list of cells for the current survey version. Using this application, users can select and combine aggregated data from different surveys for different time periods and to add additional selecting conditions. All selected data will be exported to Microsoft Excel. It is also possible to convert the selected data to a PC AXIS file format.

Figure 12. Common Macrodata base analysis application.

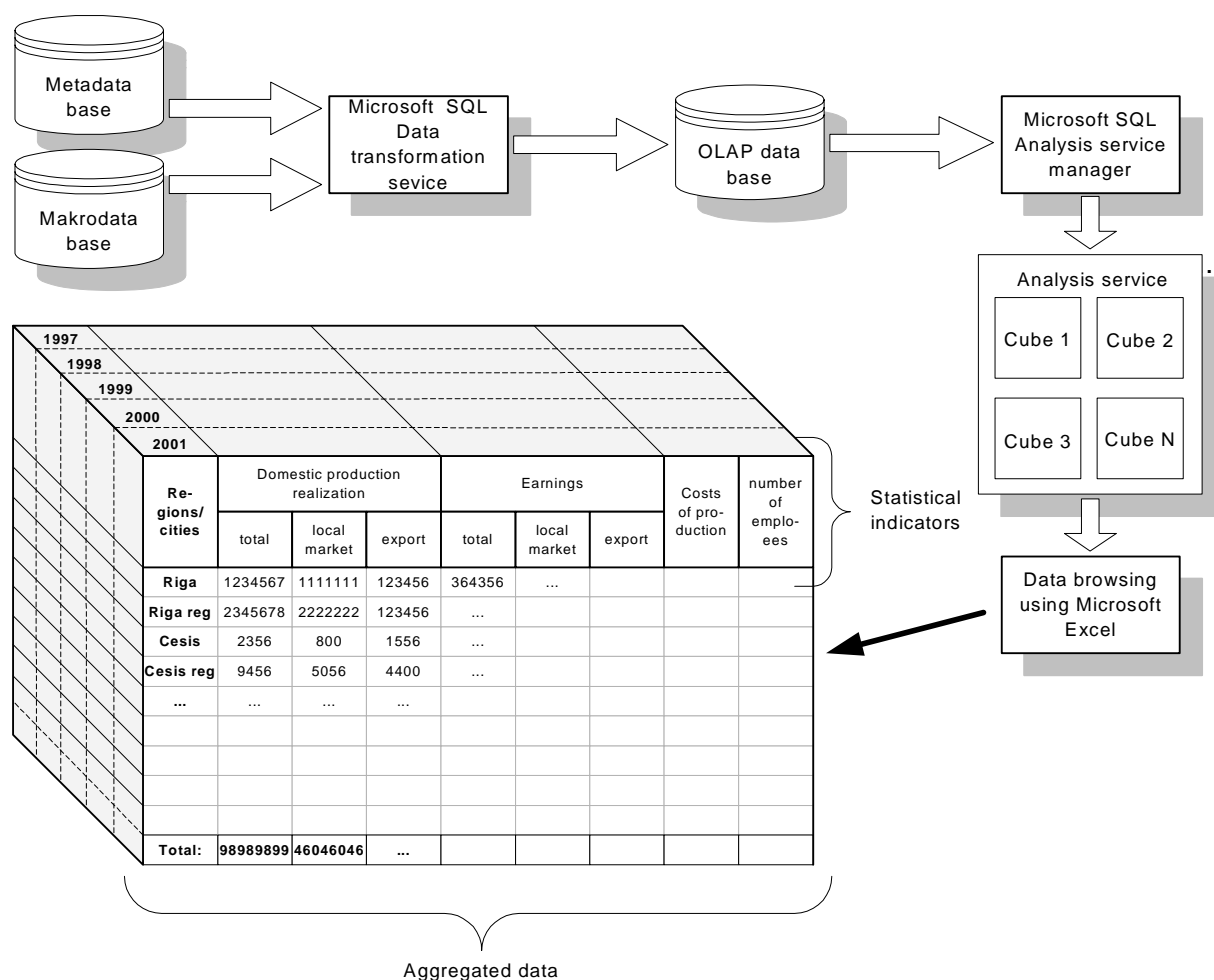


E. Data analysis module

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

60. The data analysis module is realized using Microsoft SQL 2000 component – Analyses Service for multidimensional data cube formation and storage and includes the facility for multi-dimensional statistical data analysis (see figure 13).

Figure 13. Data analysis module



61. We can divide the work with the data analysis module into two main parts:

- statistical survey data preparation for analysis;
- browsing and analysis of survey data.

62. The first part, “statistical survey data preparation for analysis”, is performed by data administrators, who are familiar with existing surveys and well introduced to the metadatabase and the macrodatabase data structures. This part is divided into two steps:

- data transformation form the metadatabase and the macrodatabase using the Microsoft Data transformation service to a special data structure format, which is required for the Microsoft SQL Analyses service;
- formation of multidimensional cubes, which administrators create with the Microsoft Analyses service manager.

63. The second part, “browsing and analysis of survey data“, will be done by ISDMS end-users using the Microsoft Excel 2000 component – PivotTable, which allows easily to view, rearrange and regroup data in different ways. It is possible to use also another Microsoft Excel component – PivotChart, which allows to create diagrams from PivotTable data.

F. Data dissemination module

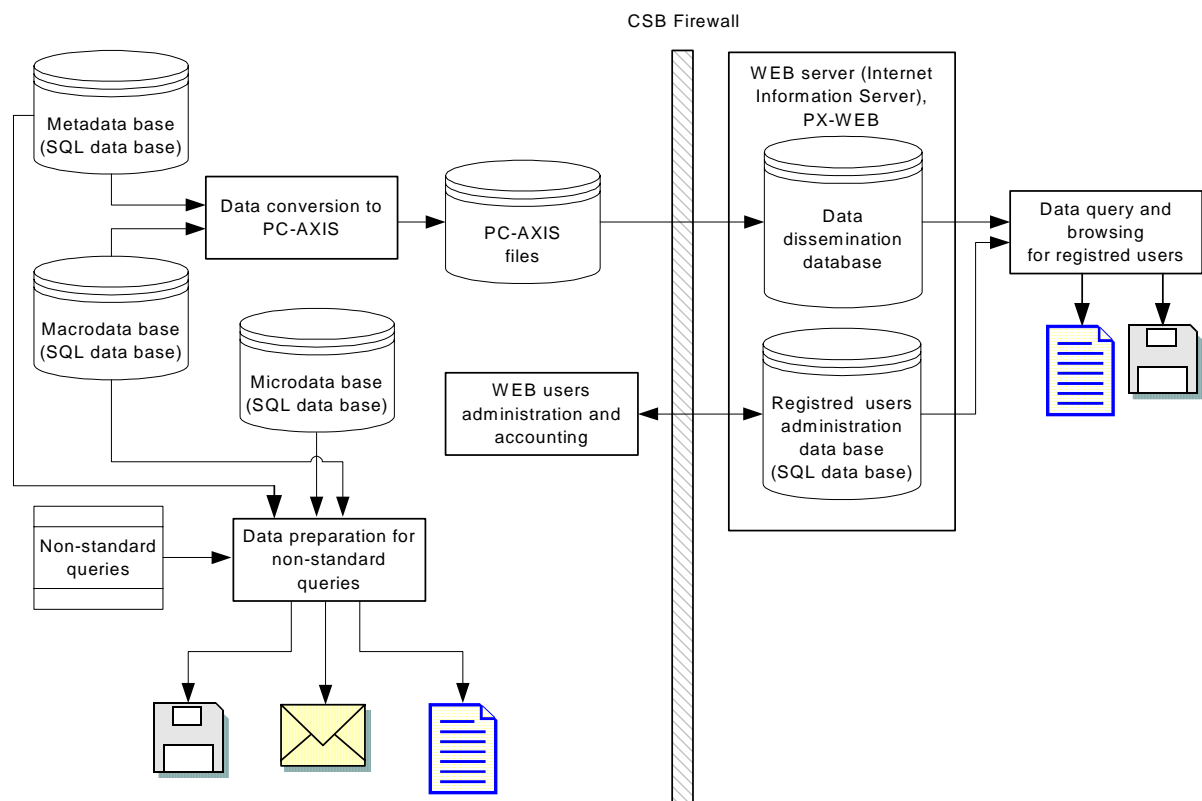
64. Following the guidelines for data publication at CSB, it was agreed that:

- publication of statistical data has to be performed using WEB environment;
- the publication tool should be PX-WEB from PC-AXIS family;

- publication should be performed from PC-AXIS files, not directly from ISDMS data bases;
- PC-AXIS files should be prepared by exporting data from the metadatabase / macrodatabase;
- data for publication cannot contain confidential information;
- data for publication are available only for registered users;
- data for publication are grouped by themes;
- users can subscribe access to themes for determined time period;
- the publication system should track users' actions.

The data dissemination module diagram is shown below (see figure 14).

Figure 14. Data dissemination module



65. Standard software PC-AXIS, developed by Statistics Sweden is used as the basis for the data dissemination module. The module provides the functions for metadata and macrodata conversion from the Microsoft SQL database (metadatabase and macrodatabase) format to PC-AXIS format.

66. Data for conversion to PC-AXIS format can be selected from aggregated data using the previously described application for aggregated data analysis (see figure 12) or can be selected from stored reports.

67. After converting the data, it is easy for the administrator to publish the data in the CSB WEB page using PC-AXIS family tool PX-WEB. Published data are not just static data tables – users can interactively select the necessary data by setting up different data queries, grouping and aggregation conditions.

68. The data dissemination module also contains applications for WEB users administration and accounting. These applications are developed using Microsoft Access 2000. WEB users administration applications cover the following functions:

- data entry and update of organizations;

- entry/update of WEB users data;
- WEB users sessions tracking;
- entry/update of organizations payments information for publications themes subscribing;
- maintenance of themes.

69. When WEB users are making a concrete theme data request, the system first checks if the theme is subscribed or not.

G. User administration module

70. The users administration module was designed and developed to provide an easier way for ISDMS users administration. The module consists of two parts:

- during ISDMS development, a lot of different applications were created with corresponding functions inside them. For each application function system developers created database roles using Microsoft SQL server 2000 administration tool - Enterprise manager, which ensures that each application function can access all necessary database objects (databases, tables, views, procedures, functions and so on) and can operate. In the user administration module a function is included, which registers information for all ISDAVS applications and corresponding functions. For each application function it is possible to describe, which database roles are necessary for this function to operate correctly. For each application it is possible to specify one or more users, who will be current application administrators;
- when the above part is done, ISDMS 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 application administrator's duty is to allocate each user group a list of application functions with which a concrete user group can work. Additionally it is possible to set rights for a concrete function in surveys for several application functions.

71. All others ISDAVS system users (who do not have administrator rights) using the administration applications can see each application functions list with which they have rights to work. They can change their own user password.

IV. ISDMS DEVELOPMENT AND IMPLEMENTATION SCHEDULE

72. Project "Modernisation of CSB – Data Management System" was started in November 1999. During year 2000 necessary IT infrastructure component parts for ISDMS, i.e., hardware and standard software was delivered, installed and configured. All necessary training was organized and performed as well. Design and development of ISDMS business application software modules took place during year 2001. From January 2002 implementation of the ISDMS was started. According to project scope, 25 selected surveys should be implemented to be the basis for ISDMS business applications software modules acceptance. These surveys were selected from business statistics field. ISDMS implementation is planned to finish by the end of June 2002.

V. VISIONS AND EXPECTATIONS FOR THE FUTURE

73. In accordance with the agreements, ISDMS development should be completed by 31 August 2002. Nevertheless, CSB plans to continue with further ISDMS amendments and the implementation of additional functions beyond the scope of project. The main areas of development will be:

- A software module for missing data imputation;
- Electronic survey data collection module;
- Software module for centralized mass data entry.

A. Software module for missing data imputation

74. Non-response in statistical surveys can never be totally prevented, but it can be considerably reduced. To achieve this, an optimal data collection design is necessary. The optimization 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 - modern and user-friendly software for missing data imputation - must be used widely.

75. The business application software module for missing data imputation should meet the following key requirements:

- software has to be compatible (interface) with the most widespread existing standard software tools used in the most advanced national statistical offices for missing data imputation;
- software has to be able to cope with several different imputation methods/techniques defined for each statistical survey. It has to be foreseen that different methods/techniques would be applied even for one particular survey;
- software has to be able to recognise each occurrence of non-response, and to divide them into:
 - o unit non-response (a whole unit has not provided the data);
 - o item non-response by type:
 - data missing by design, which means that the survey designer has decided not to collect data from particular unit in order to get practical advantages;
 - partial non-response, which means that all data is missing after a certain point;
 - item non-response, which means that there are gaps in the received data set for some units for some individual questions.
- software has to be able to work both as a total survey and stratified sample survey.
- software has to be able to cope with missing data imputation both on individual data level as well as on aggregated data level.
- software has to be able to cope with auxiliary information on causes of non-response and to provide possible solutions for reduction of non-response effect in survey data aggregates according to identified causes.
- possible imputation methods/techniques used in that application have to be based on internationally accepted methodology.
- software has to have interactive functionality in terms that application gives to statisticians a possibility to choose for each occurrence the most appropriate of proposed methods/techniques and within that – the most appropriate final variant for imputation.
- software has to have the function to compare survey results after performing imputation by different imputation methods/techniques and to provide reports on intervals of errors, variance estimations etc., as a data quality indicator in various imputation schemes.
- software has to provide a possibility for keeping comments on special cases used for missing data imputation.
- software has to be able to generate reports on finally applied imputation methods and on identifiable sources of information used for imputation (e.g., ID of donor unit, etc.) in order to archive them into metadata base.
- The report on the result of imputation in the particular statistical survey has to contain information on at least:
 - o the number of imputations in order to amortize unit non-response;
 - o the number of imputations in order to amortize partially imputed units values of imputation for each unit involved.

B. Electronic survey data collection module

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

77. Traditional methods of data collection currently used at CSB are based on paper questionnaires that can be completed by interviewers or respondents who receive a paper questionnaire via post and have

to complete and return it to CSB. Current CSB bottlenecks during processing of paper questionnaires received from respondents are:

- management of requests sent to respondents via post;
- initial analysis and validation of incoming paper forms data;
- data retyping into CSB Data management system;
- response control and management.

78. The core elements for an electronic collection system of statistical data are electronic questionnaires and WEB-based data entry and validation forms that are 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 the data quality as the data can be checked immediately.

79. The business and design goals for the electronic survey data collection module are:

- 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;
- sending automatic requests 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.

80. The key requirements to the electronic data collection software module are:

- Statisticians should be able to use one design tool for questionnaires, independent of the survey technology chosen.
- Design tool should support electronic questionnaires and WEB forms.
- Possibility to use metadata as the key element for creating a universal approach to the electronic survey system instead of the development of electronic questionnaire software specifically for a certain survey.
- A generalized solution that allows automated generation of electronic questionnaires and WEB-based data entry forms for surveys without participation of EDP professionals. In a simple, GUI supported way the statistician describes the survey and design tool generates electronic questionnaires and WEB forms.
- Modification of existing, stored electronic questionnaires and WEB forms layouts.
- Management system for electronic questionnaires and WEB forms should be created, including version control.
- Electronic questionnaires and WEB forms should offer the following features:
 - o Pre-loaded data: respondent or survey specific data (e.g. respondent's name and address).
 - o Feedback data: historical data.
 - o Auto-fill in fields: some fields can be automatically filled in, depending on values of previously completed fields.
 - o Auto calculation: columns or other fields can, for example, be summarised.
 - o Dynamic guidance: skip and hide questions not required, depending on previously entered values.
 - o Automatic validation: the electronic questionnaires and WEB forms should include validation rules.
 - o Help facilities.
- Security: it must be guaranteed that questionnaires and WEB forms data come from the responsible body and that confidential data are hidden to others.
- Automatic request be sent to respondents, including information about questionnaires, which respondent should complete. Information contains also necessary metadata about each questionnaire, which helps respondents to complete the necessary questionnaires. It is necessary to send requests via e-mail.
- Based on the received request, respondents have (after login) access to available WEB data entry forms or have the possibility to download electronic questionnaires forms.
- The software module has to ensure registration of respondents and to define detailed access rights for them.

- For periodical surveys in WEB-based applications, the respondent has to see the data from previous periods.
- During the data entry process, it is necessary to provide for both electronic questionnaires and WEB forms current in-form validation.
- With some surveys, respondents must be able to search and use classification codes such as NACE, PRODCOM, etc.
- The respondent should approve each filled and validated survey form with an electronic signature, when it will be legalized or use any other identification method.
- To control the correct data transfer, after sending the data, the respondent must receive a transmission confirmation.
- When CSB receives survey data, it must be checked (second validation) and if everything is correct, approval must be sent to the respondent.
- The response control system should check sent requests and if an answer is not received in the given time, to send automatically reminders to respondents. It is necessary to register the arrival of the responses in a database.
- When respondents log into the system, they must see information about the approved survey forms and information about reminders.
- According to CSB IT strategy, it was decided that CSB should standardize on Microsoft products. Therefore, the WEB interface should be compatible with Microsoft technologies currently presented on the market.
- The system should provide a high level of security by using user access rights control and information encryption/decryption. The three major considerations for effective IT security are confidentiality, authentication and integrity. A security strategy should be based on all three of these considerations.
- The respondent must be able to print questionnaires for internal use.

C. Software module for centralized mass data entry

81. At CSB, the current data entry process from paper form questionnaires is partly decentralized. One part of data entry is performed at regional offices and another part is performed centrally at CSB.

82. Taking into account the ongoing process of administrative reform in Latvia, it is clear, that it will be necessary to reorganize the structure of CSB regional offices. The number of regional offices will be reduced or removed completely and replaced by one central office. In any case, the need for a centralized data entry will grow and it is necessary to prepare facilities to process the huge amount of incoming data in the shortest time possible with the best quality. It is possible by creating a mass data entry centre at CSB. The special mass data entry software module should be created to fulfill necessary requirements.

83. The key requirements for the software module for establishing a data entry centre for mass data entry are:

- Automatic generation of data input forms based on the metadatabase.
- Programming-free recording of validation rules, which will permit on-line data validation.
- Data entry, validation and correction must be integrated into an interactive online business application.
- Applications for data entry must help to standardize data input activity for different surveys in the CSB.
- Applications should be user-friendly and as simple as possible for fast data entry.
- Incoming questionnaires registration. Identification of missing data using respondents lists.
- Logical division of data by individuals or other business units working in data entry center.
- Functionality of accounting entered data by individuals (as well as quality control of entered data by individuals) should be established.
- Data export to initial (raw) data database.

VI. CONCLUSIONS

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

85. The high level of flexibility of the ISDMS has been achieved using statistical metadata as the key element of the system. Any changes in the survey content and layout can be done without the participation of IT professionals and require only agreed changes in the metadatabase.

86. As the result of a feasibility study we have the clear understanding that all steps of statistical data processing for different surveys defy standardization. Each survey may require complementary functions (non-standard procedures) that are necessary for only this survey data processing.

87. To solve problems with the non-standard procedure interfaces for data exchange, ISDMS has been developed using standard statistical data processing software packages and other generalized software available in market.

88. It is necessary to establish and train special group of statisticians, who will maintain the metadatabase and who will be responsible for correctness of metadata. For the administration and maintenance of the ISDMS it is necessary to have well-trained IT staff, who are familiar with the MS SQL Server 2000 administration, MS Analysis Service, other MS tools, PC AXIS family products ,ISDMS Data Model and ISDMS applications.

89. The motivation of statisticians to move from existing to the new data processing environment is essential. Administrative restructuring could be used to move from stove-pipe data processing to process oriented data processing.

90. For the proper installation and functioning of ISDMS it 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.

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 Modeling 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.