Distr.
GENERAL

CES/AC.71/2004/11
10 March 2004

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE (ECE) CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint ECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS)
(Geneva, 17-19 May 2004)

Topic (ii): Development of IT strategies in statistical offices

METADATA DRIVEN INTEGRATED STATISTICAL DATA MANAGEMENT SYSTEM

Invited Paper

Submitted by the Central Statistical Bureau of Latvia[1]

I. INTRODUCTION

1. "The main goal in the Central Statistical Bureau (CSB) of Latvia is to use the achievements of modern information technology solutions to create a modern, effectively working IT infrastructure which provides convenient, fast and safe collection, processing, storage, analysis and dissemination of statistical data taking into account the requirements of the information society.

2. In the development of new data processing applications or subsystems, use will be made of metadata, as well as key management elements thereby ensuring an optimal end-user-friendly working environment where there will be no need to reprogram the processes in the event of recurrent changes. A gradual transition towards the development of a process-oriented data processing system at the CSB by centralizing data collection and entry procedures will be undertaken."

3. Those could be the special extracts from the IT Strategy of CSB of Latvia, which are now implemented within the working system described below.

4. The aim of this paper is to present the experience gained in the development and implementation of the new generation statistical data processing system, which integrates several subsystems, and is really metadata driven. This report is, to a certain extent, a follow-up to the invited paper on "Development of an Integrated Statistical Data Management System – the Latvian Experience" presented at UNECE ISIS 2002, held in Geneva in April 2002. Therefore, it does not include functional descriptions of the software modules already presented in the previous paper, but pays more attention to the metadata structure and the system upgrades.

[1] Prepared by Karlis Zeila (karlis.zeila@csb.gov.lv).

5. The project timetable was the following:

- From 1997 – 1999 CSB experts, in cooperation with experts contracted from PricewaterhouseCoopers, prepared Technical Requirements for the project "Modernization of CSB – Data Management System". The technical specifications represent all the technical and functional requirements for the new system where statistical metadata are used as the key element in statistical data processing.
- Development began in December 1999.

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

- Using modern IT solutions to:
  - 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;
  - Modernize and increase the quality of data dissemination;
  - Avoid hard code programming via standardization of procedures and use of metadata within the statistical data processing.
    - Development completed August 2002;
    - Implementation started May 2002;
    - To date within the new Metadata Driven Integrated Statistical Data Processing and Dissemination System (further in the text System) 59 Business Statistics surveys have been implemented;
    - System has been presented:
      - At ISIS 2002, April 2002, Geneva
      - Theoretical presentation on joint Seminar of experts from Bulgaria, Romania and Latvia in August 2002, Slivek, Bulgaria
      - Live demonstration on Joint seminar of experts from Bulgaria, Croatia and Latvia in September 2002, Riga, Latvia
      - Joint seminar of experts from Estonia, Lithuania and Latvia in October 2002, Riga, Latvia
      - Joint seminar of experts from Poland and Latvia in November 2002, Riga, Latvia
      - Joint seminar of experts from Cyprus and Latvia in May 2003, Riga, Latvia
      - Joint seminar of experts from Moldova and Latvia in June 2003, Riga, Latvia
      - Joint seminar of experts from Ireland and Latvia in September 2003,Riga, Latvia
      - METANET Project Meeting , Samos, Greece, May 2003.
      - AMRADS Final Conference, Roma, Italy, November 2003.

II. TECHNICAL PLATFORMS AND STANDARD SOFTWARE USED

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

8. The Microsoft SQL Server 2000 handles the system databases. 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 in Statistics Netherlands.

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

10. To date the system operates on the servers connected in an asymmetric cluster consisting of:

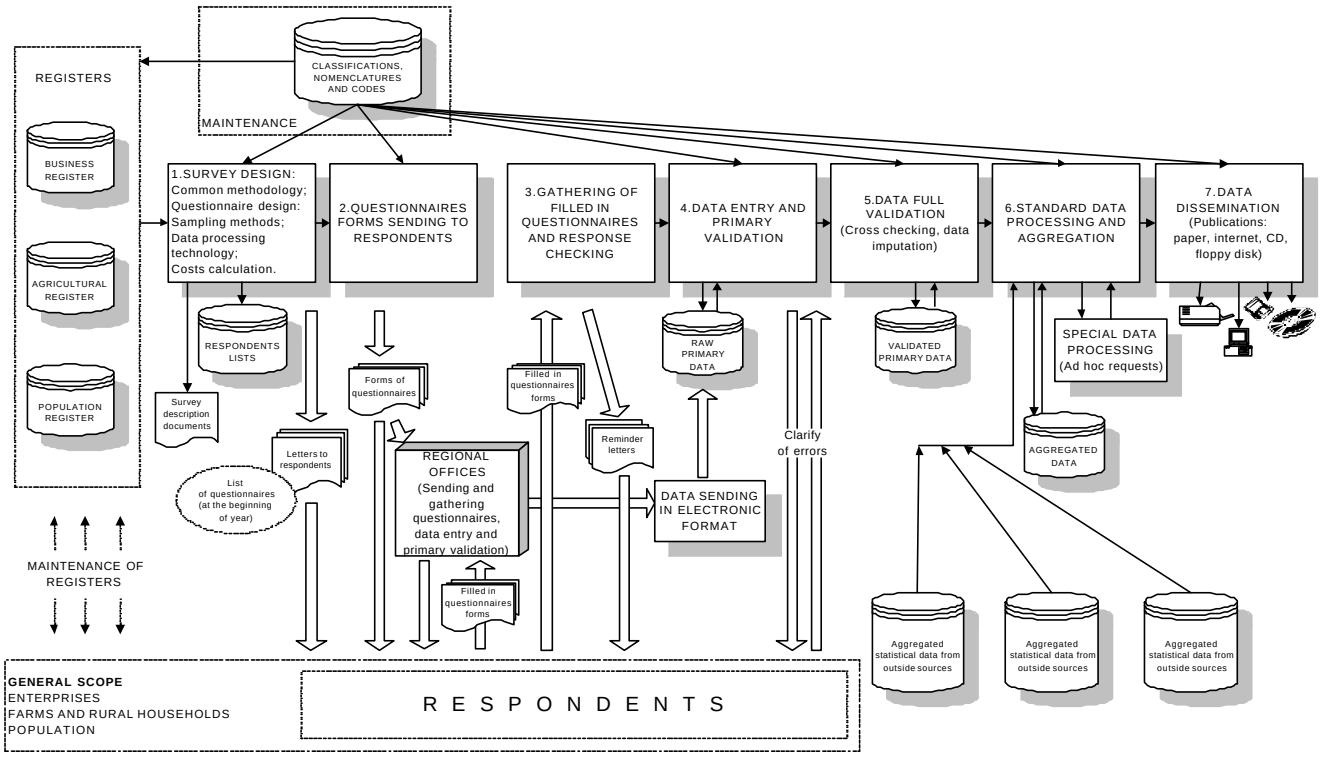
- ProLiant DL760 with 4 (Pentium III Xeon 900 MHz) processors and 8 Gb RAM;
- ProLiant 8000 with 2 processors (Pentium III Xeon 550 MHz) and 8 Gb RAM;
- Storage Works RAID Array 4000 with 8 disks of 36 Gb, 10 000 rpm.

11. The cluster system with workstations are connected within LAN Fast Ethernet 1Gbit base 100Mbit per sec. The types of workstations used are Pentiums II to IV with a required RAM not less than 128 Mb equipped with OS MS Windows 95 (better MS W-2000) and MS Office 2000. To date about 200 users in the CSB central office work with the system and more than 40 remote workstations in regional data collection and processing centres are connected on-line.

II. SYSTEM ARCHITECTURE

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

Figure 1. Standardized statistical data flow diagram



13. The theoretical basis for the system architecture was taken from the 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.

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

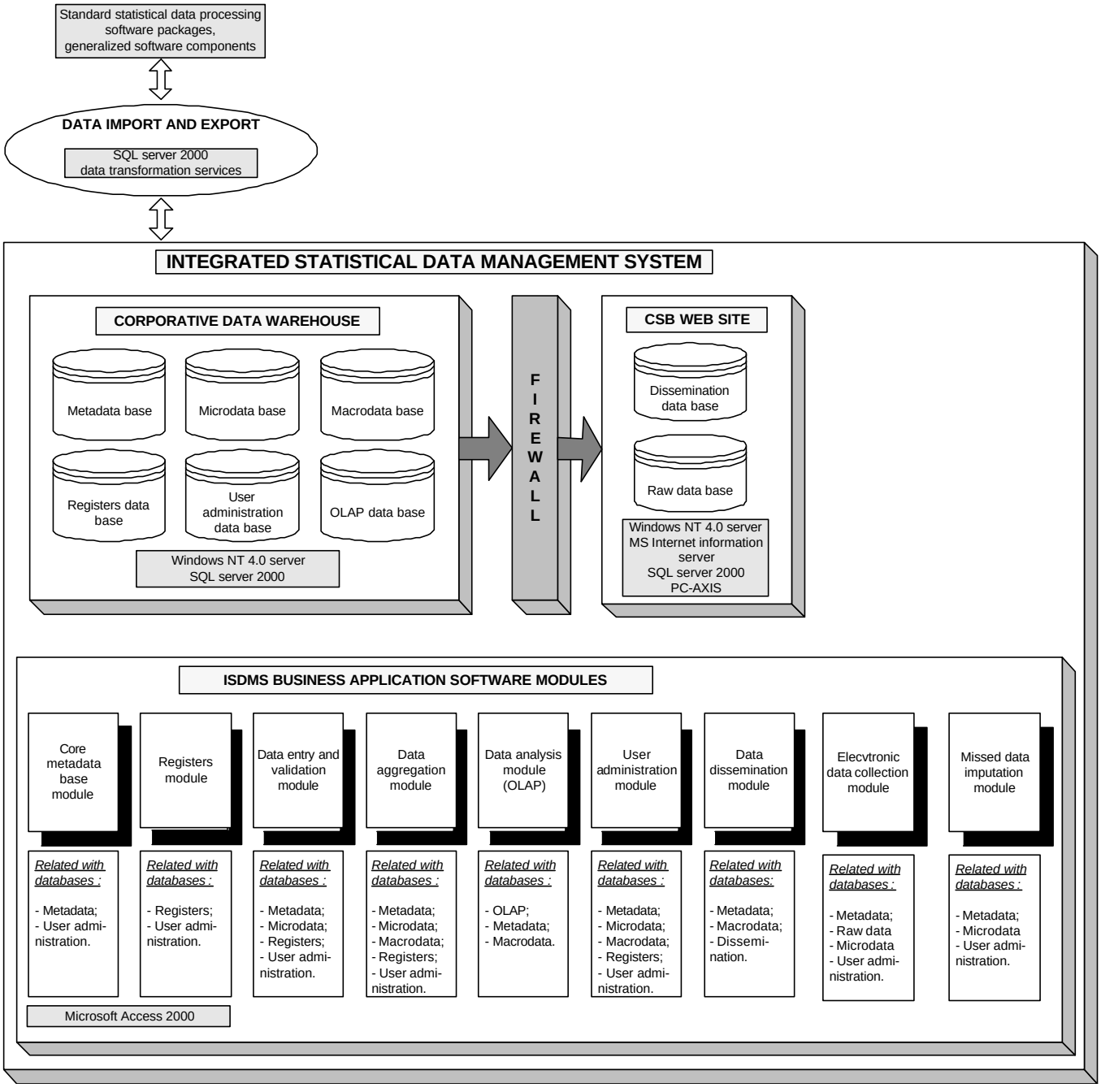
15. An important task during the design of the system was to foresee ways and to include necessary interfaces for data export/import to/from already developed standard statistical data processing software packages and other generalized software available on the market, for which the functionality was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

16. The system 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;
- WEB based data collection module;
- Missing data imputation module;
- Data aggregation module;
- Data analysis module;
- Data dissemination module;
- User administration module.

The system architecture is represented in figure 2.

Figure 2. SYSTEM architecture



III. METADATA AND METADATA BASE MODULE – CORE OF THE SYSTEM

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

18. 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".

19. The data in the meta database, in essence, are 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 meta database contains a description of statistical surveys, their content and layout, description of validation, aggregation and reports preparation rules.

- The system ensures that the meta database is used as the key element for creating a universal, common, programming-free approach for different statistical surveys data processing.
- Structure of micro data (observation data)
  [Bo Sundgren model]

20. Objects characteristics: $C_O = O(t).V(t),$

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

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

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

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

Structure of macro data (statistics)

22. Macro data are the result of estimations (aggregations). The estimations are made on the basis of a set of micro data.
Statistical characteristics: $\boldsymbol{C_S = O(t).V(t).f,}$

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

Structure of Surveys (questionnaires)

23. A new **survey** should be registered in the system. For each survey a **questionnaire version** should be created, which is valid for at least one year. If the questionnaire content and/or layout do not change, then the current version and it description in Metadata base is usable for next year.

24. Each survey contains one or more **data entry tables or chapters** (data matrix), which can be a **constant table** - with fixed rows and columns number or a table **with variable rows** or **columns number**. For each chapter we have to describe **rows and columns with their codes and names** in the Meta database. This information is necessary for automatic data entry application generation, data validation etc.

25. The last step in the questionnaire content and layout description is **cells** formation. Cells are the smallest data unit in survey data processing. **Cells** are created as a **combination of row and column** from the survey version side and **variable** from the indicators and attributes side. As an example, we could look at Retail Trade Statistics Questionnaire structure from the metadata point of view.

General information

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

Data matrix - fixed table

INDICATORS

ATRIBUTES

Goods and commodity groups	**Row code**	Turnover total 2+3+4	**Turnover from retail trade**	**Turnover from public catering**	**Turnover from wholesale trade**
Column code	B	1	2	3	4
Goods, in total	2000	***15000***	***9000***	***5000***	***1000***
Food Products(except alcoholic beverages	2010	***12000***	***5600***	***6000***	***400***
Alcoholic beverages total	2020	***3000***	***2000***	***400***	***600***
Strong drinks	2021	***500***	***300***	***100***	***100***
Wines	2022	***1000***	***500***	***200***	***300***

CELL (2010;3) contains VARIABLE = INDICATOR + ATRIBUTE

Data matrix - Fixed number of Rows and variable number of Columns
Example
Main economical indicators of the economics activity

Column heading **Row heading**	**Row code**	Total	**Title1**	**Title 2**	**n**	**Title n-1**	**Title n**
Column code		1	2	3	…	1+n-1	1+n
Attribute code			NACE 1	NACE 2		NACE n-1	NACE n
Number of employees	1010						
Net turnover	1020						
Other income	1030						

Data matrix - Fixed number of Columns and variable number of Rows
Example
Production of industry products

Product title	**Product Code**	**Row code**	**Produced in natural measurement**	**Sailed in natural measurement**	**Income in LVL**
Column code			1	2	3
Product 1	1234567	1			
Product 2	2345678	2			
……	…..	…..			
Product n-1	3456789	n-1			
Product n	4567890	n			

Creating variables

INDICATOR + ATRIBUTE (Classification) = VARIABLE

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

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

Example:
Number of employees + no attribute = Number of employees total

+ kind of activity (NACE) = Number of employees by kind of activities
+ location (Territory classification) = Number of employees in breakdown by territories

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

27. Statistical survey data processing begins with survey metadata entry in the meta database. Each new survey should be registered in the system. 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 survey content and/or layout do not change, then the current survey version and its description in the meta database is usable for the next year.

28. Each statistical survey contains one or more data entry tables or chapters. In the meta database for each chapter it is necessary to describe the table type. For each survey version chapter in the meta database 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, whose layout looks like paper questionnaires.

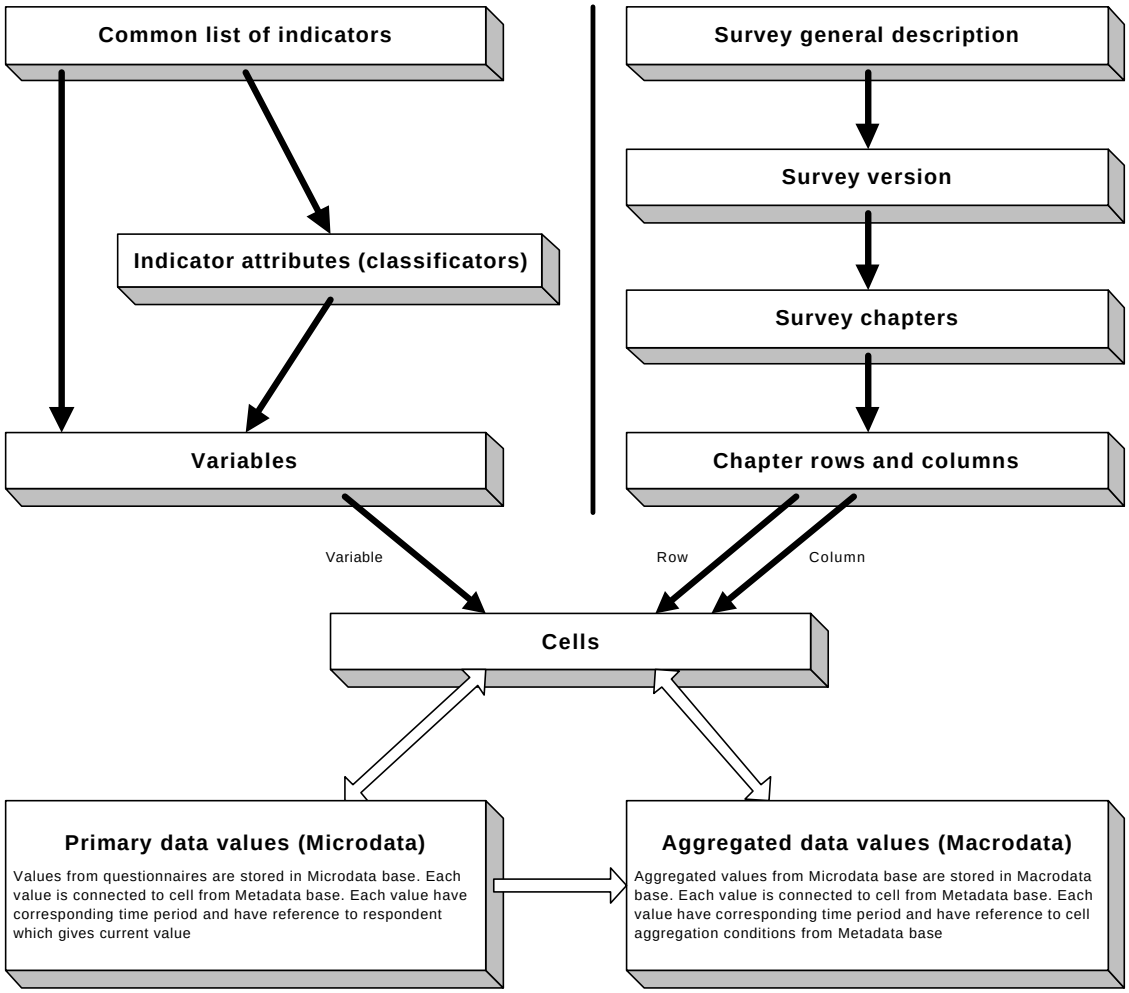
29. Information about statistical indicators is saved in the meta database. In the meta database a common indicators list is stored. Indicators are independent from surveys. This allows the possibility to attach one indicator to several surveys and to get information about one indicator from several surveys as well.

30. It is possible to define attributes – classifiers for each indicator in the system, which provides an opportunity to describe and store indicator values in a more detailed division. Indicators could be without attributes.

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

32. The last step in the survey content and layout description is cell formation. Cells are the smallest data unit in survey data processing. Cells are created as a combination of row and column from the survey version side and variables from the indicators and attributes side.

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



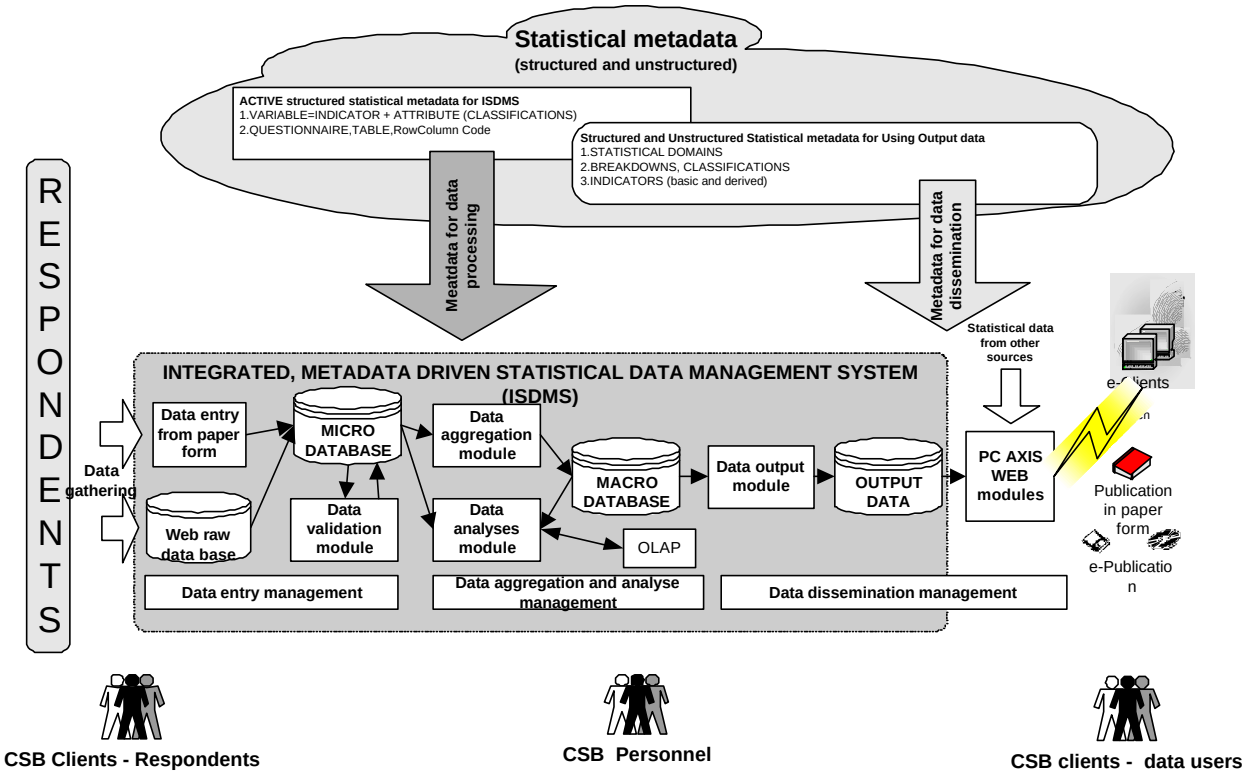
33. All survey values from the questionnaires are stored in a micro database and each value has a relation to a cell (from the meta database), which describes value meaning. Also, each value in the micro database has additional information about the respondent, which gives a current value and time period. The same situation occurs in the macro database, where aggregated values are stored. Each aggregated value refers to a cell (from the meta database), reference to each value aggregation conditions (from the meta database) and corresponding time period. Figure 4 illustrates the process chain of the statistical data production with use of different metadata profiles.

IV. ELECTRONIC DATA COLLECTION NEW FEATURE OF THE SYSTEM

34. The 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 checking and management.

35. The traditional methods of data collection currently used at CSB are based on paper questionnaires, which can be completed by interviewers or respondents. The respondent receives a paper questionnaire via post, that has to be completed and returned to CSB.

Figure 4. Statistical data production process chain



36. The current CSB bottleneck during the processing of paper form questionnaires received from respondents is due to:

- Management of requests and sending the questionnaires to respondents via post;
- Incoming paper forms data initial analysis and validation;
- Data retyping into CSB Data management system;
- Response control and management.

37. The core elements for the Electronic Data Collection Module are 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 of data quality and data can be checked immediately.

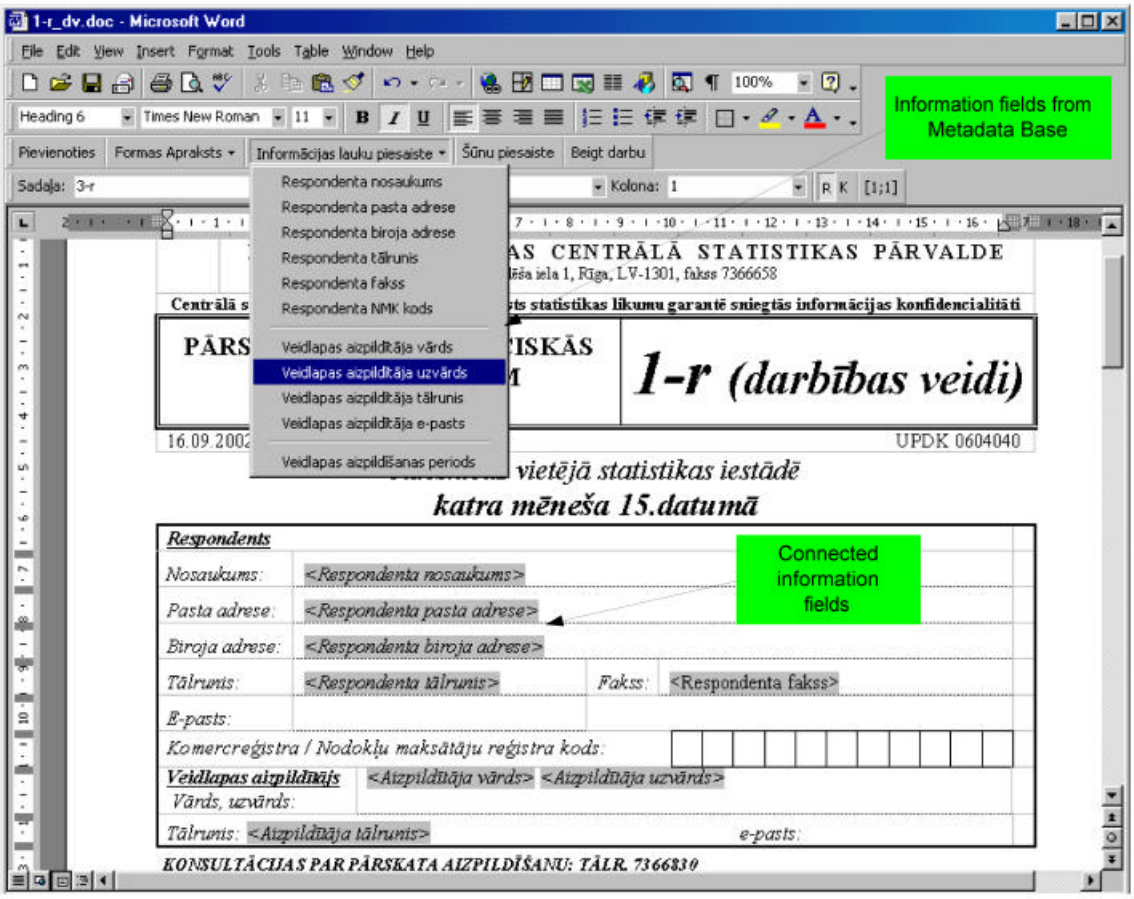
38. The business and design goals for the electronic survey data collection module were:

- Design and preparation of electronic surveys in automated mode;
- Collecting questionnaire data in electronic form from respondents;
- Improving 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 reminder systems to respondents;
- The system will remove the unnecessary stage of retyping the information by CSB personal.

39. The key features of the electronic data collection software module are:

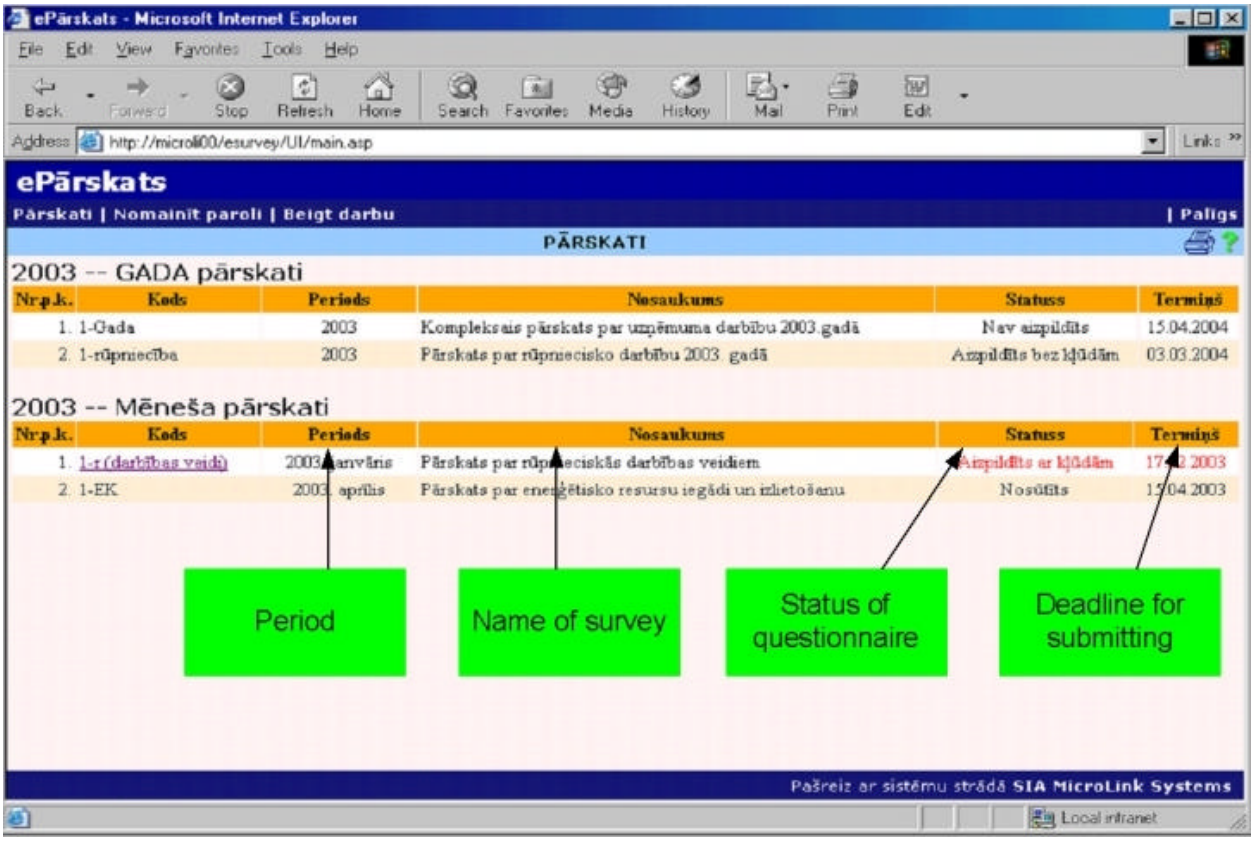
- Statisticians use the same design tool for the design of paper questionnaires as well as for WEB forms;
- Use of the metadata provides a universal approach for the generation of the WEB based data entry forms, which can be done without the involvement of IT professionals (Figure 5);

Figure 5. Preparation of the WEB form – linking with meta database



- 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:
  - Pre-loaded data: respondent or survey specific data (e.g. respondent's name and address);
  - Feedback data: historical data is available;
  - Auto-fill in fields: some fields can be automatically filled in, depending on values of previously completed fields;
  - Auto calculation: columns or other fields can, for example, be summarised.
  - Automatic validation: WEB forms could include validation rules;
  - For periodical surveys in WEB based applications respondent have a possibility to see and prove previous periods data previous periods data;
  - 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 6) as well as information about reminders;
  - The respondent is able to print questionnaires, for internal use;
  - Help facilities.

Figure 6. List of respondents surveys



- 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 database; ensures transfer of the data to the Micro data base to continue processing in the system.

V. MISSED DATA IMPUTATION FACILITY

40. 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 amount of missing data. Nevertheless, more modern strategies to cope with missing data are based on the use of modern and user-friendly software for missing data imputation.

41. Our system approach is as follows
- Extraction of the data files with missing data;
- Use of standard software packages for missing data imputation like WIDE, SOLAS or SPSS ensures imputation;
- Import of the data files/supporting meta information back into micro database/meta database.

VI. CONCLUSIONS

42. In conclusion, the following can be said:
- The design of the new information system should be based on the results of a thorough analysis of the statistical processes and data flows.
- Clear objectives of achievements have to be set up, discussed and approved by all parties involved:

- Statisticians;
- IT personnel;
- Administration.

- The initiative to move from a classical stovepipe production approach to a process oriented one has to come from the statistician, not from IT personnel or the administration;
- Improvement in the knowledge about metadata is one of the most important tasks throughout the design and implementation phases of the project;
- A clear division of the tasks and responsibilities between statisticians and IT personnel is the key point to achieving successful implementation;
- To achieve the best performance of the entire system it is important to organize the execution of the statistical processes in the right sequence;
- The new system is developed as a really metadata driven, centralized system, where all data are stored in a corporate data warehouse that allows data processing using a unified (standardized) approach for data entry and validation, data aggregation, data analysis and data dissemination for different surveys;
- The high level of flexibility of the system has been achieved 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 just accordant changes in the meta database;
- As a result of the feasibility study, we clearly understood that there are some steps of statistical data processing for different surveys, which defy standardization, some surveys may require complementary functionality (non-standard procedures), which is necessary just for data processing of that particular survey;
- To solve problems with the non-standard procedures, interfaces for data export/import to/from system have been developed to ensure use of the standard statistical data processing software packages and other generalized software available on the market;
- It is necessary to establish and train special group of statisticians, responsible for maintaining the meta database and the accuracy of the metadata;
- For the administration and maintenance of the system, 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 and system Data Model, system applications;
- Motivation of statisticians to move from an existing to a new data processing environment is essential;
- Administrative restructuring could be necessary in the move from stove-pipe data processing to process oriented data processing;
- For the proper installation and functioning of the system, 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.

43. Summing up the improvement goals and the IT strategy realised in the system, the following targets are the main ones achieved by implementation of the system:

- Increased quality of data, processes and output;
- Integration instead of fragmentation at the organizational and IT levels;
- Reduced redundant activities, structures and technical solutions wherever integration can cause more effective results;
- More efficient use and availability of statistical data by using a common data warehouse;
- Users (statistics users, statistics producers, statistics designers, statistics managers) are provided with adequate, flexible applications at their specific work places;
- Tedious and time consuming tasks are replaced by value-added activities through the more effective use of the IT infrastructure;
- Using metadata as the general principle of data processing;
- Using electronic data distribution and dissemination;
- Making extensive use of a flexible database management to provide 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.

- - - - -