

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC
COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Work Session on Statistical Metadata (METIS)
(Geneva, Switzerland, 10-12 March 2010)

CASE STUDY – CENTRAL STATISTICAL BUREAU OF LATVIA

Prepared by Jūlija Drozdova, Norberts Tālers, Jānis Linde, Central Statistical Bureau Of Latvia.

1. INTRODUCTION

Organization Name Central Statistical Bureau of Latvia (CSB)

Website <http://www.csb.gov.lv/>

Number of staff The total number of employees in 2010 is 573.
(At the beginning of 2009 there were 601 employess. 23% of all employees work in the regional structural units: Kuldīga, Preiļi and Valmiera Data Collection and Processing Centres. The interviewers and price registrars do work in different regions).

Organization structure <http://www.csb.gov.lv/csp/content/?cat=341>

Contact person (for Metadata) Julija Drozdova
Head of Statistical Meta Information Maintenance section
/IT Department
julija.drozdova@csp.gov.lv
+371 67366781

Drozdova

Metadata strategy The common strategy of CSB is available at link:
<http://www.csb.gov.lv/csp/content/?cat=4417>

In 1992 the Latvian government launched, with the assistance of the Commission of the European Communities, a programme to innovate the Central Statistical Bureau of Latvia.

We analysed the existing system of statistical indicators and harmonized with EUROSTAT Compendium. The analysis of existing processes and data flows was started simultaneously with the preparation of the data processing model which could help to define the requirements for the new IT system. From 1997 – 1999 Central Statistical Bureau of Latvia (CSB) experts in cooperation with PHARE experts prepared Technical specification for the project

“Modernisation of CSB – Data Management System”, where all technical and functional requirements for the new system were described and statistical metadata are used as the key element in statistical data processing.

Considering complicity of the project it was decided to delegate authority on development and implementation of ISDMS to outsource company with serious, long time experience in complex, large scale and large budget development projects implementation.

The idea of metadata emerged in CSB in 1999. Since 1999 metadata has been collected and analyzed. In 2002 after thoughtful analysis of data and metadata flows, Integrated Metadata Driven Statistical Data Management System (further IMD SDMS) was created.

Metadata strategy that was defined several years before was developed to cover full cycle of statistical data processing using process oriented approach instead of stovepipe approach of statistical data production.

Currently the IMD SDMS is based on following principles mentioned below:

- metadata must be created/processed/maintained in standardized environment;
- metadata must be created/processed/maintained in an integrated environment;
- metadata must be created/processed/maintained in centralized system;
- metadata must be created/processed/maintained in meta-driven system;
- metadata must be created/processed/maintained in transparent system;
- metadata must be created/processed/maintained in system, allows automated generation of user application forms;
- metadata must be created/processed/maintained in system which has a modular structure;
- metadata must be processed in system that allows closer connection to respondents.

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

- Increased quality of data, processes and output;
- Integration instead of fragmentation on organizational and IT level;
- Reduced redundant activities, structures and technical solutions wherever integration can cause more effective results;
- More efficient use and availability of statistical data by using common data warehouse (concerning IMD SDMS, see section “Current situation”);
- Users provided (statistics users, statistics producers, statistics designers, statistics managers) with adequate, flexible applications at their specific work places;
- Tedious and time consuming tasks replaced by value-added activities through an more effective use of the IT infrastructure;
- Metadata used as the general principle of data processing;
- Electronic data distribution and dissemination used;
- Making extensive use of a flexible database management provides users with high performance, confidentiality and security;

Separate storages of data and metadata in CSB should be handled by corporative repository, therefore the strategy in next years will be to focus on a corporative data and metadata repository creation, development and implementation.

One of the main aims of repository is to commonly refer to a location for data and metadata storages, providing data and metadata safety and preservation.

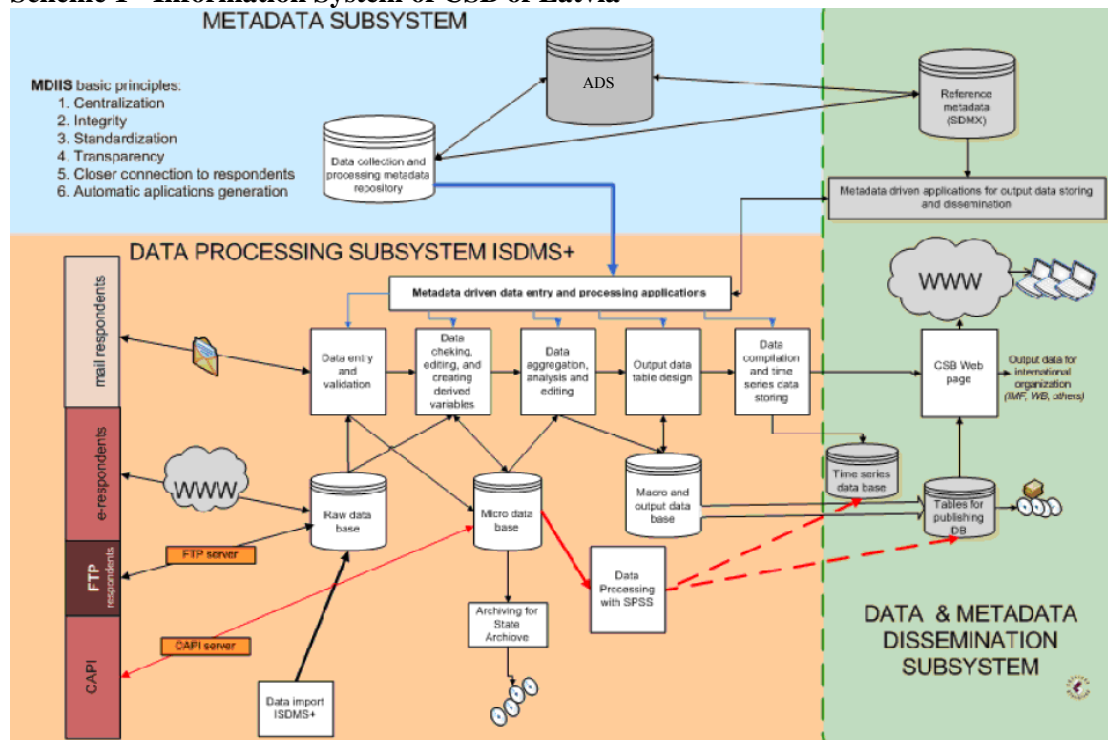
In the future the NSI of Latvia is considering to implement a project, which foresees the creation of the References metadata base.

Current situation

The Information System, which is presented in Scheme 1, is the successfully working system,

but some elements (like the ADS; time series data base; Reference metadata base (SDMX)) of this one at the moment are reworked or under construction.

Scheme 1 - Information System of CSB of Latvia



Information System of CSB of Latvia is divided in 3 subsystems:

1. Metadata subsystem. This subsystem consists of reference metadata base (the ADS). This system is intended as a reference metadata repository, which stores methodological and quality metadata in a harmonised way and the information is entered by subject matter units.

This subsystem comprises: Projects Documentation System (the ADS). The purpose of the ADS is to document the main processes of preparation of the statistical output (surveys/calculations) of the CSB, including reporting on quality according to the quality criteria of the ESS. The ADS consists of 2 parts: internal and external.

Main elements available to users are:

1. Descriptions of statistical output (methodological description and quality indicators that will be available for internal users in full amount, and selected information – for external users);
2. Thesaurus (definitions of statistical indicators).

The benefits of the system:

- centralised knowledge base at CSB on statistical output;
- comparable meta-information over time;
- metadata on quality;
- source for quality reports and other reports on statistical output;
- source for meta information both for internal & external users;
- base for SDMX.

The other data base is data collection and processing metadata repository, which is guiding

the MD ISDMS data processing processes.

2.Data processing subsystem IMD SDMS+

The data management system as such. The data base structure is a result from implementation of the Metadata Driven Integrated statistical Data Management System and regional restructuring. MD ISDMS at the time being is able to work with the Business statistics surveys.

The on-going development will result in metadata driven Social statistics data collection and in the future data processing system, which would cover Computer Assisted Personal Interviewing (CAPI), Computer Assisted Telephone Interviewing (CATI), Computer Assisted WEB Interviewing (CAWI).

Eventually data collection and processing systems at CSB will be metadata driven at the system level and would cover all means of collecting data from respondents.

3.Data and Metadata Dissemination subsystem is foreseen for storage and loading of references metadata. For the time being it is not an integrated system which holds data and metadata descriptions in completely integrated way, but nevertheless it has a connection between these instances so that the data and metadata descriptions are linked together and data user can see metadata about the particular data table available.

The main problem of the current situation is that common repository for all storages is missing.

2. STATISTICAL METADATA SYSTEMS AND THE STATISTICAL BUSINESS PROCESS

2.1 Statistical business process model

The [Statistical Business Process Model of CSB, presented in Scheme 2](#) (SBPM of CSB) defines the business processes needed to produce official statistics at CSB (for the description of quality assessment of processes at national and international level based on statistical sources). SBPM of CSB that was presented below is not an official statistical business process model.

The business process model has not adopted yet by CSB. We took a Generic Statistical Business Process Model (GSBPM) and have highlighted those model parts, which are covered by our IMD SDMS system. We mapped system functionalities with GSBPM to visualize consistency between model level and implemented data processing system in use. The SBPM of CSB presented with regard to the Metadata Management process. The SBPM of CSB comprises for levels: Level 0, the Metadata Management process; Level 1, the nine phases of the Metadata Management process; Level 2, the sub-processes within each phase; Level 3, a description of those sub-processes. SBPM shows two overarching processes as well: Quality management and Support processes that apply throughout the Metadata Management process.

2.2 Metadata system

The subject of this case study is IMD SDMS. This system provides the complete cycle of statistical data production processes for business statistics surveys.

The creation of the [Core Meta data base module – Scheme 3](#), with fundamental models of structure of Micro data/Macro data was the primary task of the IMD SDMS development. Meta data base is linked at database structure model level with Micro data base and Macro data base ([see figure 2](#)). Correctly and carefully planned databases structure model design is the basis for successful further system development and implementation.

All survey values from questionnaires are stored in Micro data base and each value has relation to cell (from Meta data base), which describes value meaning. Also each value in Micro data base has additional information about respondent, which gives current value and time period. The same situation is in Macro data base, where aggregated values are stored. Each aggregated value has reference to cell (from Meta data base), reference to each value aggregation conditions (from Meta data base) and correspondent time period.

Meta data base module contains following main applications:

- Description of statistical questionnaire;
- Description of questionnaire version;
- Description of indicators and attributes of statistical questionnaire;
- Description of content of statistical questionnaire chapters;
- Maintenance of validation rules of statistical questionnaire;
- Description of derived variables;
- Description of aggregation conditions of statistical questionnaire;
- Description of output tables conditions of statistical questionnaire;
- Grouping of classifications records;
- Common Meta data base data browsing;
- Applications of data Import/Export/Impute;
- Maintenance of Data electronic archiving system;
- Description of electronic questionnaire for electronic data collection through CSB WEB page. This is not a application, but the special medium of preparing WEB survey using predefined before metadata variable through MS Word and special tools provided by IMD SDMS;

Specially trained personnel (4 persons) of the Statistical Meta Information Maintenance Section under IT Department operate the Meta data base module. They have rights to perform Meta data entry, updating, changing and are responsible for accurateness of Meta data. It is very important, that Meta data entered into Meta data base are carefully checked and corrected, because these Meta data 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.

CSB of Latvia is using Integrated Metadata Driven Statistical Data Management System, which covers a part of GSBPM, and is intended as a system for processing of business statistics. A system with similar principles is being developed for social statistics now at the organization.

2.3 Costs and Benefits

System was launched in production in August 2002, and held 25 different surveys metadata descriptions at starting point.

Successful implementation formed basis for the CSB regional restructuring, which has been implemented within the period of two years from 2003 to 2004. Five Data Collection and processing centres replaced previously existing 26 Statistical Regional offices an city Riga office thus taking on responsibility for overall data collection and editing and decreasing amount of necessary statisticians working with data collection and editing from 180 to 115.

2.4 Implementation strategy

The project was implemented with a step-wise approach. From 1997 – 1999 CSB experts in cooperation with experts contracted from PricewaterhouseCoopers were prepared General Technical Requirements for the project “Modernisation of CSB – Data Management System”. Technical Specification embodied key technical and functional requirements for the new system where statistical Meta data should be used as the key element in statistical data processing. A lot of additional requirements appeared within the process of development

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

Using modern IT solutions:

- Increase efficiency of the main process at CSB, production of statistical information;
- Increase the quality of the statistical information produced;
- Improve processes of statistical data analysis;
- Modernise and increase the quality of data dissemination;
- Avoid hard code programming via standardisation of procedures and use of Meta data within the statistical data processing

3. STATISTICAL METADATA IN EACH PHASE OF THE STATISTICAL BUSINESS PROCESS

3.1 Metadata Classification

The CSB of Latvia doesn't have a formal classification of metadata. However it could be classified as follows (5 groups):

1. Dissemination metadata – all metadata is foreseen for end users, such as classification, data interpretation and etc.
2. Metadata on quality.
3. Metadata for data collection purposes. This metadata is used by interviewers and respondents. For example: various instructions for interviewers and respondents for coordination their activities; interviewer's guidelines and etc. This group of includes a great amount of information, therefore should be presented as a separated one.
4. Metadata for statistical data processing purposes. All metadata used in an IMD SDMS and allows producing statistical data through the cycle of statistical data processing.
5. Operational Metadata (paradata). Data about all statistical processes at NSI. There is no relation with survey's paradata.
6. System metadata – all information referring to the IT environment, including necessary information for supporting this environment.

3.2 Metadata used/created at each phase

Metadata used at each phase of [SBPM of CSB](#). The subject of this case study is IMD SDMS. For this reason only the main points for those sub-processes that are supported by IMD SDMS will be described, namely sub-processes marked in blue on SBPM.

The most important sub-processes which affect statistics quality at a greater degree are:
2.5 Design statistical processing& workflow methodology;

- 3.1 Build data collection instrument;
- 3.2 Build or enhance process components;
- 4.3 Run collection;
- 5.3 Analyze, validate, review, edit & impute
- 7.1 Update output systems
- 7.3 Manage release of dissemination products

Phase 3 – Build.

This phase provide for the real test before running of data collection in phase 4.

In the first case this phase is not complicated if:

- all data collection instruments and necessary components already have been built and it is not a first iteration of phase;
- all workflows configured previously and production system for example for concrete survey was applied when the first collection of that survey was done.

In the second case this phase a little bit complicated when some new survey is going to be collected. The second case will be described further including six sub-processes of that phase.

These sub-processes are not sequential from top to bottom; mostly they occur in parallel and some of them are iterative.

Sub-process 3.1

When a design is approved and all requirements in sub-process 2.5 (this process is strongly linked to Phase 8 – Archive) are specified the sub-process 3.1 come into force.

Sub-process 3.2

This sub-process consists of a building/improvement the various components of the production system. This process can be either time consuming if it concerns the case where a questionnaire is described in the system for the first time, and the functionality can be insufficient to cover all of the need for a given questionnaire (in a case where it is a non – standard questionnaire compared to other business statistics questionnaires) or very little time consuming in a case where existing tools are in line with requirements set by a particular questionnaire.

In that case the following steps must be realized in that sub-process:

1. **Registration of statistical survey and attachment of all necessary methodological information to survey's version.** New survey (questionnaire) should be registered in the System. For each survey a questionnaire version should be created, which is valid for at least one year. If questionnaire content and/or layout does not change, then current version and its description in IMD SDMS is usable for the next year.
2. **Description of: indicators, attributes and content of chapters of statistical questionnaire.** Each survey contains one or more data entry tables or chapters (data matrix), which could be constant table - with fixed number of rows and columns or table with variable number of rows or columns. For each chapter it's necessary to describe rows and columns with their codes and names in IMD SDMS. This information is necessary for automatic data entry application generation, data validation etc. Last step in the questionnaire content and layout description is cell formation. Cells are the smallest data unit in survey data processing. Cells are created as combination of row and column from survey version side and variable from indicators and attributes side. As an example of the fixed structure table on [Figure 1](#) we could look at Retail Trade Statistics Questionnaire structure from Meta data point of view. All necessary survey's variables with attributes or without them must be defined into IMD SDMS. Using these variables user interfaces for statisticians are created

automatically by system. The principle of creating defined by formula:

INDICATOR + ATTRIBUTE (Classification) = VARIABLE, where are

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

- Kind of activity – NACE
- Territory and etc.

Example:

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

- 3. Maintenance of validation (error detection) rules of statistical survey.** These rules described in IMD SDMS using pseudo language.
- 4. Description of conditions for aggregations and output tables of statistical survey.** On this step micro data turns into macro data. If necessary grouping of classifications records are performing. Conditions of aggregations: SUMM, COUNT (frequence), MAX1, MAX2, MIN, MEAN.

In addition the steps mentioned below can take place as well:

- 5. Description of layout for electronic statistical survey.**
- 6. Description of derived variables.** IMD SDMS calculates derived variables for defined questionnaire's version by using micro data from this version or from other survey-/s version-/s. This step is used for Structural Business statistics (SBS), where all necessary variables are combined like as a quasi survey in IMD SDMS.

Sub-process 3.3

The main objective of this sub-process is to be sure that the workflow (from data collection to dissemination) specified in sub-process 2.5 work in practice. There is no reason to check workflow with regard to the archiving phase, because if questionnaire was described in IMD SDMS there is always no problem with archiving

Sub-process 3.4

This process starts with testing of production system and ends with approval of successful operating within that system. In this work process both IT developers and statisticians need to be convinced that numerous components work together through configured workflow.

Sub-process 3.5

This sub-process always occurs in parallel with sub-process 3.4 through small-scale testing of data collection with special respondents for these testing purposes. In this sub-process a match of business process qualitative and quantitative information is performed.

This match is possible only in case if questionnaire's version has been described before in IMD SDMS and therefore all quantitative and qualitative information about this version is available (i.e., number of variables in the version questionnaire's or list of attributes attached to the version and etc.).

Sub-process 3.6

Successful results of previous processes leads to training of users and attachment of some documentation. Some kind of technical documentation can be enclosed as well.

Phase 4 – Collect

This phase is based on both: the methodology created in Design phase, namely in sub-process 2.5 and collection instruments prepared in Build phase..

Sub-process 4.1

The sample/sample is selected, validated and documented. Successful results of this work process leads to attachment of list of respondents.

Sub-process 4.2

This sub-process foresees the checking of availability for all data collection elements, for example: the list of respondents was attached, all data input user interfaces were prepared, all questionnaires were printed, electronic surveys forms were available and etc. This work had to be done by statisticians responsible for their surveys.

Sub-process 4.3

This sub-process is to evidence when the first contact between respondent and statisticians was done. The first contact occurs at the beginning of the each calendar year when the list of questionnaires and questionnaire's forms are sent out to respondents.

After that we have two scenarios of behaviour:

- respondents decide to fill in questionnaire electronically and addresses to CSB for proving of his account or fill the electronic questionnaire if he already has an the account.
- respondents provide a filled-up paper questionnaire to CSB

Sub-process 4.4

Actually, we had discussions to show this sub-process on SBPM of CSB or not. The reason is our system has the raw data base only for data collected electronically, but when data is collected from paper's questionnaires it is classified, coded and edited simultaneously and in that case the data is inputted by advanced statisticians, which simultaneously validate the data during inputting.

As the result of that is CSP doesn't have operator's data input and in such a way CSP provides a less amount of human resources for data input.

This sub-process simultaneously is related to several sub-processes: 5.2 – Classify & code; 5.3 – Analyze, validate, review, edit & impute; 8-Archive;

The main purposes for sub-process 5.3 is: to analyse, check collected data accuracy, correct it, and get a clean data set at the end. Involved organizational units and their roles: data collection unit, respective subject-matter section (e.g. Trade statistics section).

Looking into 4.4 from other perspective, in the case if electronic surveys are also used for data collection, the situation will be slightly different. Electronic respondents (who fill in electronic questionnaires) don't get full list of validation during data input process. This approach is used for avoiding of respondents burden. Therefore in this sub-process "raw data base" of data submitted by respondents through electronic surveys happens.

Phase 5 – Process

In this phase, the statistical data is analyzed, checked and "refined".

Sub-process 5.1

It consists of such activities as data integration with other source. This source can be a mixture of external or internal data sources or extracts from administrative sources.

Integration process comprises data import function, supported by IMD SDMS.

For example, some variables survey of employment statistics can be integrated with

another surveys variables of employment statistics. Using special applications of IMD SDMS matching with the aim of linking data from different sources can be done, where data refer to the same unit.

Sub-process 5.2 (see also sub-process 4.4)

In this sub-process the input data is classified and coded. In IMD SDMS numerous pre-defined classifications (local, national, international and validation classifications for calendar values) are maintained. During data input in IMD SDMS, system automatically provides a list with all classifications codes of some kind of classification with textual descriptions for each classification code. Respondents or statisticians just need to choose corresponding code, if a particular data entry cell foresees usage of classification.

Sub-process 5.3 (see also sub-process 4.4)

All activities in this sub-process are operated at micro data level. Within the sub-process d error checking is done (error detection, for example: the sum of breakdowns by NACE doesn't equal to the total sum), item non-response and miscoding, data imputation (always flagging as imputed).

Sub-process 5.4

New variables are derived using special application in IMD SDMS within this sub process, where all arithmetic formulas are described with pseudo code expressions. The responsible statistician for defined survey just needs to press the button to launch calculations. Each arithmetic formula has its priority of performance.

During the description of the questionnaire's version in IMD SDMS all (associated) derived statistical units are defined.

Sub-process 5.6

After creation of weights and their import at micro level in IMD SDMS, the system aggregates data from micro-data. Aggregation characteristics and aggregation algorithms (within one version of the defined survey) were described before at step 3.2 or can be described in the sub-process 5.6.

If there are some problems with data aggregation IMD SDMS informs users about the type of problem (for example: aggregation field is not filled with certain statistical units).

The main point of aggregated data – they are aggregated at lowest classification level (for example: NACE family (4 digit); CPA (6 digit); PRODCOM (8 digit) and etc.). It should be underlined that data analysis for aggregates will be available at lowest classification level as well.

Sub-process 5.7

This sub-process provides macro data sets (or output tables), which are used as the input to phase 6 Analyze. All versions of provisional or final macro data sets are stored in IMD SDMS and have attached to them calendar dates, which show when macro data sets were created.

Phase 6 – Analyze

In this phase statistics are produced, analysed and prepared for dissemination.

Sub-process 6.1

This sub-process can be divided into two parts: the first part is covered by IMD SDMS and the second one covered by other processing tools.

This sub-process brings together the results of output tables (created after data aggregations) or/and results of production of additional measurements such as indices, trends or seasonally adjusted series, recording of quality characteristics.

IMD SDMS has a possibility to create output tables (within one version of the defined survey). The main point of output tables – they are summarized at highest classification level (for example: summarized data by NACE section “F” -

Construction). In that case analytical data for output tables will be available at highest classification level. This case occurs for data, which is published in absolute values. Others processing tools like as Demetra, Access, SQL and etc., cover another part of this sub-process.

Sub-process 6.2

In this sub-process statisticians validate the quality of the outputs produced at micro and at macro levels. The divergence from expectations is analyzed. This sub-process performed by following components: IMD SDMS analytical tools, OLAP (Dealing with OLAP data cube. Using OLAP is it possible to get to micro data from macro data during analyzing the divergence), SQL procedures, SPSS tools.

In this sub-process only IMD SDMS analytical tools will be described.

IMD SDMS analytical tools are foreseen for macro and micro data express analysis. Analytical tools for Microdata allow easy to create any kind of data requests from individual data in different breakdowns, for different periods. These tools provide the export possibilities to XLS or ACCESS for further processing.

Analytical tools for Macrodata allow data requests of aggregated data sets at different levels of aggregation. The results of requests can be exported to XLS or ACCESS for further analysis.

Sub-process 6.4

This sub-process covered by IMD SDMS as well. IMD SDMS automatically applies confidentiality rules for macro data sets, making checks for primary disclosure.

There are three main primary confidentiality conditions and each of them is marked in provided data set (output tables) group by its own color, therefore it is very handily for statisticians. The confidentiality conditions are described in details in CSP Confidentiality handbook.

Phase 7 – Disseminate

This phase manages the release of the statistical products to customers. This phase deals with checking data and metadata readiness for dissemination.

Sub-process 7.2

In this sub-process PC-Axis tools are widely used, as it helps to map data and metadata for putting into dissemination output file system.

Phase 8 – Archive

This phase dealing with micro data and meta data using Data electronic archiving (DEA) system's applications in IMD SDMS.

DEA system performs preparation of statistical documents (surveys) in electronic format for their deposition to the State Archive of Latvia.

This phase is made up of four sub-processes, which are generally sequential from top to bottom:

Sub-process 8.1

This sub-process determines the archiving rules, namely; conditions and the medium of archiving.

The conditions under which data and meta data should be archived:

- data and associated metadata for year three years ago (for example: in 2010 data will be archived for 2007);
- archived data must correspond to the data structure of IMD SDMS and must be matched with corresponding structure by using special IMD SDMS application, if it does not. This is valid in cases when data not stored in IMD

- SDMD has to be archived;
- archiving process must be carried out from external data in a file or from data which is stored in IMD SDMS
- respondents data which is collected in sub-process 5.3 but without data which has been imputed

The medium of archiving:

- DEA provides a medium by which the archiving document is created (this archiving document includes different kinds of information like as respondent's data; surveys questionnaire; thesaurus - structured content of archiving information and etc.)

Preservation of data and associated metadata:

- data is prepared by DEA in HTML 4.01 using Baltic-1257 coding, data is burned on data carrier (CD-R);
- data preserved on IS server and is available for viewing in IMD SDMS

Sub-process 8.2

This sub-process includes the match of data structure for archiving;

Sub-process 8.3

This sub-process provides the following activities:

- identifying data and meta data for archiving in line with the rules defined in 8.1
- if necessary formatting those data and metadata for the repository after matching
- transferring data and metadata to the repository;
- create the archiving documents
- verifying that the data and meta data have been successfully archived

Sub-process 8.4

- Data and associated metadata is disposed off in line with rules in 8.1 and is prepared by DEA in HTML 4.01 using Baltic-1257 coding. Data is burned on data carrier (CD-R). Archived data is easy retrieved from DEA system. The special flag in IMD SDMS has been done and data is available for viewing in IMD SDMS.

3.3 Metadata relevant to other business processes

Apart from Metadata management process for statistical data processing purposes all others business processes use metadata as well.

The Business processes conditionally can be named as follows:

1. Meta data dissemination processes
2. Quality management process
3. Metadata management process for data gathering purposes
4. Metadata management process for statistical data processing purposes (described in details especially within this case study)
5. Operational processes
6. System metadata

First of all it should be noted that 4.Metadata management process for statistical data processing purposes was described in details in section 3.2 for particular case within IMD SDMS.

4. SYSTEMS AND DESIGN ISSUES

4.1 IT Architecture

Before development and implementation of the system classic Stove Pipe data processing approach with all appropriate technical incompatibilities existed as a consequence of the wide range of technology solutions that were in use.

As the result of the analysis of processes, data flows, user requirements and situation mentioned above it turned out that most of statistical surveys have the same main steps of data processing starting with survey design and ending with statistical data dissemination. The division was necessary between surveys filled in by respondent and surveys filled in with assistance of interviewer. The main difference was found in both data obtaining methods and data aggregation algorithms obtaining data from businesses and from persons & households. Business respondents are filling in questionnaires are either mailing them to CSB or enter the data in electronic survey system. Data from persons & households are obtained via interviewers service. Statistics structuring in the Central Statistical Bureau of Latvia is presented on a high level diagram as it is shown on the [Figure 3 - Statistics Structuring in CSB based on the Process Oriented data processing.](#)

A typical statistics production high level workflow can be seen as very simple diagram on [Figure 4 - Typical statistics production high level workflow.](#)

Looking deeper in the statistical processes taking place in Statistics Latvia we can define them as in [Figures 5](#) and [Figure 6.](#)

The corporative data warehouse of CSB is presented [in Figure 7.](#)

As the theoretical basis for system architecture “Information systems architecture for national and international statistical organizations” elaborated by professor Mr. Bo Sundgren (Statistics Sweden) and issued by UNSC and ECE and approved by Conference of European Statisticians as Statistical Standard was taken.

New system contributes harmonization and standardization and is developed as centralized system, where all data are stored in corporate data warehouse. The approach is by using advanced IT tools to ensure the rationalizing, standardization and integration of the statistical data production processes.

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

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

- Meta data base module;
- Registers module;
- Data checking, editing and derivation module;
- Missing data imputation module;
- WEB based data collection and administration module;
- Data aggregation module;
- Output tables module;
- Data analysis module;
- Data dissemination module;
- User administration module;
- DEA module;
- Respondents response and reminder system.

4.2 Metadata Management Tools of IMD SDMS

All metadata management tools are provided by IMD SDMS. The modules (described in Section 4, see description of modules, which have to cover and to support all phases of the statistical data processing) provide the management tools for metadata.

4.3 Standards and formats

The metadata standards and file formats being used within CSB metadata systems:

1.The ADS. This system at the moment is under implementation. ESS documents on quality reporting (Standard Quality Report and Standard Quality Indicators) have been used as the base for the development of the structure for ADS projects.

2. IMD SDMS, based on: guideline “Information systems architecture for national and international statistical offices, guidelines and recommendations, United Nations, Geneva, 1999” applied by CSB for metadata production. In particular: fundamental concepts: “statistical characteristic” and “estimated statistical characteristic”, aspects of the metadata infrastructure of a statistical organization, strategy for the development and implementation of a metadata infrastructure for a statistical organization”; Complies with: ISO/IEC 11179, Information technology – Specification and standardization of data elements, national standards on metadata and SDMX standard). File formats: *.px; *.xls; *.dbf; *.xml; *.html, *.doc

3.Data and Metadata Dissemination subsystem. Files-structured storage. Reference metadata structure is based on SDDS; the standard template is used for preparation of reference metadata within publication table. File formats: *.px; *.xls; *.xml; *.html

4.4 Version control and revisions of IMD SDMS

Metadata systems are controlled and revised permanently by responsible staff . The versioning of the system has no set rules, instead, system updates project may be launched if there is a reasonable requirements that the system does not meet. As for the version control of metadata descriptions, the version of questionnaire IMD SDMS is defined within one-year period, therefore each version with associated metadata is revised once per year.

At the moment CSB has the fourth version of the IMD SDMS. In comparison with the first version they are significant differences: new functionalities were built up and more user friendly interface was provided.

4.5 Outsourcing versus in-house development

IMD SDMS is developed by outsource company. After the eight years of the successfully exploitation of the IMD SDMS we found that system functionality should be reasonably increased.

Since 2009 a project has been launched for the IMD SDMS to cover Social statistics domain.

4.6 Sharing software components of tools

-

4.7 Technical platform and standard software used

The first version of the system, to be in line with the CSB IT strategy, existing computer and network infrastructure for the system development the Microsoft SQL Server was taken to handle system databases. All applications comply with the client/server technology model, where data processing performed mostly on server

side. Other components of Microsoft Office are used as well. For multidimensional statistical data analysis is used Microsoft OLAP technology. As tool for data dissemination was chosen software product PC-AXIS developed by Statistics Sweden, which is widely used in different statistical organizations in different countries.

The last version of the system has been upgraded to MS Server 2003, MS SQL Server 2005 and applications reprogrammed in .net in 2008.

5. ORGANIZATIONAL AND WORKPLACE CULTURE ISSUES

5.1 Overview of roles and responsibilities

There are several organizational units involved in development and maintenance of metainformation systems.

<i>Contact person</i>	<i>Phone; E-mail</i>	<i>Job title</i>	<i>Unit</i>	<i>Associated metainformation system</i>
<i>Norberts Talers</i>	<i>+371 67366650 Norberts.Talers@csb.gov.lv</i>	<i>Vice president</i>		<i>Responsible for IT and data dissemination at CSB of Latvia</i>
<i>Uldis Ainars</i>	<i>+371 67366920; Uldis.Ainars@csb.gov.lv</i>	<i>Director of the department</i>	<i>Information, publishing and printing department</i>	<i>Data and Metadata Dissemination subsystem</i>
<i>Jolanta Minkevica</i>	<i>+371 67366629; Jolanta.Minkevica@csb.gov.lv</i>	<i>Head of the division</i>	<i>Statistical Methodology and Organization Division</i>	<i>Surveys & calculations documentation system</i>
<i>Julija Drozdova</i>	<i>+371 67366781; Julija.Drozdova@csb.gov.lv</i>	<i>Head of Statistical Meta Information Maintenance section</i>	<i>IT Department</i>	<i>Integrated Metadata Driven Statistical Data Management System</i>

The information with regard to the roles and responsibilities of the staff is available in Annual Report 2008 of Central Statistical Bureau at link:

<http://www.csb.gov.lv/csp/content/?cat=4601>

5.2 Training and knowledge management

CSB staff training for working with IMD SDMS was realized by CSB staff. The training provides all necessary knowledge for all subject matter units. This training includes the detailed considering within full cycle of system production processes.

In general, the employees have opportunity to improve their own skills in the training of European statisticians. In the training they can acquire knowledge on current statistical subjects, as well as, basic knowledge of European statistics. The above courses mainly offer knowledge that is not possible to acquire in each individual country. The courses are conducted by highly qualified experts in the respective areas both from the member states of the European Union and various international organisations and institutions. The courses take place across a broad geographical area - in Sweden, Norway, Finland, Luxembourg, etc.

6. LESSONS LEARNED

6.1 The list mentioned below provides key points of “lessons learned” from planning developing and maintaining metadata management system:

- Design of the new information system should be based on the results of deep analysis of the statistical processes and data flows;
- Clear objectives of achievements have to be set up, discussed and approved by all parties involved: statisticians, IT, Administration;
- As the result of feasibility study we clearly understood, that all steps of statistical data processing for different surveys allows standardization, while each survey may require complementary functionality (non standard procedures), which is necessary just for this exact survey data processing;
- For solving problems with the non-standard procedures interfaces for data export/import to/from system has been developed to ensure use of the standard statistical data processing software packages and other generalized software available in market;
- Within the process of the design and implementation of Metadata driven integrated statistical information system both parties - statisticians and IT specialists should be involved from the very beginning;
- Clear division of the tasks and responsibilities between statisticians and IT personal is the key point to achieve successful implementation;
- Both parties have to have clear understanding of all statistical processes, which will be covered by the system, as well as Metadata meaning and role within the system from production and user sides;
- Initiative to move from classical stove-pipe production approach to process oriented have to come from statisticians side and not from IT personnel or administration, therefore motivation of the statisticians to move from existing to the new data processing environment is essential;
- Improvement of knowledge about Metadata is one of the most important tasks through out of the all process of the design and implementation phases of the project (knowledge of theoretical aspects);
- It is necessary to establish and train special group of statisticians, which will maintain Metadata base and which will be responsible for accurateness of Metadata;
- To achieve the best performance of the entire system it is important to organize the execution of the statistical processes in the right sequence;
- Data electronic archiving reduces human resources (at the moment 2 persons), time of archiving and physical amount of archiving information (In 2000, the amount of the archiving information of Population Census has occupied the space in 21 m³ which was equal to 4 DVD). It should be highlighted that the expenses of CSB for deposition in the State Archives of Latvia are reduced as well;
- IT developers must draw an attention on Sub-process 3.6 and get submission from statisticians that this process is being tested, because statistician is the best in their field;
- Taking into consideration experience of CSB of Latvia for creating Metadata system, which is based on MS products, the following key points are actual:
 - For the administration and maintenance of the system it is necessary to have well trained IT staff, which is familiar with the MS SQL Server administration, MS Analysis Service,
- other MS tools, PC AXIS family products and system Data model, system applications;

7. THEORETICAL BACKGROUND

7.1

- Handbook on Design And Implementation Of Business Surveys. Edited by Ad Willeboordse. October 1997;
- Guidelines for the modeling of statistical data and metadata. UNITED NATIONS. Conference of European statisticians. Methodological material. Geneva, 1995;
- Information systems architecture for national and international statistical offices. Guidelines and recommendations. Conference of European statisticians. Statistical standards and studies – no.51. United nations. Geneva, 1999;
- Guidelines for statistical metadata on the Internet. Conference of European statisticians statistical standards and studies – no. 52. United nations. Geneva, 2000;
- ISO/IEC 11179-1 Information technology — Specification and standardization of data elements - INTERNATIONAL STANDARD. First edition. 1999;
- Developing and implementing statistical metadata systems. Bo Sundgren. 2003-04-06. Draft version of MetaNet WG3 deliverable;
- Sundgren B. (2003). Statistical Metadata and Data Warehouse. Metaware IST-1999-12583;
- Terminology on Statistical Meta data. Conference of European Statisticians, Statistical Standards and Studies No 53;
- Willeboordse A. Towards a New Statistics Netherlands. Blueprint for a process oriented organisational structure;
- Towards an SDMX User Guide: Exchange of Statistical Data and Metadata between Different Systems, National and International. By Bo Sundgren, Statistics Sweden, Christos Androvitsaneas, ECB and Lars Thygesen OECD. OECD Expert Group on Statistical Data and Metadata Exchange. April 2006. Geneva (*the text contains a reference that the most ambitious project has been labelled as “metadata-driven statistical data management system”*).