

Case study: Central Statistical Bureau of Latvia

Case study: Central Statistical Bureau of Latvia Central Statistical Bureau of Latvia: use of GSBPM

Contact person*	Julija Drozdova
Job title	Head of Statistical Meta Information Maintenance section /IT Department
Email	julija.drozdova@csb.gov.lv
Telephone	+371 67366781

Summary*

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

In 2010-2011 IMD SDMS has been modified to cover statistical data collection and processing for Social Statistics as well. IMD SDMS has received the new name - MetaData Driven Integrated Statistical Data Management System - Computer Assisted Survey Information System (further, MDD ISDMS - CASIS).

For today MDD ISDMS - CASIS covers both: Business and Social statistics. MDD ISDMS - CASIS is capable to replace completely "BLAISE" (e Survey processing system). The Population Census 2011 of Latvia has been started in MDD ISDMS - CASIS. Further it is planned to start others social statistical surveys like Labour Force Survey, EU-SILC and etc.

STATISTICAL DATA COLLECTION, PROCESSING AND DISSEMINATION processes, which are presented in Scheme 1, is the successfully working system, but some elements, like as Common dissemination data base (time series data base); Reference metadata base (SDMX)); the links between metadata bases at the moment are under construction or planned to be developed.

Scheme 1 - STATISTICAL DATA COLLECTION, PROCESSING AND DISSEMINATION.

STATISTICAL DATA COLLECTION, PROCESSING AND DISSEMINATION processes of CSB of Latvia are managed by 2 systems:

- 1.MDD ISDMS - CASIS
- 2.Data and metadata dissemination system.

1.MDD ISDMS - CASIS. This system is Integrated Statistical Data Management System - Computer Assisted Survey Information System for statistical data management, collection and processing purposes, which covers Computer Assisted Personal Interviewing (CAPI), Computer Assisted Telephone Interviewing (CATI), Computer Assisted WEB Interviewing (CAWI).

It is necessary to mention that despite that the majority of statistical surveys (including business and social statistical surveys) take place the operation cycle through MDD ISDMS - CASIS, nevertheless at present some statistical surveys are processed (including data collection as well) through this MDD ISDMS - CASIS partially or are not processed at all. For such cases, depending on a stage of individual processing of the survey, such means as - MS Access, BLAISE, SPSS, MS Excel are applied.

Main principles of MDD ISDMS - CASIS:

- 1.Centralization
- 2.Independence of individual programming
- 3.Integrity
- 4.Standardization
- 5.Transparency
- 6.Closer connection to respondents
- 7Automatic applications generation

- 2.Data and metadata dissemination system.

This system is foreseen for both: to maintenance the common dissemination data base and to 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.

Since 1 January 2009, the CSB of Latvia has introduced a reference metadata repository that describes contents and quality of statistical data. The Project's Documentation System (ADS) includes the following information on surveys and calculations of the CSB beginning with 2008 annual statistics and 2009 short-term statistics in Latvian:

- 1) structured descriptive information according to the production process of statistics (identification of data demand, project preparation, data collection, data processing, data analysis, data dissemination);
- 2) ESS quality and performance indicators;
- 3) Thesaurus (definitions of statistical indicators).

Currently ADS is accessible just for internal users. It is planned that selected information will be accessible to external users in mid-2011.

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.

Metadata system(s)

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.

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.

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

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.

Metadata Management Tools

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.

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

Version control and revisions

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.

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.

Sharing software components of tools

Overview of roles and responsibilities

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

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

Metadata management team

.

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.

Partnerships and cooperation

.

Other issues

.

Lessons learned

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;

Links:

Attachments