WP 20

29 January 2010

ENGLISH ONLY

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)

MANAGEMENT OF METADATA IN INTEGRATED STATISTICAL INFORMATION SYSTEM (ISIS)

Giedrė Vaišnoraitė, Statistics Lithuania (giedre.vaisnoraite@stat.gov.lt)

I. INTRODUCTION

1. The strategy of Statistics Lithuania for 2008–2012 gives sufficient attention to the management and development of information and communication technology. The main objectives are to interconnect information systems of Statistics Lithuania into a flexible and safe information infrastructure, to provide conditions for better management and more rational use of processes and resources; to completely interconnect partial processes of statistical information development: collection of statistical data, processing and dissemination of statistical information.

2. Aiming at implementation of these objectives, the basic goals of the strategic period are the introduction of an Integrated Statistical Information System (ISIS), the use of Information and communication technologies (ICT) potential for improvement of public services and strengthening the relations with users of statistical information and respondents.

3. The aim of this paper is to acquaint participants with the structure of metadata in ISIS and show how metadata are managed.

II. OBJECTIVES OF ISIS

4. A new Integrated Statistical Information System (ISIS) was initiated at Statistics Lithuania in 2004. The key drivers for the system development were:

- a mix of centralized and custom-developed IT solutions for statistical systems;
- obstacles in merging data from different surveys;
- efficiencies to be gained from standardized software (especially for common statistical processes);
- need to use metadata and common classifications;
- efficiencies to be gained from modern technologies.

5. The project was oriented to the re-engineering of the applications designed for base statistical information production processes and did not covered electronic data collection and dissemination areas. Specific project tasks were:

1) Statistical surveys area: to standardize survey preparation, data collection, processing and dissemination processes; to minimize the number of developed custom survey applications; to improve the quality and to reduce the costs and time of statistical data processing; to improve the quality of collected statistical data; to provide opportunities for respondents to submit data in various forms and ways; to use administrative data more efficiently

2) Communication, improvement of staff activity and management area: to ensure the means of monitoring; to create conditions for incident management and analysis; to facilitate communication between the survey teams

3) Databases improvement area: to improve the accessibility of statistical data and information; to integrate all databases into a common system; to create a common metadata base integrated into the system; to reduce data duplication in the system; to enable saving intermediate states of data; to enable data archiving

4) General system functionality improvement area: to provide the means for the staff of regional statistical offices to work with the system; to ensure integrity with the existing components of the system and those planned to be developed; to develop software satisfying the system's objectives and users' requirements; to implement hardware and software necessary for proper functioning of the ISIS; to implement means to improve data reliability, integrity, confidentiality.

6. The first surveys were started to be implemented into ISIS in 2007.

III. STRUCTURE OF THE SYSTEM

7. Main ISIS components:

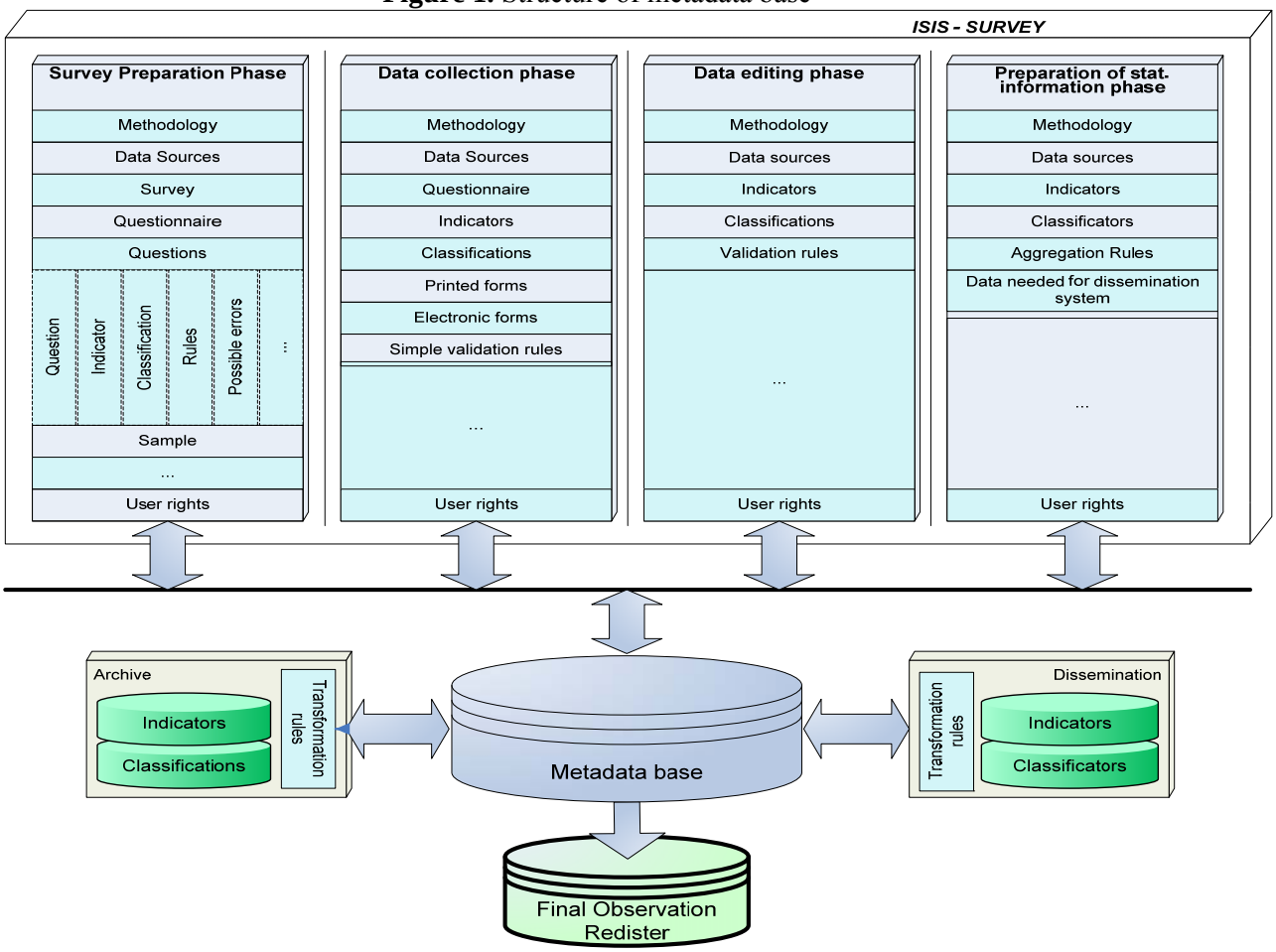
- Core Metadata Repository for Storage of unified metainformation of all statistical surveys (variables, code lists, editing and tabulation rules, etc.);
- Software for metadata management;
- Uniform software for data manipulation using metadata (survey management and operation);
- Interfaces to basic registers and Dissemination DB;
- Interface to Electronic Archive;
- Processes and results quality management tools.

8. Metainformation base ensures the systematic use of metainformation inside and outside the statistical information system and provides tools for the security of the internal and external integration of the ISIS. Metadata base includes methodology, data sources, questionnaire, indicators and variables, classifications and code lists, metadata quality, validation rules, electronic forms and printed forms. The statistical metainformation base is focused mainly on the process of creation and provision of statistical information. The model of metainformation determination of statistical data

ensures a unified description of statistical data in all stages of this process (from the formulation of the statistical survey to the dissemination of statistical information to the users).

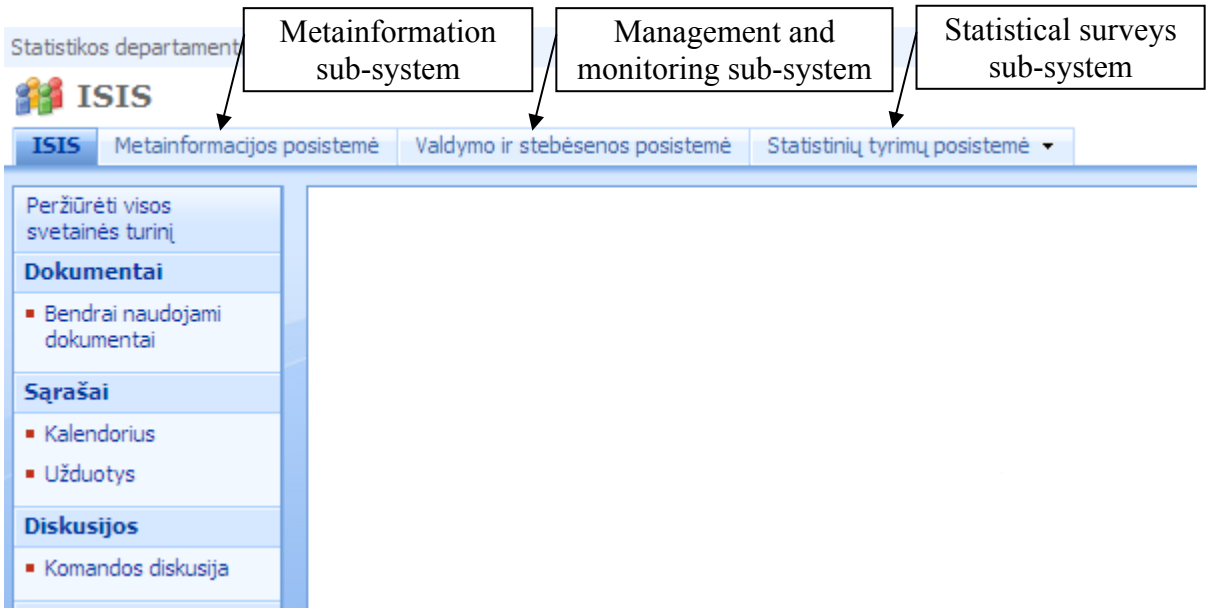
9. Structure of metadata base consists of four main processes (see figure 1): survey preparation, data collection, data editing and preparation of statistical information.

Figure 1. Structure of metadata base



10. Integrated Statistical Information System consists of three main sub-systems (see figure 2): metainformation sub-system, management and monitoring sub-system and statistical surveys sub-system.

Figure 2. Main sub-systems of ISIS



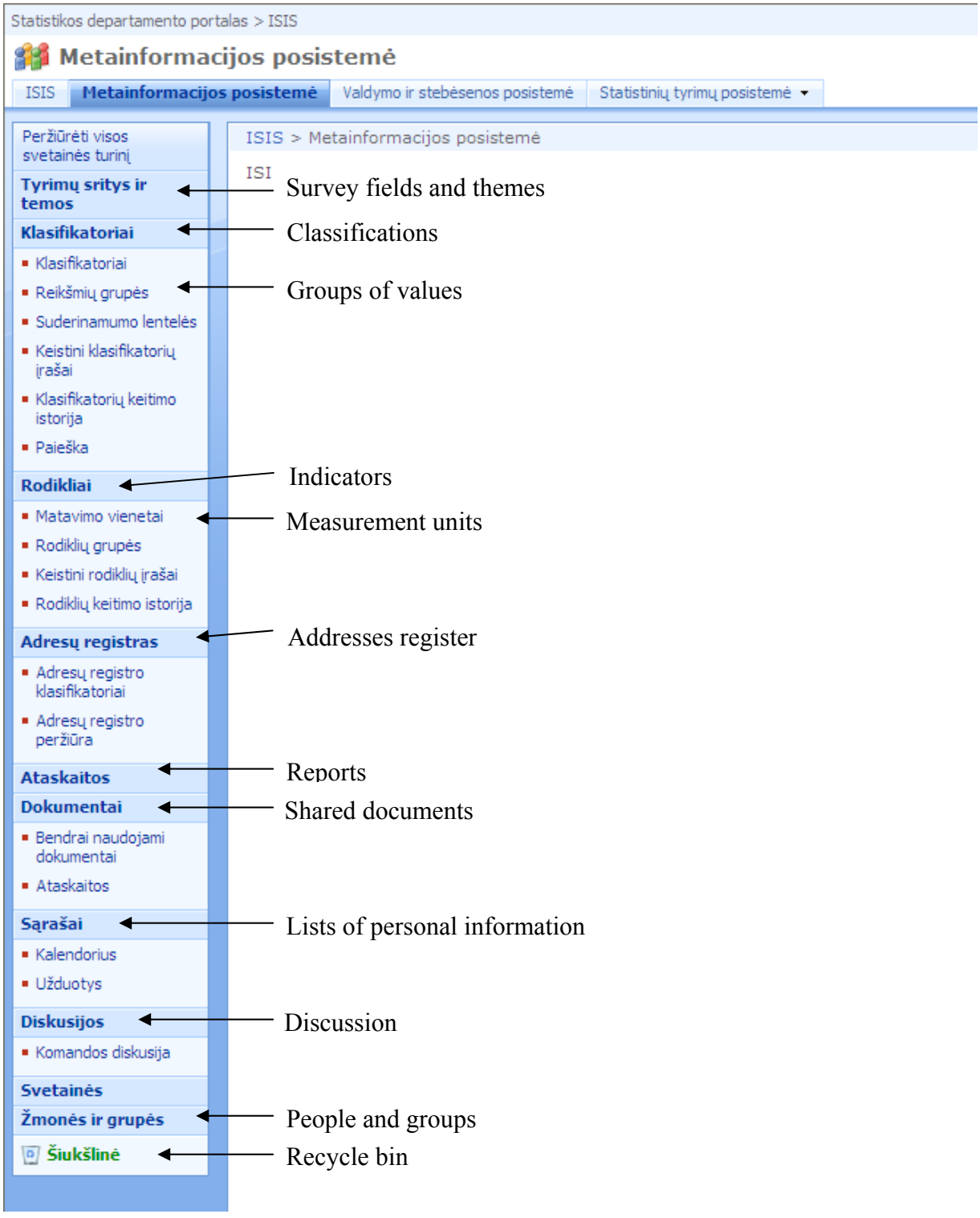
11. Management of the metadata will be discussed in a more detailed way: indicators, variable, classifications, preparation of a questionnaire template, validation rules, electronic forms and printed forms, data entry/import, validation and export.

IV. MANAGEMENT OF METADATA IN ISIS

12. Statistical survey data processing begins with survey Metadata entry into the Metadata base. Each new statistical survey should be registered in the system. For each survey it is necessary to encode its own metainformation, i.e. first of all, to enter basic information about survey (name, periodicity, indicators and variables, classifications and code lists, to create questionnaire templates, to describe validation rules, to associate electronic form with questionnaire).

13. Metainformation sub-system is represented in figure 3, which consists of classifications, groups of variables, indicators, measurement units, addresses' register, etc.

Figure 3. Structure of metainformation sub-system



There is possibility to edit, delete, import and export metadata.

14. Metadata concepts are categorized into sections or areas according to the logical relationships between them. These sections can be described as follows:

- Classifications;

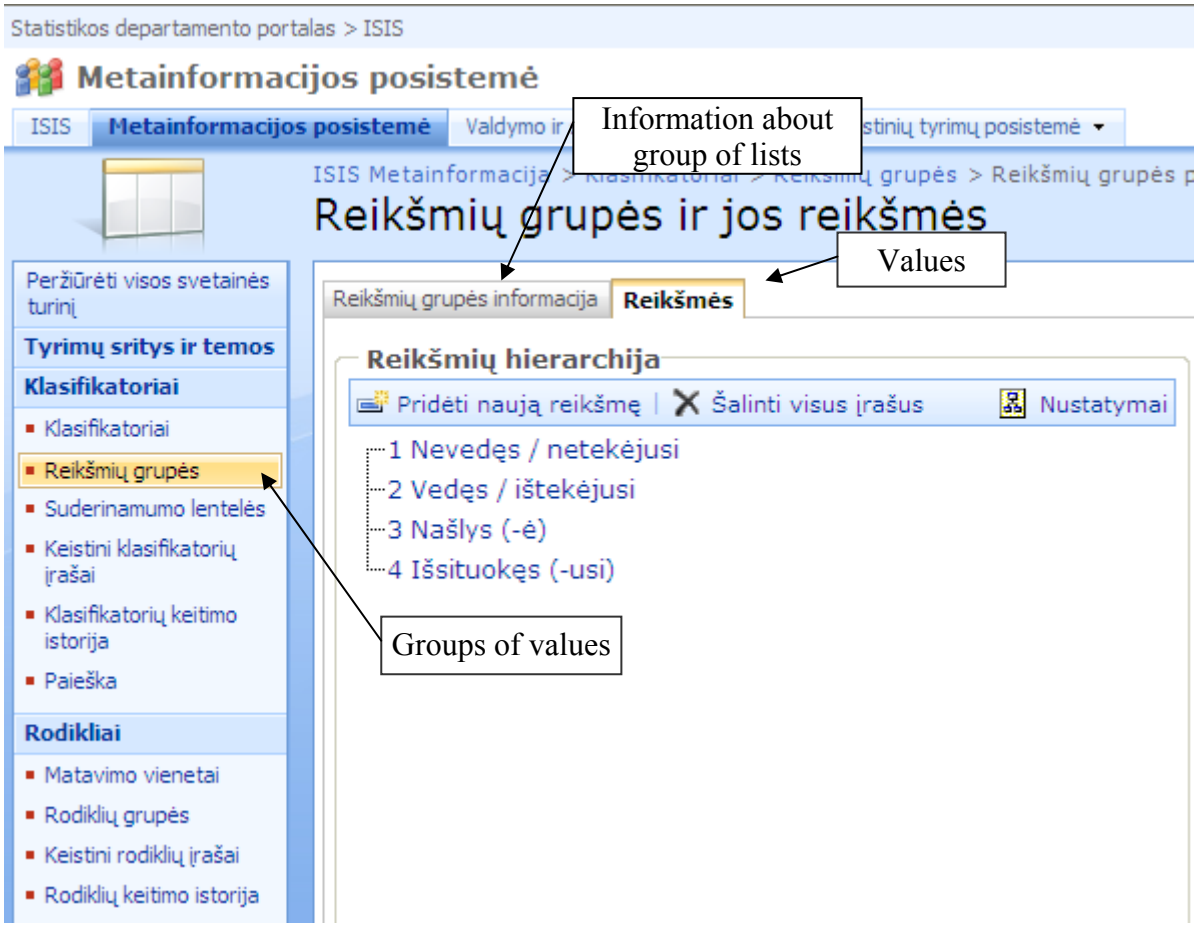
This section contains all metadata concerning classifications (National, European and International classifications are catalogued and most of them are stored in the Metadata base).

Classifications are stored in a centralized way and can be downloaded. Examples of metadata concepts in this section are Classification family, Classification, Classification version, Classification item, etc.

- Groups of values;

Value used in the system is a value of indicators for the exact item of classification. Created values are attached to survey. Same as indicators, there is possibility to select existing values from common group of values and to attract it to several surveys.

Figure 4. Metainformation sub-system – Groups of values



- Indicators;

This section contains all information about primary and aggregated indicators (see figure 5). They are stored in Metadata base in the common group of indicators. Indicators are characterized by code, name, periodicity, measurement unit. It is possible to describe new indicator, also to select existing indicator from common list of indicators and to attract it to several surveys.

Figure 5. Metainformation sub-system – Indicators

15. It is possible to see indicator hierarchy: child and parent indicators, which have relationship with current indicator. All indicators have codes harmonize with Statistical Data and Metadata eXchange (SDMX).

16. In our case, structure of indicators is "3-X-YY-ZZ-VVV-UUU-NNN-MM-…", where first level "3" means statistical domain "Environment and multi-domain statistics", second level "X" means statistical areas within domains, third level "YY" describes group of Statistical survey, fourth level "ZZ" describes subgroup of Statistical survey, fifth level "VVV" describes number of statistical survey, sixth level "UUU" describes number of indicators' group of specific statistical survey, seventh level "NNN" describes number of indicators of group "UUU", eighth level "MM" describes number of sub-indicator and so on.

17. This classification provides a useful high-level scheme for organizing metadata in many types of applications.

- Other metainformation in ISIS;

Next step in the survey content layout description is preparation of a questionnaire template. Users create data entry templates themselves. There is opportunity to generate questionnaire templates of e-questionnaires; possibility to have versions of created templates.

When questionnaire templates are created, users can descript data validation and editing rules themselves using the tools embedded in the system. There is an opportunity to form/describe respondents' and questionnaires' packages, allocate tasks to operators. Received questionnaires are registered (mode and date of submission). There is an opportunity to record the history of communication with the respondent.

- Data entry/import, validation and export;

Data import can be performed by users from various formats. All data entry forms of different surveys have a common user interface. There is an opportunity to view the scanned questionnaire and data of other periods or surveys. Users can view data errors while editing history. Data validation can be performed for all or part of microdata. A responsible user can export microdata from the system to various formats.

- Aggregation and preparation for dissemination;

For microdata aggregation, the prepared tabulation rules are used. Users can form tables and analyze data. There is an opportunity to import an out of system formed table. Users can harmonize code lists of formed tables with code lists of the Data Dissemination System.

V. MONITORING QUALITY OF THE PROCESS

18. One of the project objectives is to improve the quality and to reduce the costs and time of statistical data processing and to improve the quality of collected statistical data. Metadata on data quality and processing are stored in a common database.

19. Quality indicators of the statistical production processes are characterized by code, name, description of methodology, quality unit, person responsible for measurements, periodicity, expire from and expire end dates. There is possibility to specify calculation algorithm and formula of indicator, calculation periodicity.

20. For all quality indicators of the statistical production processes, critical values are defined, which allows managing risk, identifying and solving problems on time.
Users have an opportunity to define and form the required quality report. They can analyze obtained information at different levels and in different ways

21. There is opportunity to see calculated quality indicators in diagram, i.e. representation of the current situation.

22. Users can create quality indicator templates themselves and save in the system (XML, HTML, PDF, etc). Description information of statistical survey and quality indicators should be kept in Lithuanian and English languages.

VI. METADATA PREPARATION FOR LONG-TERM STORAGE

23. Data from the surveys elder than two or three years are stored in archives. To ensure high and stable performance of the system, that archived data is useable by future generations, we need to include not only the data itself but also information such as the survey indicators and variables definitions, classifications, questionnaires, quality indicators, definition of methodology.

24. Each statistical survey has possibility to create a new description of archive, to edit its information by indicating the archiving settings, to archive data and to delete micro data.