Distr.
GENERAL

CES/AC.71/2004/20
30 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

INTEGRATED META DATABASE AS A BASIS FOR DEVELOPMENT

Supporting Paper

Submitted by the Central Statistical Bureau, Croatia[1]

I. INTRODUCTION

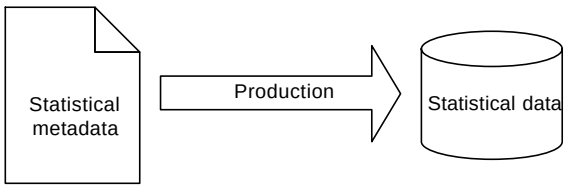
1. The aim of this paper is to describe the general model of the Integrated Statistical Meta database and its use in statistical survey processing within the Central Statistical Bureau of Croatia (CROSTAT). The model is under development but partially implemented.

2. Placing metadata in the central role of development illustrates our belief that the development of a statistical system is the development of the metadata. If metadata exists or is organized in the form of a database, this could define a good basis for development of the whole statistical system. A centralized repository of standards, classifications, methodologies and other subject matter knowledge allows for the faster development of the statistical system and its immediate availability for public use.

II. GENERAL ARCHITECTURE

A. Main elements

3. The very basic picture that explains our approach is very simple and as we believe gives the shortest possible description of the use of metadata driven development. In the logical sense all of the necessary



[1] Prepared by Zdenko Milonja (zmilonja@dzs.hr)

information for statistical data production is stored in the metadata, of course, the traditional method of storing is not appropriate for any type of automatic processing. Metadata is stored in books, papers, minutes of meetings, official and unofficial documents, sometimes as general knowledge of the experts. This form of metadata is not suitable for automatic processing and we would like to develop an integrated meta database that will contain all the necessary data for production and dissemination.

4. The general architecture comprises several main elements in the integrated meta database. It consists of a description of the organizational structure, annual plan and publishing calendar, survey descriptions, classifications and some other elements. Beside the meta database stands a statistical database, in micro data and macro data versions. The third element is the statistical production system that connects statistical data and statistical metadata.

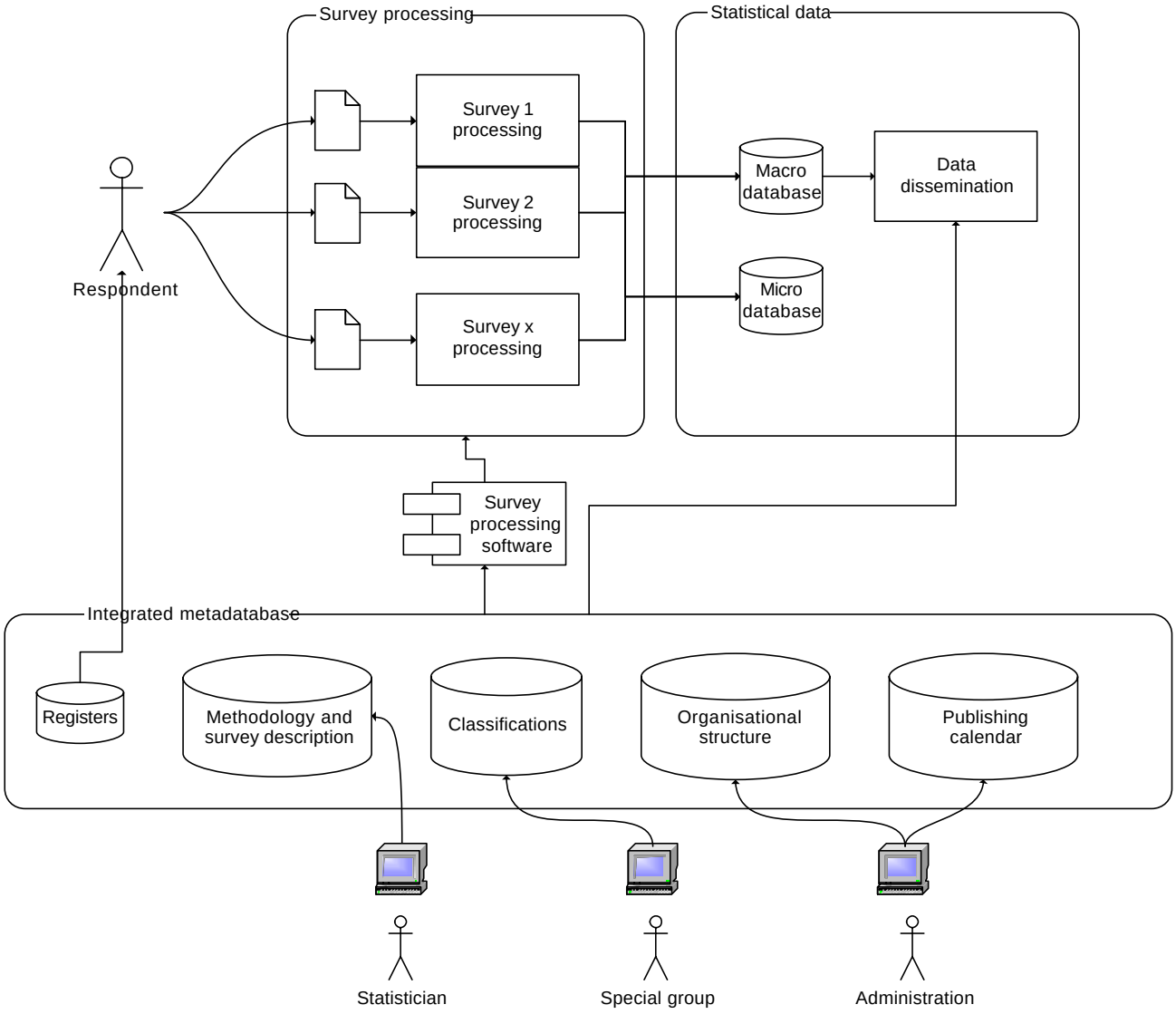


Figure 1. Basic system components

5. The importance of information technologies in statistical survey processing and dissemination is increasing and consumption of the resources is significant. Through the integrated meta database we open

the possibility to increase productivity in various parts of IT implementation. Separating the data from programs, but also metadata will probably lead to more adaptable programs.

6. The essential question we would like to inspect is the productivity of software production in the statistical survey environment. Under software production we mean the production of necessary software to make the final observation register. This software is always tailored to the questionnaire and depends very much on the data collected. Software for statistical processing (SAS, SPSS, etc) already exists as general purpose software not connected to a specific questionnaire and its data.

7. The software for "questionnaire processing" includes data validity checking which depends on the characteristics of the data on the questionnaire. The data checks are described by the Boolean formula that is stored in the meta database together with other methodological explanations. This separation of metadata as a feature of survey allows us to create the program without any data connected with specific questionnaire.

B. Basic components

8. The content of the metadata is spread over different departments in the statistical office. The methodological issues comprise the main part of statisticians (subject matter specialists) work. This work is probably the most important and most popular part of the metadata. We consider this as the metadata. The two most important segments are semantical and technical metadata. Semantical metadata is, for example, a statistical object or statistical variable. The technical metadata is, for example, the number of decimal places for a certain statistical variable.

9. The second important segment of the metadata is the classifications and coding lists for qualitative statistical variables. Classifications as metadata must be separated from surveys because it is important that if two surveys use the same, or sometimes similar quantitative variable, they use the same classification. The use of the same classifications gives us the opportunity to connect two surveys (two data sets) into one data set and the possibility to compare the separately collected data. This provides us the basis for making the statistical surveys without collecting new data. There are some classifications that are not connected to a specific survey but are used by two or more surveys (NACE as typical example). Those clasifications are maintained by a specific group to prevent the adaptation of the classification to specific survey needs and thereby loosing the possibility of connecting the data from different surveys.

10. The next component of the integrated meta database is the organizational structure. The organizational structure is important because the surveys are performed by certain departments, there is a work flow among many participants who are in charge of the various phases of the whole process. All complex tasks need planning and the data about every participant in the process is stored in the database. This provides the basis for planning activities in statistical surveys production.

11. In the planning process after recognizing who is in charge of running the statistical activity, the second component is time. The time dimension is usually based on the publishing calendar. To be on time according to the publishing calendar means activities are going smoothly and there is no need to act in any direction. The difficult question is how we know that everything is going smoothly and will be on time. It is important to foresee the unexpected delay as early as possible. Comparing actual dates with the publishing calendar will give us the early warning to act. The publishing calendar includes all the obligations for data delivery regardless of the fact that the data is published for public use. This could mean delivery to IMF or EUROSTAT.

III. BASIC SYSTEM FUNCTIONS

A. Creating and maintenance of organization structure

12. The organizational structure of the institution is the basic metadata concerning planning and management process. The organization structure can be shown on diagram:

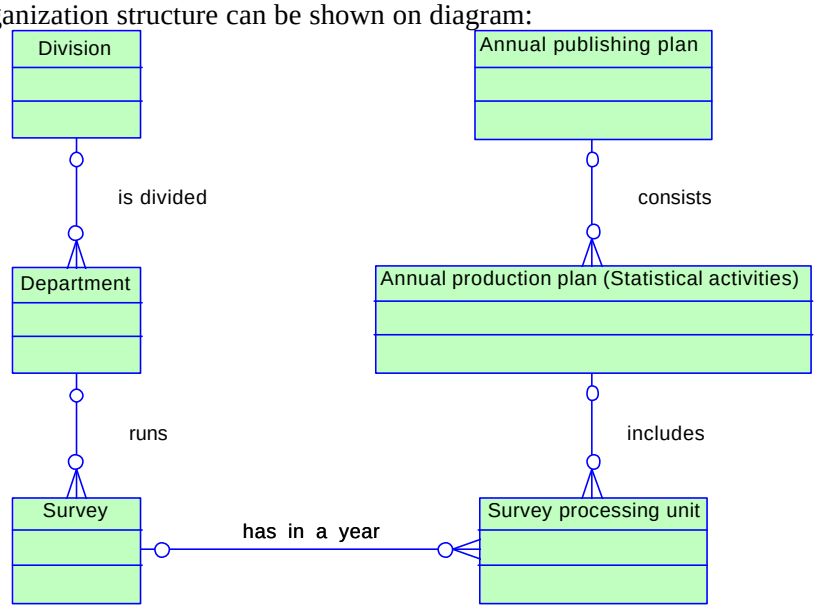


Figure 2. Organisational structure in the database

13. The organizational structure in CROSTAT comprises divisions (of which there are 8), and departments. There are 3-5 departments in a division. Not all of them are included in statistical activities but the data exist in the database. All surveys are in the database and each survey is attached to the department which is in charge of running the survey. In the database there is the difference between the survey and survey processing unit. For a monthly survey, we say there are 12 survey processing units.

14. Collection of survey processing units defines the annual production plan (collection of statistical activities). There are some surveys in the preparation stage or major restructuring phase. For such a survey where there is more methodological work rather then the usual collecting and processing of the data, there is still a survey processing unit, but without repetition and the task is, for example, to produce the revision of the methodological manual. Some of the statistical activities are involved in production of the published output.

15. The planning system is based on planning data on survey processing units. Various data are stored according to: planned date of the data collection, planned date for finishing data validation, etc. During actual processing of the survey processing unit the actual data are collected and stored into the database. It is possible to compare the planned data with actual data and get the status of work in progress.

B. Registers and respondents lists production

16. The most important register, the register of business entities is currently under development in CROSTAT. We plan to develop the statistical business register as a separate organizational unit which will provide all surveys with data for survey sample frame. Various respondent lists will be extracted from the register and the data collected from respondents will be used to update the register. The data in the register will be used as a part of the survey data but the register maintenance will be conducted in special organizational unit.

C. Classifications maintenance

17. Classifications are also considered as a part of the meta database. For the classifications database, we use the BRIDGE system which is developed according the Neuchatel group recommendations. Classifications are extracted from the surveys and stand as a separate task with separate organizational unit. In fact, if a survey is the only user of the classification, this survey will maintain the classification, but if there are more users for the same classification it will be maintained by a special unit. Through application of this principle we hope to achieve better homogeneity of the different surveys.

D. Planning

18. In the planning process we can recognize two separate phases. In the first phase we produce the data about planned activities and time when we expect them to happen. We developed the planning module that will be used for entering the data about planned activities in the next year.

19. The second phase is collecting the data about real events and actual time. Normal production in the statistical office can be divided into projects such as activities and survey processing. Survey processing comprises activities that are basically the same in each period. The data about dynamics of processing will be captured mainly automatically during the process of survey life cycle.

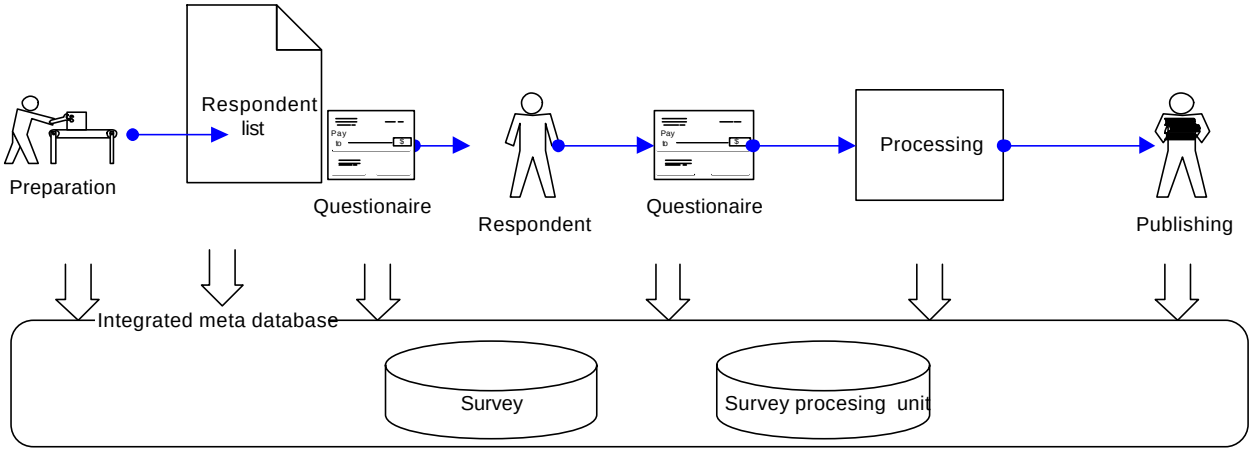


Figure 3. Planning data collection

20. Having collected data about the planned and actual time for the processing activities, it is possible to compare these values and develop information for operational management to keep the production within the set time limits as much as they can. The strategic management have the opportunity to analyse the collected data find the possible improvement of the survey life cycle.

E. Software development

21. Software development is a costly and time-consuming process. In the statistical office there is always a need for software tailored to specific needs for certain surveys. Data validation and editing is dependent upon the content and layout of the questionnaire and usual practice is to develop a specialized software for such purposes. The bad feature is that any change in a questionnaire requires a change in the software. As we know the need for change is not a rare event, it is an everyday practice and is characteristic of modern technology and modern society.

22. There are two possible approaches in handling this problem:

- to develop one program which is so flexible that it is able to process all of the requests from every survey. Any new survey could be added to processing without the need to change the program.
- to develop a program which is able to generate another program automatically, by machine itself, and this program will be capable of processing the survey.

In both cases there is a clear conclusion that must be a meta database which has the data which is needed by the program. Regardless if the program is so "universal" (1. option) to process the new survey and take the data (from the metadata) to "reorganise" itself, or, the program will produce another program (2. option), which is tailored to the new survey and get the data from the meta database.

23. Having that in mind, the important part of the meta database includes a "technical" description of the survey.

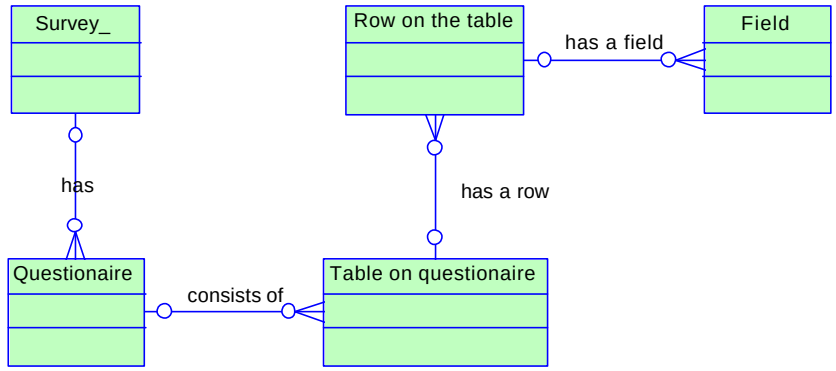


Figure 4. Technical metadata for questionnaire structure

The "technical" part of the meta database says that there are surveys which can have questionnaires. On the questionnaires there are some tables that have rows and in each row there is one or more fields where the respondent could put the data. Generally with some attributes this describes the "technical" structure of statistical questionnaire.

F. Methodological and semantic metadata

24. Together with technical metadata in the database, there is a space for semantic metadata. This metadata is the basis for putting the methodological work into the database. We develop the database schema for describing the semantic part of the survey.

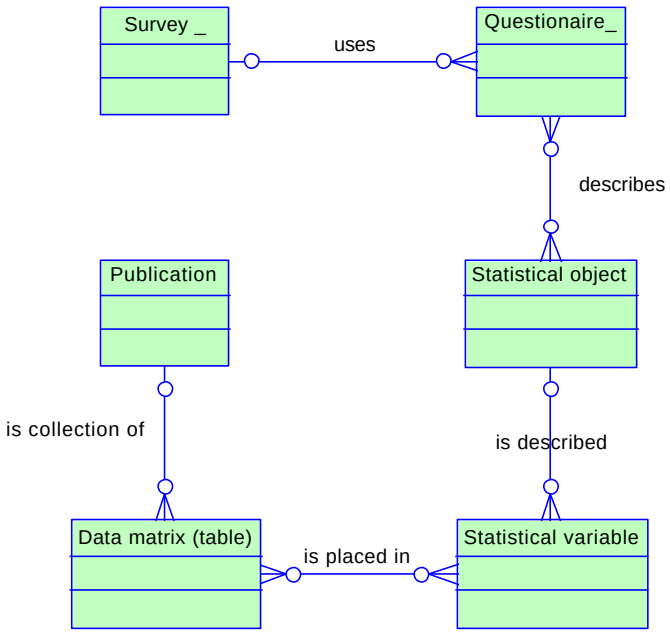


Figure 5. Semantical metadata for questionnaire structure

25. The semantic part is organized around the statistical object. The statistical object "exists" on each questionnaire, at least one statistical object, but it is not unusual that a few statistical objects are described by one questionnaire. Of course, one statistical object could be present on several questionnaires. The questionnaire belongs to a survey. The statistical object is described by one or more statistical variables. One statistical variable can be attached to only one statistical object even it is the "same" variable. The colour of the car is something (semantically) very different from the hair colour even though we speak about "colour".

26. Statistical variables are placed in data matrices (tables) and again one statistical variable could be placed in several tables and one table consists of several statistical variables. The publication we describe as a collection of matrices.

IV. PROJECT STATUS

27. We are aware that the large and complex projects cannot be developed very quickly. The main idea was not to achieve a very rapid development of statistical system but we planned to draw a base line for future development for a long period. Our major concern was to establish a long lasting platform for development because we would like to save the money invested in software development.

28. At the moment we develop some parts of the overall project according to CROSTAT priorities. We have in-house software development and consultancy support from Statistics Sweden.

V. ACKNOWLEDGEMENT

29. We wish to express our gratitude to the Swedish Agency for International Cooperation (SIDA) and Statistics Sweden for their support and active involvement, as well as substantial help from the Central Statistical Bureau of Latvia and the Swiss Federal Statistical Office.

- - - - -