

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (iii): New methodological and technological developments in statistical data editing

**NEW DEVELOPMENTS IN THE DATA COLLECTION TECHNOLOGY IN THE CZECH
STATISTICAL OFFICE**

Submitted by the Czech Statistical Office¹

Contributed paper

1. In the course of the last year, the DataMan system has been developed to make it compatible with the Oracle database and to make the product stable. A DataMan version has been created which may work both with its own data format and with the Oracle database engine, converting one data structure to another as needed.
2. The development of ProjektMan during the last year was mainly dealing with the finalisation of the system in order to automate the connection between the user and the programmer. Now the main development is focusing on the implementation of the system in the population census in 2001. As the preparation for this action is currently going on, the functions described in the paper are in the planning stage, and not yet implemented and verified in practice.
3. The population census data collection and editing is planned to be carried out in the following phases:
 - a) Collection of the completed census questionnaires;
 - b) Scanning and recognition of questionnaires;
This activity will be carried out by an external company. The questionnaires are scanned and converted into electronic images at a specialized workplace. Optical Character Recognition (OCR) programs are used to convert electronic images into digital text form. Links are maintained between the OCR data and the graphic images of questionnaires.
 - c) Storing images and converted data in the database;
It is planned to create two databases with the same structure: an input and an output database.
 - d) Data checking and corrections;
 - (i) Input database - comprises the OCR data and graphic images. It is used for checking procedures and data correction activities. It keeps the original data, i.e. each adjusted figure has a shadow item behind, where the original value is stored. Clean, error-free data are transferred into output database.

¹ Prepared by Dusan Loutocky.

(ii) Output database -includes only corrected OCR data. It does not contain images. Data stored in this database must not be changed. If the data needs to be corrected, the checking procedures and corrections are run again on the input database and clean data are reloaded in the output database.

4. The checking algorithms are defined in the ProjektMan special language. A semifinished product in the Oracle (PL SQL) language is generated, using the output from ProjektMan as a source. After inclusion in the program frame, a program is generated which can process data from the input database on the central server. This activity may be carried out interactively or in batch mode. Some incorrect items are corrected automatically, some are gathered into a package for visual control. The completeness of the questionnaire, correctness of identification keys, completeness and permissibility of data values and validity of links among attributes are checked. The checking process results either in the correction of the errors according to a prescribed procedure or, when the detected errors require human evaluation (if a graphic image has to be looked through), incorrect data are sent for interactive corrections.

Data corrections at specialized workplaces

5. This activity is interactive and is performed by trained specialists at interactive correction workplaces, the number of which is estimated to be about 70. Each workplace comprises one PC connected via LAN to the central server where the input database is kept. The workplace has access to data and procedures stored in the database. The interactive control stage follows the automatic checking process only when individual records are flagged for interactive corrections. The process is controlled by the computer, i.e. for various error levels the correction procedure is programmed in advance, the staff only verifies the values.

6. Data corrections requiring visual control of graphic images are carried out at the interactive workplace. The digital value, image and data from classification tables, are displayed concurrently. Suspicious data are marked. Different colours are used to indicate if the data item is in the original questionnaire, or if the data item was corrected manually, or corrected by the auto-correction procedure.

Automatic correction

7. Automatic correction is carried out by a program which automatically eliminates errors that cannot be corrected at the interactive stage. This operation can affect the data quality more than the manual operation. It is performed in a batch mode and represents the last part of the correction activity. The auto-corrections are performed after making all manual operations. The program performing auto-corrections will be described in the ProjektMan special language, after which the semifinished program product will be generated in the PL SQL language.

Creation of outputs

8. Creation of outputs is the last part of data processing. The data will be processed using both the database tables containing input questionnaires and the tables which include derived data.

Use of ProjektMan and DataMan systems

9. ProjektMan will be used in all activities where the collaboration between the user and programmer is necessary. It concerns, first of all, checking procedures in the course of input processing. The creation of checking programs means the construction of algorithms which check data validity and follow links among individual data items. The ProjektMan allows to describe the control procedures in a standard format; its adjustment concerns only changes in the target language. Other activities realized in the ProjektMan include an auto-correction program, a program for creating derived data and finally a program for the creation of outputs. All these activities result in the generation of a semifinished program product. We suppose that its next adjustment will be manual. At present we are unable to predict to what extent this manual work will be required.

10. An obvious precondition for all these activities is to convert ProjektMan to the 32-bit environment. As the DataMan and ProjektMan systems represent a substantial part of the technological environment in the Czech Statistical Office, we expect their further development and greater linking with the technology of statistical surveys and at the same time further harmonization with UN and EU standards.