Data entry and editing of Slovenian Agricultural Census

Pavle Kozjek, Statistical Office of the Republic of Slovenia

1. Introduction

Agricultural Census was one of the main actions of the Statistical Office of the Republic of Slovenia (SORS) in the year 2000. Data were collected by field interviewers on traditional paper forms. Scanning or other character recognition techniques were not used, and relatively large amount of data (more than 120 000 forms, each with 828 data fields) was a good reason to develop an efficient solution for data entry and editing in Blaise. Census applications have to meet some requirements (e.g. speed of data entry) that are not so important with some other surveys, but should be obtained for census data processing.

Agricultural census data entry and editing was processed completely in a local area network Windows NT environment, with Blaise 4 Windows as a main supporting system. This became possible with new hardware and software equipment installed in the data entry department at the beginning of the year 2000. The census data entry application was integrated into a new Blaise database management and administration system, developed at SORS for the general needs of high-speed data entry in a LAN Windows NT environment.

2. Data entry

Census data entry was not integrated with data editing. Main reason for that was traditional organization of data processing at SORS, with two specialized departments: one for data entry and another for data editing. Following this division of labour, the census application was made up of two main data models: one for data entry and another for data editing. Editing data model was developed first, checking the data and retrieving all external information necessary to complete the process of administrative data editing.

Data model used for high speed data entry was a "copy" of the data editing application - structure was the same, but without reference files, and only some formal data checking criteria were enforced. This "top down" sequence of development is perhaps unusual, but in this way compatibility and data transfer between entry and editing databases was ensured in a simple way. Data entry screens were adjusted to the needs of high speed data entry - they were designed as much as possible like paper forms, to help data entrists navigate between questions and fields on the screen. Access to data and applications on a server was enabled through the VB user interface (part of the generator for high-speed data entry applications), using login and password procedures. Administration of data entry was based on physical batches of paper forms, each containing about 50 forms. One batch represents one partial Blaise database that was later in the process added to the final database. This administration was based on generated Blaise datamodels and Manipula setups for handling different Blaise databases and ASCII files, using parameters entered through the user interface. Special attention was devoted to the end users: although some new components were used in the census data entry application, it took only a few days before data entrists learned to use it completely.

2.1 Role of the Generator of high speed data entry applications

The generator of high-speed data entry applications (GEntry)[1] is an in-house developed tool, made for two types of end-users:

- Data entry administrator specifies and generates data model (together with all setups necessary to complete the data entry process), implements it in a production environment and finally sends entered data to the archive and further processing.
- Data entrists use generated application to enter and verify data, using the method of double keying.

The generator is based on functionality of Blaise 4 Windows and Visual Basic. It enables non-EDP people specifying survey data models, and supports and organizes high-speed data entry, verification and process administration on a LAN. Of course, generator can only be successful and efficient with some in-house standards and conventions concerning design of paper forms and data models respectively.

With regular statistical surveys, the procedure of preparing and using the data entry application is highly automated. Some "special cases" need to be improved and corrected "manually". With census, the main data model was not generated, but the rest of the data entry process (administration of data and users) was automated like in the other surveys with high-speed data entry, using GEntry production environment.

3. Data editing

Compared to data entry, the data editing application represented another view of the census database. Blaise screen was different (interviewing mode, with basic meta information on the screen), all data editing rules were included and 4 reference files were used to check and to complete the census data: the address register of farms, the code list of farms, the code list of municipalities and the code list of house numbers. A primary key was defined (in data entry application only secondary) to obtain easy multi-user access to data forms, collected in a single Blaise database (more than 500 M). In the final stage some problems with response time appeared, probably also related to relatively large reference files (up to 500 000 records) and a number of criteria (632 checks and signals)[2]. The editing application also provided the possibility of data entry, for some cases where data were not entered by the high-speed application.

During data entry, some subject matter inconsistencies between collected data and SORS registers and code lists were discovered. Most of them were solved during the data editing process.

4. Findings and their importance for future work

In general, data entry and editing application functioned successfully, without serious problems and the job was finished respecting deadlines. Users (data entrists and data checking staff) accepted it very well.

---

[1] GEntry-Generator of high-speed data entry applications was presented on a 6th International Blaise Users Conference (IBUC) in Cork, Ireland in May 2000. It is used in SORS production since April 2000, and in May 2001 it covers about 80 % of SORS high-speed data entry (70 surveys).

[2] Repetitions of the same check on a different field are also counted.

This was the first time at SORS that department for high-speed data entry used form-based data entry instead of traditional record-based data entry (which is used with the old mainframe solution). It seems that users have no problems to accept it.

With census application tried to optimize the situation, where (due to specialized staff, or other reasons) high-speed data entry is used as the first step of statistical process, and data editing is the following step. We are aiming to integrate both steps, where it is applicable, but currently we have to deal with two specialized departments. The only exceptions are some important surveys running in a CATI studio.
Comparing to processes of other surveys with traditional high-speed data entry on a mainframe, some important advantages were noticed:

- Complete process (including high-speed data entry) is defined in a single (LAN Windows NT) environment - easier and more efficient process management, maintenance etc.
- Entry and editing processes can be separated or integrated (entering data directly into editing application)
- Less re-structuring: basically the same data model is used for entry and editing; at the same time it provides structuring and metadata for further processing (e.g. tables in a relational database)
- Form-based data entry provides easier and clearer specification of data editing rules (although it sometimes complicates data modeling).

Lessons learned are already included into the project of building SORS general data editing system, running in a LAN Windows NT environment and based on Blaise and VB. Process management is highly automated, and development and maintenance can be (in most cases) defined as set of rules and templates. This is extremely important when you have to deal with a large number of surveys and a very limited number of application developers.

5. Conclusions

The application for agricultural census data entry and editing represents one of the possible solutions how to integrate high-speed data entry into the statistical process in a LAN environment. At SORS we are using it as a reference for surveys, where high-speed data entry is followed by interactive data editing in Blaise. Rational approach to development, implementation and maintenance is highlighted: we are looking for the efficiency of the complete process of a survey, not just a data entry or data editing phase. On the other hand, solutions should be also general, to cover the wide range of different surveys and to include "special cases" in an acceptable way.

Furthermore, data entry and editing applications should provide an important part of metadata for the next steps of data processing. The project of building general base of SORS survey metadata is under way, and we are working together to find the best solutions. In this way we are trying to move step by step towards standardized and integrated statistical survey processing environment.

References

- Keller, W. J., Preparing for a New Era in Statistical Processing: How new technologies and methodologies will affect statistical processes and their organization, Strategic reflection Colloquium on IT for Statistics, Luxembourg, 1999.
- Kozjek, P., Blaise Generator for High Speed Data Entry Applications, 6th International Blaise Users' Conference, Cork, 2000.

- SORS and Statistics Sweden, Feasibility study on the architecture of information systems and related equipment issues, Study implementation and Hardware/Software Specification for Tendering, September 1997.
- Sundgren, B., An Information Systems Architecture for National and International Statistical Organizations, CES/AC.71/1999/4, Meeting on the Management of Statistical Information Technology, Geneva, 15-17 February 1999.