Blaise III: Changing the data model after implementation

Pavle Kozjek and Marko Sluga, Statistical Office of the Republic of Slovenia

1. Introduction

Recent development of statistical methods and techniques has opened many new possibilities in survey processing. Modern tools were developed to help statisticians and informaticians to get results - statistical information - in shorter time and with less effort. But, on the other hand, less and less time, money and human resources are available to get the same statistical information. Timeliness became a problem for many developers, and it usually results in lack of time to completely design and test the application. With CAI (particularly CAPI) applications this can lead to serious problems, when such an application is installed to many laptops, and some data are already collected and edited. Changing data definition means incompatibility of data files, which usually can not be solved without developers.
This paper describes one of the possible solutions for this problem, when the survey application is already in production. It can be used with the Blaise III applications. At the Statistical Office of Slovenia (SORS) it was successfully used with the continuos Household Budget Survey, conducted since February 1997, with the Pilot Census 1998 and some other surveys.

2. Slovenian Household Budget Survey - a short description

HBS in Slovenia has been carried out since 1963 on a larger sample (every five years about 3,200 households) and on a smaller sample (every year about 1,000 households). Since 1997 the Slovenian Statistical Office started with a continuos HBS. Every year about 1,200 (net) households shall be interviewed. Since this sample is too small for more detailed calculations, the data shall be aggregated together for 3 years. Every year the "oldest" 1,200 households shall be eliminated from the sample and the "newest" 1,200 households shall be included.
Becoming continuos, the survey was radically changed, with new survey and questionnaire design, new number of interviews, diaries were introduced and complete data processing was defined in a different way. A pilot survey (which was contemporary a Blaise III test survey at SORS) was conducted in October 1996. In February 1997 we started with the continuos HBS.
Compared with the annual Household Budget Survey, the continuos HBS was expected to have some advantages:

- Always current data of higher quality and lower costs
- Better organisation of the fieldwork and methodology of the survey because of the continuity
- Better trained and experienced interviewers
- More detailed results, derived from more subsequent years aggregated data

The main questionnaire is divided into two parts and interviewers visit each household twice. At the first visit, the first part is asked and the diaries (two types on paper forms, recording period 14 days) are explained. After two weeks the second part of the main questionnaire is entered and the completed diaries are collected.

3. The survey application design

From developer's point of view, the main problem with the Household Budget Survey was lack of time for application development. HBS is one of the most complex surveys: paper questionnaire consists of 72 pages and over 3,000 possible fields to be entered. The final version (February 1997) of the CAPI Blaise instrument consisted of 98 screens, mainly tables. But there were only about four weeks and one person available to develop the Blaise instrument, together with the entire case management.
Processing of the HBS main questionnaire data is completely independent from processing of diaries. The main questionnaire is processed in Blaise III as a multi-mode survey: there are 22 interviewers with laptops, using CAPI instrument and 4 data entrists, entering data from paper forms (PAPI/CADI). After coding of PAPI/CADI data (CAPI data are coded by interviewers), all data files are combined into one file and sent to SAS system for further processing.
CADI application for diary processing is still a Blaise 2.5 application. Coding is the most important part of it, so all three available types of coding are enabled: alphabetical, hierarchical and trigram.

If we wanted to follow the terms and to realise the education for interviewers, the application had to be installed to the laptops on the same day the last question was coded into Blaise language. So absolutely no time for testing was available. Furthermore, due to lack of time only the most important edits were included in the first production version of the Blaise instrument.
First small errors were discovered and removed already on the first day of interviewer's training. Since the second training day was a week later, time was used to improve and re-install the application and interviewer's interface. The interviewers were expected to start data collection and editing immediately after the second meeting, so that should be the final version.
But some more errors were found after the second installation, mainly not very important for data collection and editing (e.g. some unnecessary questions with an "empty" attribute were on the route). More important was the absence of some edits, which should be performed during the interview, but were not included in the first production version. Interviewers could start their job in time, but it was really necessary to improve the data entry and editing application.
After a month the HBS Blaise instrument was finally satisfactory and tested enough to prevent unexpected surprises. Some questions were removed from the route and many new edits were added. Of course, this changed data definition, and already collected data were not compatible with it.

4. Re-installing the survey application

When re-installing the HBS application we followed the instructions from the Blaise III Developer's Guide. Due to some difficulties in newly established modem communications between interviewers and the Statistical Office, we decided to send a new installation diskette to each interviewer, together with short printed instructions.
We were a little bit afraid of data conversion because of complexity of the questionnaire and especially because of additional interviewer's burden. Having this in mind, we prepared the re-installation procedure very carefully and made it as simple as possible, a real "push-one-button" solution. Interviewers were only asked to start the batch procedure, which installed some files and the new main interface menu, where they had to choose the new option called "Conversion to a new data definition". This option executed the following steps:

1. Creating the OLD directory and copying the existing data files into it
2. Installing the new improved survey application to a working directory
3. Installing and executing of Manipula conversion setup (from old to new data definition)

The Manipula setup to convert data from the old to the new data definition can be generated automatically any time during survey processing on the basis of metadata definitions, which are always present. Therefore, the conversion procedure can be executed more than just once, if necessary. Automatic linking of different data models is the key of the entire procedure: development of the conversion application would be much too time consuming.
When working, the interviewers use the new application exactly in the same way as the old one. A backup version of old data and metadata was automatically saved in a separate directory.

Conversion to the new data model in production was quite successful. Only with two interviewers problems occurred, mainly because they didn't exactly follow the instructions. Re-installation procedure was developed in a such way, that it was possible to repeat it without loosing the necessary files, so also these problems were solved relatively fast.
With the new data definition and new editing rules, some errors appeared also in previously clean records and the interviewers had to check all the records again. Since changes did not require another contact with already interviewed households, it was not too hard a job.
The same procedure was performed on the centralised part of the survey application, on PAPI/CADI data files at the Statistical Office. The old and the new data definition were the same as in CAPI instrument, and new data files from both modes of data collection were combined into one file again. New SAS setup was produced from the new data definition and data were sent to further processing.

5. New possibilities for application developers

A unified strategy for an improved editing process starts at the design phase and ends only when the final results are published. The desired final result is high quality statistical information. The aspects of quality of statistical information are accuracy, timeliness and costs[6]. It's not hard to conclude that the possibility of changing the survey design after implementation could improve especially timeliness, which usually affects also the other aspects. And which benefits we can get from using this possibility? We found some from our experience with the HBS:

- The survey data collection can start in time even if the data model is just roughly designed and tested
- The subject matter people can make last corrections in survey contents even after implementation (if really necessary)
- Interviewer's remarks can be considered and used in the final version of the application
- Edits can be added or removed during the survey data processing
- Development and production of the survey can run concurrently

There are some disadvantages, too:

- Additional interviewer's burden (even if small...)
- Adding new fields usually means repeated interview
- More survey administration

In case of our HBS the advantages were clearly prevailing, so this approach or a similar one will be probably used also in some other surveys at SORS, especially when rapid application development is necessary.

W.A.M. de Jong: Designing a Complete Edit Strategy; Combining techniques

The same approach was used with the 1998 Pilot Census. Like in the HBS, it was impossible to anticipate and define all the relevant edits before data entry had started. Since it was a CADI survey, data conversion between different models was relatively simple.

Both applications fully utilised the possibility of ASCII-relational reading blocks of data out of Blaise, which enable efficient loading of Oracle database and analysis of individual blocks of data. This is specially important with large data models. Concerning relations to other systems, this option seems to be one of the most important improvements in Blaise (of course, we are all looking forward to ODBC in Blaise 4 Windows).

5. Conclusion

Growing needs for faster and more accurate statistical information are often in conflict with developer's possibility to complete and test the survey application in time. With tools and methods, which enable completing and finalising the survey even after its implementation, some extra time could be acquired. In many cases only little extra time for development means the difference between interviewer's exhausting hard work and easy, user-friendly data collection and editing. The possibility of redefining the survey after implementation can improve timeliness as well as the quality of collected data.

6. References

Bethlehem, J. (1995), "A Control Centre for Computer-assisted Survey Processing". Report, Voorburg: Statistics Netherlands.

De Jong, W.A.M. (1996), "Designing a Complete Edit Strategy; Combining Techniques". Research paper no. 9639, Voorburg, Statistics Netherlands.

Schou, R. (1995), "Developing a Multi-Mode Survey System". Third International Blaise Users' Conference, Helsinki.