Working Paper No. 4
17 April 2003

ENGLISH ONLY

UN STATISTICAL COMMISSION and
UN ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE-EUROSTAT Work Session on Population and Housing Censuses
(Ohrid, The former Yugoslav Republic of Macedonia, 21-23 May 2003)

Session I – Supporting paper

REGISTER BASED 2002 CENSUS OF POPULATION, HOUSEHOLDS AND HOUSING IN SLOVENIA AND NEW SOLUTIONS IN DATA PROCESSING

Submitted by the Statistical Office of the Republic of Slovenia[1]

I. INTRODUCTION

1. Between 1 and 15 April 2002, the Statistical Office of the Republic of Slovenia carried out the first population census in independent Slovenia. The preceding censuses took place within the scope of former Yugoslavia. The methodology was prepared by the Federal Statistical Office, with statistical offices in individual republics having to apply unified methodology with the possibility of adding some (very few) questions by themselves. All census data from 1948 to 1991 in former Yugoslavia took into account population with permanent residence (de iure), so that persons residing abroad (even if they stayed there for years) were also included as inhabitants.

2. Since its independence, Slovenia has joined several international organizations. The most important was a step to towards the European Union. In addition, Slovenian statistics rapidly began harmonising methodology in different fields with international standards and recommendations. The definition of population according to the Recommendations for the 2000 Censuses of Population and Housing in the ECE region was for the first time, and the most significant difference to the previous census, fully applied in the 2002 Census in Slovenia.

3. The National Statistics Act is the legal act which makes possible the provision of individual data from administrative registers, linking them together and developing the statistical registers which can be used only for statistical purposes, bearing in mind the confidentiality of data. Only aggregated data can be disseminated.

4. In preparing the 2002 Census implementation, we took into account the following starting-points:

- register orientation of Slovenian statistics will be followed and the greatest possible part of contents will be provided from the existing statistical and administrative registers;
- the costs of data processing will be cut by introducing new procedures and the processing time will be reduced;

[1] Paper prepared by Danilo Dolenc.

- the final control of micro data shall be carried out online and without paper (only with the help of scanned questionnaires).

5. The Statistical Office for the first time selected, according to the public tender, a private company for the data processing (Cetis Celje), which also includes printing of questionnaires, pre-print of data to questionnaires and distribution of questionnaires to and from the field enumeration. We concluded that the decision to outsource was a very good one in terms of data quality, time needed for processing and costs.

II. USE OF REGISTERS AND PREPARATION OF QUESTIONNAIRES FOR FIELD DATA COLLECTION

II.1 Preparation of questionnaires and pre-print

6. For the 2002 Census, the change from the classical method of preparing census questionnaires and data processing was dictated by great technological changes in the past decade and the development of high quality statistical and administrative registers in Slovenia, which we were able to use in preparing the data for pre-printing the questionnaires. In this way, we considerably decreased the burden of enumerators collecting the data and the population providing data, improved the quality of collected data and reduced the costs.

7. Pre-print of questionnaires means:

- a unique identification number for recognising the questionnaire in the later stages of processing (bar codes) is printed;
- identification of buildings, dwellings, households and persons was printed for linking enumeration units together (e.g. the same household number connecting persons to each household and the household with the dwelling);
- printing basic data on buildings, dwellings and persons alphabetically (address);
- printing of names, family names and PIN's to questionnaires for persons;
- marking (with 'X') which contents on the questionnaires for persons have already been obtained from the sources and exist in the pre-census database.

2.2 Pre-census database

8. The pre-census database combined data from various administrative and statistical sources that were used in the process of data collection and processing.

9. Before we started designing the questionnaires, we had to answer the following questions according to the demand on content of census data defined in the UN/ECE Recommendations and in the Census Act, adopted by the Slovenian parliament:
i) what administrative and statistical sources of data exist in Slovenia;
ii) the quality of data in sources;
iii) what census content can be obtained from them;
iv) to what extent the data can be obtained (for the whole population, only for a part of the population or just for some groups, etc.);
v) which data in sources are changeable and how regularly are they updated;
vi) which data are constant and therefore can be extracted only once (e.g. place of birth);
vii) are there key variables which allow us to link data from different sources (e.g. PIN).

10. I must mention here that we were able to obtain data only on persons and this will be the main topic of my paper. In Slovenia we do not (yet) have a register of dwellings and buildings. However, in our Register of Spatial Units (RSU), there are very detailed data for each house number – each registered building (addresses, census districts, codes, coordinates, etc.). On the basis of the RSU we pre-printed all addresses of buildings, but the number of dwellings in each

building (and thus the number of pre-printed addresses to questionnaires for dwellings) was estimated by the use of statistical assumptions.

11. In the next stage we defined whether or not the data in sources are available to cover some content in total or only partly. The contents which were not collected in the field have to fulfil certain conditions:

- data are available for the whole population;
- data are available for the whole segment of the population that is included (e.g. employed persons);
- there is a key variable and the processing is relatively simple and inexpensive;
- the quality is good enough for further processing.

12. The contents that we were able to provide only for a part of the census population were entered on the questionnaires; however, we marked on the questionnaire which questions need not be answered by the respondents because the data will be taken from the pre-census database.

13. The most important data sources for preparing the pre-census database were:

- Central Population Register (the basic source for demographic data on citizens);
- Permanent Population Register (foreigners);
- Register of Spatial Units (addresses, territorial division);
- Statistical Register of Employment (employed and self-employed persons, persons with pension and disability insurance in Slovenia);
- Business Register of Slovenia (data on business subjects);
- Unemployment Register of the Employment Service of Slovenia (registered unemployment);
- data from the Pension and Disability Insurance Institute (pensioners);
- data of the statistical survey on students and graduates (education data);
- 1991 Census data.

14. The contents that were taken over from the pre-census database:

1. **Entirely:**
place of birth
last migration
citizenship
marital status
field of education
employment status
activity
occupation
usual working hours
place of work

2. **Partly:**
sex
address of the residence one year before the census
first residence after birth
education program
place of education
activity

15. It was very important that we decided not to collect two variables which are difficult to ask, difficult to answer and difficult to code: occupation and activity. Both these data were provided from the register also for the 1991 Census. The most pretentious work in the pre-census database processing was the extracting of the status of activity. We were able to cover 80% of data from the pre-census database on that topic.

II.3 The usage of the pre-census database

16. For preparing the pre-census database (over 2 million entries on people and almost 500,000 entries on buildings), over 10 statistical and administrative registers were used. The pre-census database was used:

A. in the organization period of field enumeration for:

- mathematically determining identification numbers (barcodes, other identification);
- above-mentioned pre-printing;
- calculating quantities of all census questionnaires and other census material;
- helping organise fieldwork (number of enumerators, field-costs, etc.).

B. in the stage of processing data for:

- checking the coverage and return of questionnaires;
- putting together data from questionnaires and from the pre-census database;
- controlling data accuracy and logical consistency.

17. The last version of the pre-census database that was used for input and logical controls corresponds to 31 March 2002 (the reference date of the 2002 Census). Only changeable data were updated and the last update was done at the end of June 2002.

II.4 Methods of data collection

18. 2002 Census data were thus obtained in two ways: from administrative and statistical sources; and with fieldwork where two data collection methods were used:

- self-enumeration, when respondents themselves were able to fill in a part of the questionnaire;
- classical enumeration, when all census questionnaires were filled in by specially qualified enumerators.

III. DATA PROCESSING

19. Data on household and family structure were coded in the field by enumerators. After the questionnaires were collected from the field and delivered to the chosen company, there was no correction or checking of data, but the scanning process began at once. Such an organization solution was applied for the first time in the post-census operation. Mistakes that were made by enumerators were corrected in the last stage of editing (consistency checks). As we expected, the majority of mistakes corresponded to wrong coding of families. Different methods of data collection required adjustment of all procedures in electronic data capture and control of data.

III.1 DATA CAPTURE

20. With electronic data capture we provided:

- the optical photo archive of all census questionnaires, which could be used at later stages of controlling consistency of data (paperless control);
- recognition, interpretation and verification of field collected data;
- automated coding of texts;
- SQL databases for final consistency control.

21. We set up five basic objectives for electronic data capture:

- very efficient transfer of data from paper to electronic form (scanning);
- simple output of data for later processing;
- adequate quality of capture (les than 1% of errors);
- minimal influence of the operator;
- low costs.

22. The scanning of all questionnaires was done by four highly operable document scanners. The total number of scanned images was over 14,000,000 A4 pages. The capacity of images is 600 GB of hard-disk space. The whole process of scanning, verification of scanned data and transfer of

data to the SQL database was done in 92 workdays in two shifts by 48 persons. The verification took place on only 12 workstations per shift (PC's).

23. All textual data were coded already at the verification stage. In addition to computer assisted coding, we applied to the greatest possible extent the methods of automatic coding, so that operators needed to intervene in only 3% of textual data. Recognition rate for scanned signs (mark 'X' in that field) was 99.99%. The method of automatic coding was applied for the first time in the census operation in Slovenia (coding of nationality/ethnicity, religion and all territorial texts – countries, municipalities, settlements). Via the bar code, pre-printed questionnaires were taken over directly from the database, which accelerated the verification and improved the quality.

24. We used two types of codebooks:

- standard codebooks (e.g. ISO, spatial units);
- open codebooks (nationality/ethnicity, religion) where we successively added all new texts that appeared in the questionnaire beside the texts that were defined in advance. Because of the Slovenian grammar - which has a lot of cases, possible male or females forms or adjectives - we prepared for every code massive versions of texts and added them to the codebook.

25. We also improved the verification process by adding the versions of names in standard codebooks that appear repeatedly (e.g. BIH for Bosnia and Herzegovina) but, of course, we did not change the codes.

26. The most important technical novelty in processing the 2002 Census data was the optical archive, the images of all census questionnaires. By this solution the paper questionnaires were taken to the warehouse immediately after the successful scanning. From then on, all operations (verification, correction of data, coverage) were done only on the basis of images.

III.2 Correction of data

27. Usually the most demanding part of processing consists of formal checks, logical checks and inconsistency checks at the level of variable, two or more variables in the same questionnaires and two or more variables in different questionnaires. Errors may occur from respondents, but the most common are the result of enumerator's fieldwork. Some errors also occur in the editing phase, but their share is normally less than 1%. A particularity of the 2002 Census was also the possibility of inconsistency between the data from the pre-census database and the data collected in the field.

28. For the first time data accuracy control was implemented without paper, only on the basis of photos of questionnaires (online). The manipulation of the large extent of census questionnaires, folders for questionnaires and a great number of required staff (operators, junior clerks, etc.) always caused logistic problems.

29. In comparison with the processing of previous censuses, for the 2002 Census an extremely large number of controls was introduced as a consequence of the demanding and sophisticated data collection methods. Even more important was the solution to correct errors mostly automatically. The main advantages are the uniform approach to corrections on the basis of exact criteria and the shortage of time needed for corrections. All in all more than 1,300 different formal, logical and consistency checks were made in every set of questionnaires (package), more than 90% were in the case of errors done automatically without any intervention of the operator.

30. Preparation of the software and all other technical solutions for online operation of the system of checking and correcting is exclusively the result of domestic knowledge. Experts of the company Cetis Celje, which was selected at public tender, did their work on preparing the application with methodological support from the Statistical Office in a very short time (less than 4 months) and with exceptional innovation. The application was designed user-friendly and allowed us to change, add or remove some controls in a very simple and quick mode. For rapid operation of

the application, it was necessary to optimise the work with the database and use a simple and quick user interface, which was designed so that the person processing the data got on screen all the questionnaires necessary for correcting the errors.

31. The application allows only correction of those data that are precisely defined for an individual control and are active on the questionnaire image. It does not allow the operators to change in any way other data. In addition to the graphic presentation of photos, the operators also had at their disposal a tabular presentation of data. For control, the operators always had on screen the data on the current execution and course of controls.

32. The very sophisticated system of linking together the data from three different databases, and at the same time the transfer of four types of images (more than 14 million images in total) to the screen, also had a very good response time measured in milliseconds. The checking of an average enumeration district (about 150 persons, adequate number of households, dwellings and buildings) lasted depending on the quality of input data from 15 to 25 minutes. Also in this phase only 12 working stations (PC's) were used. There was about two hours of training needed for an operator to handle the software and about one day of methodological training before the operator could start to work independently. Because the whole system worked mainly automatically, the number of expected operators was reduced to the minimum. After one week of training, one operator was able to handle three PC's at the same time. Most of the work on checking and correcting of input data was done in only two months with 16 operators working in four shifts day and night.

III.3 Coverage control

33. The entire concept of the methodological design of the census and data processing also enabled us to control for the first time – after the control of data accuracy and consistency – the duplication of enumeration of individual people and to control which people were for various reasons not enumerated, and thus to decrease non-response. This final phase was even more important because of the discrepancy between persons registered and persons actually residing, where we found out that more than 100,000 persons did not live at the addresses that were pre-printed on the questionnaires. Therefore, it was possible that:

- some people were enumerated twice as inhabitants (at the address where they actually live and at the pre-printed address);
- people who had moved to another place in Slovenia were not enumerated at all.

34. For the first case we prepared criteria for which double record should be deleted (and of course it was necessary to check these units again), while for the second case we used the statistical methods and the data from the pre-census database. Almost 1% of the population was counted twice and slightly less than 2% of the population was added to the population census file because of non-enumeration.

IV. CONCLUSION

35. The most important improvements in the processing of data compared to the 1991 Census were:

- composition of a pre-census database which was used for printing data to the questionnaires and for planning all organization activities;
- uniform identifications and barcodes, and printing of all identifications to the questionnaires;
- full acceptance of some census topics from the pre-census database (without field collection);
- an optical archive of images of all questionnaires;
- simultaneous verification and coding;
- automatic coding of territorial data and data on nationality and religion;
- online consistency check supported by images of questionnaires;

- very short time for the whole data processing (less than 8 months).

36. The register-based 2002 Census was also the first but important step towards the “virtual” Census which is, no doubt, the future challenge for the Statistical Office of the Republic of Slovenia.

37. The Statistical Office of the Republic of Slovenia published the final data of the 2002 Census on 16 April 2003, only one year after the last day of enumeration. The press conference took place in the Slovenian Parliament. This was possible such short time after the Census because of the great enthusiasm of staff involved and because of the optimisation of all processes in the company Cetis Celje, which provided technical support, and, of course, also in the Statistical Office (derived variables, imputation, re-coding, tabulation, publication, dissemination).

38. The users who are interested in 2002 Census data in Slovenia can find numerous data, together with extended methodological notes, on our web site www.gov.si/popis2002 in English.