UNITED NATIONS **E**

Economic and Social Council

Distr.
GENERAL

ECE/CES/GE.41/2007/7
26 March 2007

ENGLISH
Original: ENGLISH/RUSSIAN

ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Group on Population and Housing Censuses

Tenth session
Astana, 4-6 June 2007
Agenda item 3 (b) of the provisional agenda

CENSUS TECHNOLOGY: RECENT DEVELOPMENTS AND IMPLICATIONS ON CENSUS METHODOLOGY

Experience in using scanners to process data from the 1999 population census

Submitted by the Statistical Agency of the

REPUBLIC OF KAZAKHSTAN

The meeting is organized jointly with Eurostat

GE.07-21252 (E) 230507 300507

SUMMARY

At its meeting held in Washington, D.C. (United States of America) on 19 and 20 October 2006, the Bureau of the Conference of European Statisticians (CES) approved the revised terms of reference of the Steering Group on Population and Housing Censuses and the plan for future CES activities on population and housing censuses. The CES Bureau also agreed that the Steering Group would coordinate the work of various meetings. The present paper was prepared at the request of the Steering Group on Population and Housing Censuses for presentation and discussion at the Joint UNECE/Eurostat Meeting on Population and Housing Censuses in Astana (Kazakhstan), to be held from 4 to 6 June 2007. The paper provides a substantive basis for discussion at the meeting devoted to "Census technology: recent developments and implications on census methodology".

I. INTRODUCTION

1. The processing of data from Kazakhstan's first national population census in 1999 was the most difficult but, at the same time, most successful project conducted by the Computing Centre of the Statistical Agency of the Republic of Kazakhstan. In order to implement the project (1998-2001), modern tools were used to develop application programmes and a "client-server" technology, and competent data processing specialists were trained. The database created using census results is unique in Kazakhstan and is one of the assets of the Statistical Agency. Previous censuses were conducted by the Central Statistical Office (CSO) of the Kazakh Soviet Socialist Republic, which established centralized databases. The results of these censuses have been kept only in the form of printed handbooks. Now, when preparations for the next population census are under way, a synthesis of the results of the work conducted, an analysis of all the measures involved in processing the materials of the first population census, and an analysis of the errors are of great importance.

II. ESTABLISHMENT OF SECTIONAL DATA PROCESSING CENTRES

2. Upon conclusion of the census (25 February-4 March 1999), all census materials were verified at instructor stations and later at census offices; the materials were subsequently transmitted to district statistical divisions. After an additional check, the census materials were sent to provincial (municipal) statistical offices for coding; the offices sent the materials to sectional centres for subsequent automated processing, in accordance with the schedule. A total of five sectional centres were established:

(a) The Almaty centre (Computing Centre of the Statistical Agency of the Republic of Kazakhstan) - to process data of North Kazakhstan and Almaty provinces and the city of Almaty;

(b) The Aqtobe Aktyubinsk centre (Aqtobe) - to process data of Aktyubinsk, Atyrau, Mangistau, West Kazakhstan and Qyzylorda (Kyzyl-Orda) provinces;

(c) The East Kazakhstan centre (Ust-Kamenogorsk) - to process data from East Kazakhstan and Pavlodar provinces;

(d) The Qaraghandy (Karaganda) centre (Qaraghandy) - to process data of the Karaganda, Akmola, Qostanay (Kustanay) provinces and the city of Astana;

(e) The South Kazakhstan centre (city of Shymkent) - to process data from South Kazakhstan and Zhambyl provinces.

3. The sectional centre within the Computing Centre of the Statistical Agency also operated as a national centre, i.e., all data from sectional centres were transmitted to the Computing Centre for further processing. After all the census forms were optically read, they were returned from sectional centres to the provincial statistical offices for safekeeping.

4. From the end of April, the centres started receiving materials for processing; by the end of August 1999, data entry was completed in all centres. The main data on the amount of material processed are set out in table 1. It can be noted that between 15,000 and 20,000 forms were loaded into one scanner in one day.

Table 1

Main indicators of the amount of data processed

Processing centres	Number of scanners, *units*	Number of census forms, *units*	Number of folders, *units*	Number of forms processed with one scanner, *units*	Length of processing, *days*	Number of forms processed in one day, *units*
Almaty	3	4 743 300	15 800	1 581 100	102	46 503
Aqtobe (Aktyubinsk)	2	3 402 191	11 420	1 701 096	110	30 929
East Kazakhstan	2	3 000 123	10 040	1 500 062	90	33 335
Qaraghandy (Karaganda)	2	4 590 000	14 027	2 295 000	120	38 250
South Kazakhstan	2	3 710 657	12 847	1 855 329	90	41 230
Total	11	19 446 271	64 134			

5. A total of 467 temporary staff were recruited for data entry and initial data correction (see table 2). The average monthly salary of the staff working in the data processing centres was 6,000 tenge.

Table 2

Number of workers who took part in data processing

Data processing centres	Staff responsible for data correction	Staff preparing forms for the scanner	Technical support specialists	Total, *persons*
Almaty	90	36	3	129
Aqtobe (Aktyubinsk)	60	21	3	84
East Kazakhstan	68	12	2	82
Qaraghandy (Karaganda)	66	16	2	84
South Kazakhstan	68	16	4	88
Total	352	101	14	467

III. DATABASE TECHNICAL AND SOFTWARE SUPPORT FOR THE CENTRES

6. Each centre set up local computer networks. In order to optimize the work, the networks were divided into segments consisting of one server, one scanner, one personal computer (PC) for scanning, one PC for data recognition and 10 PCs for data correction. Thus, each local computer network included two servers, two scanners and 24 PCs. All complexes specially created for data processing operated in the Ethernet network on the Windows NT Server.

7. The ScanStar 5045C colour scanner makes it possible to scan 50 forms (format A4, weight 70-80 g/m²) a minute at a resolution of 200 dpi (dots per inch). The input tray is loaded simultaneously with 100-150 forms (in theory, 300 is the permissible load); various types of forms can be mixed together in one batch. In practice, 27-30 forms were scanned a minute

because the paper did not always have the permissible thickness; moreover, after every 1,000-1,200 forms, the scanner had to be cleaned. Data entry was also slowed down owing to the low speed of recognition, which could be regulated (the lower the recognition speed, the higher its quality).

8. Colour scanners ScanStar 5045C and software BUSY, Image Port, JobScan and RecoStar are manufactured by Computer Gesellschaft Konstanz, which is a branch of Siemens Nixdorf (Germany). The total cost of one scanner with all the database software was US$ 67,770. The total cost of the project, which included the delivery of 11 sets of equipment, its installation in sectional centres, the setting up of a spare parts depot, and training, was US$ 995,680.

9. The database software has the following uses: BUSY - software for automated processing of the flow of documents in local computer networks and development of additional applications; Image Port, JobScan - a software complex for controlling the scanning process and connecting the scanner to the processing system; and RecoStar - character recognition software.

10. The recognition speed depends on the capacity of the computer on which the RecoStar programme has been installed and can reach 100 characters per minute. Recognition errors account for no more than 1 per cent, provided that the necessary requirements are met with regard to the quality of paper and the printing of high-quality forms at the printing house. The information density of the forms was:

1B form - 22 characters, 1 mark;

2P form - 334 characters, 14 marks;

3C form - 141 characters, 21 marks;

4I form - 142 characters, 6 marks.

11. The main function of the BUSY system is to enable the user to control the workflow of documents and their individual processing. The system also has tools for developing additional applications and for administration. For the development of additional applications, the BUSY system offers a variety of its own functions, which are required in order to verify the data entered, and gives the programmer an opportunity to develop independently and activate specific procedures written in the Visual C++ language. The shell of the BUSY software can control all data-processing processes - the scanning of census forms and the recognition, verification, correction, conversion and electronic archiving of data.

12. With a view to developing additional applications, specialists were trained not only to use the new Image Port, JobScan, RecoStar and BUSY systems but also to work with the Visual C++ package and with network operating systems. Specialists from Germany trained our programmers to work with all software products delivered with the scanner. In addition, our specialists trained users from the sectional centres to work with the newly developed software complexes. Detailed user instructions were prepared, with a full description of all phases of work and ways of solving all types of problems that might arise at each processing phase.

IV. STAGES OF THE TECHNOLOGICAL PROCESS OF DATA PROCESSING

13. The technological process of data processing that we selected is fairly complicated and labour-intensive, but makes it possible to achieve high processing speed and obtain data of sufficient accuracy and quality. The programmes are based on algorithms that make it possible to detect not only recognition errors but also errors made by the counter because the forms had been filled out carelessly or hastily. All stages of processing were highly automated.

14. The entire process of automated processing of population census materials consists of two levels. The following tasks were carried out by the sectional centres:

(a) Scanning census forms;

(b) Recognition of the content of the forms;

(c) Correction in the BUSY medium of incorrectly recognized or unrecognized information in three stages: initial, main and using the image (of the scanned document);

(d) Verifying data consistency in terms of the arithmetic, the logic, and by comparing forms with each other;

(e) Automated data coding and conversion;

(f) Uploading converted data into DMS Access (Database Management System);

(g) Correcting data in the DMS Access medium using the image;

(h) Verifying district information against the main indicators;

(i) Archiving, copying to CD-ROM or diskettes and transmitting to the national centre.

15. A diagram of the technological process of data processing in one segment of a local computer network, established by connecting one scanning and recognition system is contained in annex I.

16. The following tasks were implemented at the national level:

(a) Receiving information from district centres;

(b) Verifying the completeness of district information and checking it against the main indicators (DMS Access);

(c) Establishing a database for each province (DMS MS SQL Server);

(d) Drawing up tables by sections;

(e) Making spare copies;

(f) Analysing data and obtaining specifications from provincial offices;

(g) Downloading the database on provinces into the central database;

(h) Drawing up summary tables and generating reports.

17. Once the census forms are loaded by scanning, the data is automatically recognized. The content of the forms undergoes special tests. The recognition results are compared with data from dictionaries, which leads to a significant improvement in the quality of recognition. Once recognition is completed, the user has the option of correcting the census forms. Data from each folder are copied into a separate file and, if necessary, undergo several stages of verification and correction.

18. **Organizational features of the correction process.** A multiphase system of verification and correction has been developed; all data are first verified in terms of consistency, then in terms of the arithmetic and logic, and are subsequently checked against the tables.

19. The correction process consists of three consecutive stages - initial, main and correction by a specialist. At the initial stage of correction, all errors linked to poor character recognition are corrected. The presence of an accompanying form is verified, or the form is duplicated. Then the territory code in the “B” form is checked (if the territory code is correct, it is automatically copied to the other forms). In addition, at the initial stage, indicators, the meaning of which can be found in the dictionary (place of birth, nationality, citizenship, State language), are verified. A check is also carried out to ensure that the numbers of forms have not been duplicated and that none of the forms are missing. Forms can be renumbered in blocks, if necessary.

20. After the initial stage of correction, the process moves to the main stage, at which the logical control mechanism starts functioning immediately; the programme analyses errors and, where possible, corrects them (marks and values are automatically inserted according to a specific algorithm developed by methodologists).

21. If, during the main stage of correction, problems arise that cannot be solved by a simple corrector, the process moves to the next stage, in which errors are corrected by a specialist, a population census expert.

22. Thus, the following methods were used to optimize data processing. First, the process of correction was divided into several stages: starting with simple mechanical correction, the results of which do not need to be analysed, and ending with a stage of correction that requires a certain level of knowledge and special training on the part of the person responsible for data correction. Secondly, during the correction process the user is offered correction alternatives: the user simply needs to select the desired corrections and confirm his or her choice. A corresponding text from the dictionary appears alongside the recognized text to enable the data to be compared, or the programme offers “pop-up” lists so as to enable the user to choose the text to be corrected. The BUSY and RecoStar system facilities make it possible to use dictionary technology for automatic data correction in handwriting recognition. A dictionary is consulted during the recognition process; the results of text recognition are compared with the corresponding dictionary data and are automatically corrected, if necessary. The recognition results are also

compared with data in the trigram tables, which contain the most frequently used three-letter combinations; once the comparison is completed, the data are automatically corrected. After the recognized text has been checked, the data are automatically coded; instead of the text, a numerical value of the code is inserted into the file.

23. Once correction has been completed, the data are automatically coded. The data from each folder, which have been verified and rectified, are downloaded to the DMS MS Access database, where they are again verified and, if necessary, corrected. Here, preliminary results within one district with regard to the main indicators can be obtained and, if necessary, the processed folder can be returned from this stage back to the BUSY system for further corrections; images of forms from the file that has been returned are kept in the BUSY system until the correction process has been completed. The fully corrected data are transferred from DMS MS Access to the DMS MS SQL Server, where information is stored and a copy of the database is made. The DMS MS SQL Server is one of the most powerful database servers and makes it possible to use "client-server" architecture. In this architecture, the client application transmits a request to the database server, on the basis of which all commands are implemented. The results of the commands are sent to the client for use, viewing and printing.

24. Data accumulated in territorial subdivisions are downloaded for transmission to the centre via communication channels and on compact disks. A centralized census database has been created at the national level. It operates on the basis of the DMS MS SQL Server. Annex II provides information on the size of the databases.

V. APPLICATION PROGRAMMES DEVELOPED FOR THE PROCESSING STAGES

25. The introduction of new and fairly complex technologies and the use of multiple stages to process census materials required the use of a large number of new software products. It became necessary to develop many sets of additional applications, namely:

(a) A set of additional applications using the Image Port, JobScan, RecoStar and BUSY systems:

(i) Generating document descriptions;

(ii) Developing a system for verifying uploaded information;

(iii) Developing data correction programmes;

(iv) Developing data conversion programmes and programmes for downloading information from the BUSY system to the final database.

(b) A set of additional applications using Visual C++, Access, MS SQL:

(i) Programmes for verifying and correcting data downloaded from the BUSY system to Access;

(ii) Developing accessory programmes for processing census materials;

(c) A set of additional applications designed to control the overall flow of information in the system:

(iii) Data conversion programmes;

(iv) Programmes for creating and managing reference information and information on regulations;

(v) Programmes for regulating the registration of information and its transfer from branches to the central office.

(d) A set of programmes for creating tables of regulations:

(i) Developing programmes for creating technological files with aggregated census information;

(ii) Programmes for generating reports.

26. The large number of newly developed application programmes indicates the level of complexity of the project. It also indicates the high level of programming specialists, who managed, in a short period of time, to master and introduce information technologies that were rather complex for their time. It was here that “client-server” technology was used for the first time and this demonstrated in practice the advantages of this method in the processing of large data flows.

VI. MAIN PROBLEMS ENCOUNTERED IN PROCESSING POPULATION CENSUS MATERIALS

27. Recognition errors depended a great deal on the quality of the census forms, which had been poorly printed.

28. The excessively tight time limits affected the quality of the programmes developed. The training of specialists was completed only five months before the beginning of the census.

29. There was no possibility of studying the ways in which other organizations used similar equipment in a population census since the countries of the Commonwealth of Independent States did not have such experience.

30. Other forms had been used in a pilot census; other technology had been envisaged. As a result, a prototype of the future system had not been developed and no testing had been conducted.

31. Database software delivered with the scanner also has its shortcomings. The software can be used for a narrow range of operational requirements related to its application in commercial banks; however, it had not been used in a census. The software has built-in functionality but it is very sensitive to the introduction of additional applications developed by our programmers

(pop-up windows, folder identification system, etc.), which cause frequent failures and slow down the system. No verification is carried out within the batch of documents (in the past, forms were not compared with each other on this system).

VII. CONCLUSION

32. Despite organizational problems and difficulties, data from the census forms were entered earlier than planned. For the first time, all participants in the project could see and appreciate in practice the advantages of using scanners to process census materials. The experience acquired in the implementation of this project was used to process materials of an agricultural census that was conducted in two stages in 2006 and 2007.

Annex I

DIAGRAM OF THE TECHNOLOGICAL PROCESS OF DATA PROCESSING USING SCANNERS

1. Scanner ScanStar 5045

2. PC with ImagePort

3. PCs for further processing

4. SCSI

5. Local computer network Ethernet-LAN

6. JobScan software

7. RecoStar software

8. BUSY software

9. Recognition station

10. File server

1. To enable scanning, scanner ScanStar 5045C is connected to a PC on which the Image Port and JobScan software have been installed.

2. Once it has been scanned, the batch of documents (file) is sent from the scanning PC to the file server, on which the server part of the BUSY software has been installed.

3. The file is transferred automatically (or manually) from the server to the “Recognition station” PC, on which the RecoStar programme has been installed.

4. Once recognized, the file is automatically returned to the server.

5. Correction is carried out on all “computers for further processing”, on which client components of the BUSY software have been installed.

6. DMS Access is installed on one of the “computers for further processing”; district data is corrected and verified for completeness.

7. To help the operator, an image of the census document appears on the PC screen when items 5 and 6 are being carried out.

8. Once verified, district data are entered into the database for each province; the database operates on the DMS MS SQL Server.

Annex II

TABLE. INFORMATION ON THE SIZE OF THE DATABASE FOR THE POPULATION CENSUS

	Provinces	Number of entries in the database	
		3C form	2P form
1	Akmola	843 245	832 267
2	Aqtobe (Aktyubinsk)	690 859	679 438
3	Almaty	1 569 571	1 533 758
4	Atyrau	450 694	437 857
5	East Kazakhstan	1 539 184	1 524 635
6	Zhambyl	998 445	980 310
7	West Kazakhstan	623 931	610 052
8	Qaraghandy (Karaganda)	1 422 851	1 389 921
9	Qostanay (Kustanay)	1 024 333	1 001 850
10	Qyzylorda (Kyzyl-Orda)	602 248	593 315
11	Mangistau	319 442	312 998
12	Pavlodar	809 883	803 257
13	North Kazakhstan	730 056	722 280
14	South Kazakhstan	1 990 444	1 972 187
15	Astana (city)	324 758	318 769
16	Almaty (city)	1 156 806	1 096 483
	Republic of Kazakhstan	15 096 750	14 809 377
	Database size, Gb	4.5	1.62

The table contains information on the size of the database created after the initial processing of the forms - reading and correction. The size of the database created after the final processing is 13 Gb. Every entry in the database corresponds to one physical person and consists of 66 fields; data can be called up by a populated area, district or province. The population census information system operates on the DMS MS SQL Server and has the following components:

(a) Databases of 16 provinces (main resource);

(b) Database containing reference information and information on regulations (classifiers, dictionaries, reference books);

(c) Modules that generate output forms.
