United Nations ECE/CES/GE.41/2017/8

Economic and Social Council

Distr.: General
12 July 2017
English
Original: Russian
English and Russian only

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Nineteenth Meeting
Geneva, 4-6 October 2017
Item 3 of the provisional agenda
Innovations in census technology

Issues in protecting raw data when implementing technological innovations in the census

Note by the Russian Federation Federal State Statistics Service

Summary

In the Russian Federation, the 2020 census will use a combined technique to collect information on the population: respondents will independently fill in electronic questionnaires on the Internet; enumerators will canvass and fill in electronic questionnaires on tablet computers; and enumerators will also canvass and fill in traditional paper questionnaires. Innovations pose new challenges for the protection of data confidentiality. They introduce computerization not only at the stage where the collected data are processed, but throughout the chain of preparation and implementation of the census in the field. This has a significant impact on the cost of the census as a whole, making its technological component more expensive.

GE.17-11637 (E) 040817 150817
1711637

Please recycle

I. The 2010 Russian population census: practice

1. The 2010 census was conducted using the traditional method of making rounds to residences and collecting information on the population through a survey. The data were entered in machine-readable census sheets by specially trained enumerators; the sheets were subsequently scanned, machine-read, verified and consolidated into common information sets by region and for the country as a whole.

2. The federal Population Census Act includes a provision that population data in census sheets are limited-access information, are not to be publicly disclosed or disseminated and are to be used only for the purpose of establishing official statistics. Machine-readable census questionnaires bore the marking "CONFIDENTIAL (confidentiality ensured by the person receiving the information)".

3. The law on censuses includes mandatory provisions to ensure that information is protected: under the law, persons with access to census sheets bear responsibility for the loss or disclosure of such information and for its falsification. The obligation not to disclose confidential census information is set out in census workers' contracts.

4. Owing to the need to monitor the validity of the data collected by enumerators and the consolidation of data on persons in households, in addition to the machine-readable census sheet forms, non-machine-readable lists were drawn up of persons living in each residential unit. During the data collection and verification phases it was possible using these lists, to identify, in each unit, the people among the respondents who were registered in the households under numbers 1 or 2, etc., and to determine who lived together in single households and who lived in separate ones.

5. For the depersonalization of the data collected during the population census, the census forms contained no blanks to fill in family names, given names or patronymics; they only had blanks for the number from the list of residents. This made it possible when processing the census sheets to make the data anonymous, separating the machine-readable census sheets from the non-machine-readable lists of residents.

6. The census sheets are processed in a way that ensures that they are protected against unauthorized access and that prevents their theft, loss, forgery or other falsification. This is ensured through organizational and technical solutions incorporated in the census infrastructure and in the infrastructure's information protection certification by the authorized State bodies.

7. The 2010 national census was the target of a clone of the census's information website. On the first day of the national census, a bogus census site appeared on the Internet. The imposter site copied the official national population census site, broadcasting information that was either simplified or distorted. In addition, navigating from any page to the site's homepage produced a poor quality, mobile text version that was heavily simplified. In addition to using an indistinguishable style of presentation, the domain name of the bogus site was similar to the real ones. Nonetheless, representatives of the Federal State Statistics Service immediately announced that the bogus site domain, PERIPIS-2010.RU, had nothing to do with the official domain, PEREPIS-2010.RU.

8. The developers of the official site of the national population census warned users not to click on suspicious links and called on them to exercise caution and not to download information from sites until they were sure of their authenticity.

9. The bogus site did not cause any direct damage, but it undermined many citizens' confidence that the census could be conducted securely. If such a situation is repeated with a census using the Internet, there is the possibility that people not very well versed in modern technology could send their information to scammers, resulting in a lack of confidence in the census.

II. Innovations in the 2020 census

10. The 2020 census will utilize both paper and electronic questionnaires (with identical questions) and three ways to collect the data about the population:

- self-administered completion of electronic questionnaires by respondents, via the Internet (online)
- a survey with electronic questionnaires filled in by enumerators on tablet computers
- interviews and traditional completion of paper questionnaires by enumerators

11. In the light of international experience, the online census usually takes place first. This is so that the statistical agencies can establish specifically what part of the population and what units and households have been counted over the Internet. Thereafter, the second phase starts — the traditional population survey on paper supports or using tablet computers. According to the plan, 45% of the reporting districts will use tablets and 55% will use paper census sheets.

12. In terms of economics, combining the techniques will make it possible to substantially reduce expenditure for the production of paper forms, the purchase of scanning equipment, labour costs for the entry of machine-readable census sheets, data coding and data quality verification and technical backup for such procedures.

Table 1

Comparison of the overall cost of the 2020 national census with the use of different data collection techniques

(*million rubles*)

Options for the 2020 national census	*Printing machine-readable census sheets (% of option 1)*	*Wages of hired census personnel (field workers and automation engineers) (% of option 1)*	*Software development, equipment and licence procurement; processing (% of option 1)*	*Other costs (outreach/awareness campaign, equipment, procurement/ logistics, training) (% of option 1)*	*Total (% of option 1)*	*Excess (for minimum expenditure) over the cost of option 1 (% of option 1)*
Option 1: Three collection methods (portable computers, machine-readable census sheets and via Internet)	371.7 (100.0%)	33 153.3 (100.0%)	9 360.3 (100.0%)	9 364.8 (100.0%)	52 250.1 (100.0%)	–
Best option for conducting the census: 10% conducted via Internet, 40% by tablet computers and 50% using machine-readable census sheets						
Option 2: Traditional method (machine-readable census sheets)	743.4 (200.0%)	39 957.1 (120.5%)	6 454.6 (69.0%)	9 364.8 (100.0%)	56 519.9 (108.2%)	8.2%
Difference from option 1: printing of machine-readable census sheets for 100% of the population (thus, doubled); heavier workload on all employee categories (increase of 20% in staffing costs); minimal expenditure on software development, but more work stations at the district and regional levels for processing of machine-readable census sheets (100% usage)						
Option 3: Two collection methods (portable computers and machine-readable census sheets)	371.7 (100.0%)	36 627.4 (110.5%)	8 909.2 (95.2%)	9 364.8 (100.0%)	55 273.1 (105.8%)	5.8%
Difference from option 1: 60% using enumerators with machine-readable census sheets, 40% using enumerators with tablets, heavier workload for employees in the field (10% more for wages), no expenditure on software development for conducting the census on the Internet, additional workstations for processing of the 10% more of the population using machine-readable census sheets						
Option 4: Two collection methods (machine-readable census sheets and Internet)	743.4 (200.0%)	38 292.2 (115.5%)	6 642.3 (71.0%)	9 364.8 (100.0%)	55 042.7 (105.3%)	5.3%
Difference from option 1: 90% of the population covered with enumerators using machine-readable census sheets, with the print run doubled; staffing costs 8% higher, greater burden on enumerators and automation engineers, more workstations at the district and regional levels to process the additional quantity of 100% machine-readable census sheets						

13. The 2020 national population census is planned to take place from 1 to 10 October 2020 for the online census.

14. People wishing to take part in the census via the Internet must not be inadvertently excluded or counted twice in the process. Every participant must thus be identified. To solve this problem, there are plans to make use of already existing means of personal identification on the Public Services Portal of the Russian Federation, through the Uniform System for User Identification and Authentication (the Uniform System).

15. The Uniform System is a reliable public resource that aggregates the population's personal data from various departmental data sources. Its distinguishing characteristic is that it contains data only about people who have registered with it of their own volition. It is precisely such people who prefer using electronic government services rather than having traditional personal interaction with State bodies.

16. Once respondents visit the census's Internet page, they will be requested to select their place of residence from a list of addresses prepared by the Federal State Statistics Service in accordance with the census districting. One of the persons permanently residing in the place of residence can fill in the electronic census sheets for himself or herself and also for the other residents who authorize him or her to submit their information for the census.

17. The electronic census sheet should also include a question whereby respondents consent to the transmission of their data over the Internet.

18. The Uniform System makes it possible to fill in some of the answers in advance on the basis of respondents' personal data already in the system, such as:

- sex
- date of birth
- place of birth
- citizenship

19. With the respondent's consent, such information can be transferred to the electronic census sheet automatically. Respondents will be able to verify and correct the data while the federal statistical survey is underway.

20. During the online census phase and for four days after it ends (until 15 October 2020), the Federal State Statistics Service will update the electronic address lists, noting the residence addresses where the population took part in the Internet census and entering the number of men and women and the number of households recorded.

21. From 16 to 31 October 2020, enumerators will visit all the residences in their reporting districts and will confirm participation in the online census; they will then fill in the electronic or paper census questionnaires for the population that has not taken part in the online census.

22. Respondents who have taken part in the online census will receive confirmation codes that can be entered in the information for the corresponding residences in the address records. When residences with confirmation codes are visited, the enumerators will check the respondents' codes against the codes for their addresses. If necessary, they will correct the information (for example if one of the household members was not registered online). The confirmation codes will contain encoded information identifying the households, the sex of the respondents and their family relationship with the members of their households. This will make it possible to verify the information on the members of the household who have been registered online and by the enumerators. The respondent's name will not appear in the census sheet or the address list.

III. New challenges in protecting confidentiality while introducing innovations in the 2020 population census

23. The main aim in protecting confidential information when carrying out the population census using portable computers and the Internet is to protect raw census data by securing it against unauthorized access.

24. From the technological point of view, there is the problem of the deployment of the online census software: it can be embedded into the Public Services Portal of the Russian Federation, or it can be placed on a specially developed site for the online census, with a link from the Public Services Portal.

25. The main benefits of placing the online census software on the Public Services Portal is that the portal is already equipped with the technical tools to protect confidentiality and personal data and to defend itself against denial-of-service attacks; and it also has sufficient capacity. If the software is deployed on a site specially developed for the online census, additional investments will be necessary to ensure that data confidentiality is protected, to protect the site against denial-of-service attacks and to provide the required capacity for the software.

26. The main benefit of placing the software on a specially developed website for the online census consists in the fact that the appearance and functionality of the electronic census sheets and the algorithms used to verify the responses to the questions on the census sheets are not subject to any technical limitations. In practice, the use of software for federal statistical surveys has shown that some of the rules for the interactive verification that responses have been filled in are directly changed during the process itself, when the actual data received differ from what was expected by the developers. The process for the approval of changes and making changes in the electronic forms on the Public Services Portal involves many participants and requires an extended period of time. Such delays could have a critical impact on the census.

27. Thus, we have chosen the option of a specially designed website for the online census, using the Uniform System as the means of access.

28. In order to ensure the security and confidentiality of data collection during the online census, the plan is to have respondents undergo a three-phase authorization procedure:

- Respondents will first undergo authorization through the Uniform System’s records. Their authority will be verified by virtue of their ability to access the full range of public services in electronic form.
- They will then pass a Turing test (CAPTCHA), thus protecting against the entry of data by machines.
- They will then be identified using unique residence and respondent numbers, thus establishing a link between the respondents and all their electronic census sheets with the list of addresses of residences for the 2020 national population census.

29. Other types of protection may also be used, as follows:

- a personal identification code, making it possible to fill in the online census sheet in several Internet sessions (i.e., in more than one session)
- a secret question created by a member of the household and making it possible to interrupt and resume filling in the online census sheet

30. For responses submitted online, a number of methods will be developed for data quality control and for correcting the data: automatic sequencing of corresponding questions, interactive editing to detect problematic responses and the use of drop-down lists so as to ensure that entries are in a permissible format. In some cases, other methods, such as automated coding, will be used.

31. For the use of tablets, several security-related issues must be resolved beforehand:

- How will the data be stored (on the device, or with immediate transfer to the central system)?
- How will the data be transmitted from the personal computer to the workstations at the regional level and from the regional server to the central server?
- How will the security of data stored in personal computers be ensured?

32. Android, a widely distributed free-of-charge operating system which has been successfully used by the Federal State Statistics Service in its surveys since 2012, is slated for installation on the personal computers for the 2020 population census.

33. In accordance with the national requirements for the protection of mobile equipment used in government computer systems, in addition to the basic Android operating system, the tablets will be equipped with a separate Android protection system:

(a) that does not allow the installation of third-party, unauthorized software without the consent of the Federal State Statistics Service;

(b) that makes it possible to work without a connection and without Bluetooth and Wi-Fi wireless networks;

(c) that ensures the confidentiality of the information on the population.

34. If a personal computer is lost or stolen, the question of how to ensure the security of the data stored on it becomes critical.

35. In such cases, two means of protection will be used.

36. First, the data will be encrypted. All data located on the device will be in encrypted form, and a decryption “key” will be necessary to view such information. There is also the possibility that once the form has been filled in, the information will be encrypted so that even the enumerator will be unable to enter corrections.

37. Secondly, passwords will be used for protection. Access to the software will be granted by password, to the enumerator responsible for the reporting district in question. An alternative to the use of passwords is the use of a fingerprint reader. Such devices provide a high degree of protection but significantly raise equipment costs.

38. To ensure the safe-keeping of the information collected on the personal computers during the census, the devices will be supplied with removable micro SD memory cards for backing up the information. The enumerators will periodically hand them in to their monitors in exchange for blank memory cards. Information from the tablets will be transferred into the system using wire (USB) connections to workstations on closed, protected, certified circuits, or on SD cards with the use of certified notebook computers.

39. Channels of communication within the automated system are protected by an information encryption system.

40. For the 2020 population census, the system used for the 2010 census will have to be perfected and inevitably tested with a trial run, in 2018. In addition to updating existing software and equipment, entirely new units and modules will have to be developed for:

- automation of functions at the field level (census, instructor and reporting district levels) and at the level of online census data collection
- multimedia training tools, including for online respondents
- collection and processing of online data
- data protection hardware and software

41. Innovations in the population census thus pose new challenges for the protection of data confidentiality. They introduce computerization not only at the data processing stage, as was the case for the traditional census, but at all stages of the preparation and implementation of the census in the field. This has a significant impact on the overall cost of the census, making its technological component more expensive.