United Nations ECE/CES/ GE.41/2012/13

Economic and Social Council

Distr.: General
23 April 2012

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Fourteenth Meeting
Geneva, 24-25 May 2012
Item 5 of the provisional agenda
Internet data collection

Internet data collection in the 2011 Population and Housing Census in the Czech Republic

Note by Czech Statistical Office[1]

Summary

Population censuses on the territory of the contemporary Czech Republic have been performed for several centuries. But the 2011 Population and Housing Census offered, for first time, the option to use electronic questionnaires and the Internet. All inhabitants and owners of houses and dwellings could decide if they would use paper or electronic questionnaires. At the first visit of an interviewer every household got a set of paper questionnaires. Every questionnaire was bearing a special PIN for opening of the original electronic questionnaire over the Internet.

The whole system of electronic questionnaires was developed and operated by an external IT company. The supplier received main requirements for the system in terms of its capacity and safety. The delivered actual capacity of the electronic system was never used at more than 1/3 thereof at a time and no safety accident occurred.

All users could use both paper and electronic questionnaires at the same time for four weeks. This fact was the core of an educational campaign. The electronic questionnaires were used by 27.5% of inhabitants and the share of electronic questionnaires in all completed questionnaires was 25.5%.

[1] Prepared by Stanislav Drápal.

I. Electronic Forms

1. When designing complex solutions for electronic data collection systems, the crucial elements are the type of e-forms and the technology applied within the e-form system. Like conventional paper forms, e-forms are used for collecting various types of data. It is important that the e-forms look very much like the paper forms. In order to fill in the e-forms properly, a variety of interactive elements are offered (text fields, check boxes, soft-keys, etc.). Their appearance and functionality are very close to what users already know from their everyday activities. In addition, the advantages of electronic processing are included (for example, e-forms may be automatically populated with a part of the data collected from other sources, etc.).

2. Further crucial elements of the solutions for e-forms are the technologies necessary for distribution of the forms, as well as for the collection of them. Among these technologies are, for example, the presentation of e-forms on the given website, verification of the data entered by users, interaction with the user, transfer of the data to databases or another data repository, confirmation sent to the sender, and much more.

II. Electronic Forms, their Types and Technological Fundaments

3. It was agreed that the best intersection between functionality and openness seems to be the Adobe technology, whose e-forms have traditionally been connected with the PDF format and the Adobe Acrobat application. Using the application, you can create forms whose control elements and interactivity compare to, or are better than, the options offered by HTML forms (for example, the print output, the saving or the self-support, which makes e-forms equivalent to paper forms). In addition, the possibility to apply PDF forms in combination with technologies for electronic signature and security was really worth the attention. Moreover, PDF representation options are better than those in HTML (and in other form-related technologies as well), which makes the PDF platform even more attractive. Finally, the application you need in order to view and fill in the PDF forms, the multi-platform Adobe Reader, is available for free.

III. Solutions for e-Forms

4. A solution based on the technology of Adobe PDF forms was used for distribution and collection of e-forms during the 2011 Population and Housing Census. The reason why this solution was preferred, when comparing with online HTML forms, was the complex fulfilment of the following requirements:

- Minimised communication flow – during the research done before the census, a significant interest of the respondents [2] in online submissions was identified.
- Self-support of the forms – each form must comprise all functionalities and complex information so that the respondents can fill in their answers correctly, with no need of any interaction with central systems.
- Identical look – It was required that the e-form and the paper questionnaire must be alike so that the general public can understand the instructions and fill in the

2 Approximately 10 % of respondents

data easily. The respondents receive the same instructions before they decide which type of the questionnaire to use.

- Evidentiary origin – electronic forms must comprise a mechanism to prove their origin, in other words, to demonstrate that this is the questionnaire provided by the CZSO (Czech Statistical Office).
- Technological consistency – the technology to be used by the public when filling in the questionnaires must be widely known, free and available throughout standard platforms so that the questionnaires can be used consistently by the widest public.
- Privacy policy – the technology must support the protection of personal data provided by the respondents, and it must be clearly indicated that the data are not saved in the computer without the respondent's knowledge. This approach opens an option to use computers available for the public in study rooms, libraries and municipal offices for citizens who cannot use their own computers.
- Confirmation of submission – the solution must comprise a functionality to confirm that the respondent has filled in the questionnaire, which may be saved in electronic form or printed out, and cannot be falsified.

5. Thus the solution following from the Adobe PDF technology was divided into the following segments, which were addressed separately, and at the end, the partial solutions were integrated in order to form a complex solution.

- Preparation of questionnaires – development of PDF documents that represent the specific types of questionnaires including the required functionalities and options and the possible final check by the respondent.
- Pre-population of the questionnaires – preparation of a mechanism which will pre-populate respondent's questionnaires in conformity with the paper ones.
- Information presentation – provision of information related to downloading, filling in and submitting the questionnaire.
- Distribution of questionnaires – the electronic questionnaire is only available to the respondent whose data are filled in.
- Collection of questionnaires – collection of the data provided by filling in the form and confirmation of the submission.
- Technological infrastructure – development of necessary technological infrastructure to ensure the distribution and the collection in conformity with security requirements on its operation.

IV. Preparation of Enumerating Questionnaires

6. The electronic questionnaires were developed as separate PDF documents for each type of form, i.e. the census person questionnaire, the buildings questionnaire and the dwellings questionnaire. The questionnaires were elaborated so that they very much looked like their paper versions. In addition to the unification of the information provided to the respondents in the forms of either type (electronic or paper), this similarity covered the situation when a user prints a blank electronic questionnaire, fills in, and hands it over as the paper questionnaire: the identical look allowed digitising.

7. The questionnaires included a program code to offer contextual help for each field to be filled in. In addition, the code comprised verification of the data at the level of the

specific fields (for example, you cannot enter letters in a field which requires numbers), then there was a logic check across conditioned fields (e.g. sex versus the number of children – a male cannot give birth to a child) and, in case of some specific fields, a selection of values to be filled in the form was offered. Before sending the data to the CZSO, an automatic check for the completeness of the data filled in the form has been made. The above-mentioned checks greatly increased the quality of the data collected by means of electronic questionnaires.

8. The forms were signed by an electronic certificate issued for the CZSO by VeriSign, a renowned certification body. The public was informed in necessary detail which way the author of the document may be verified in the Adobe Reader application. This way, the uncompromisingness and trustworthiness of the downloaded questionnaire was provided, eliminating the danger of attacks.

9. With respect to the requirements for functionality, PDF documents designed for relatively new versions of the Adobe Reader application were used. It proved to be a good decision to apply the Adobe technologies as the respondents had absolutely no problems when upgrading or installing Adobe Reader, which confirmed the advantages of the universal tool, in comparison to (previously considered) solutions based on HTML or X-Forms.

V. Pre-populated Data in the Questionnaires

10. With regard to the fact that the paper questionnaires were provided to the respondents including the data already known from administration sources in order to make the filling in process easier and improve the quality of the collected data, this concept was applied in the case of the electronic questionnaires as well. In addition, each electronic questionnaire was unambiguously matched to the respondent.

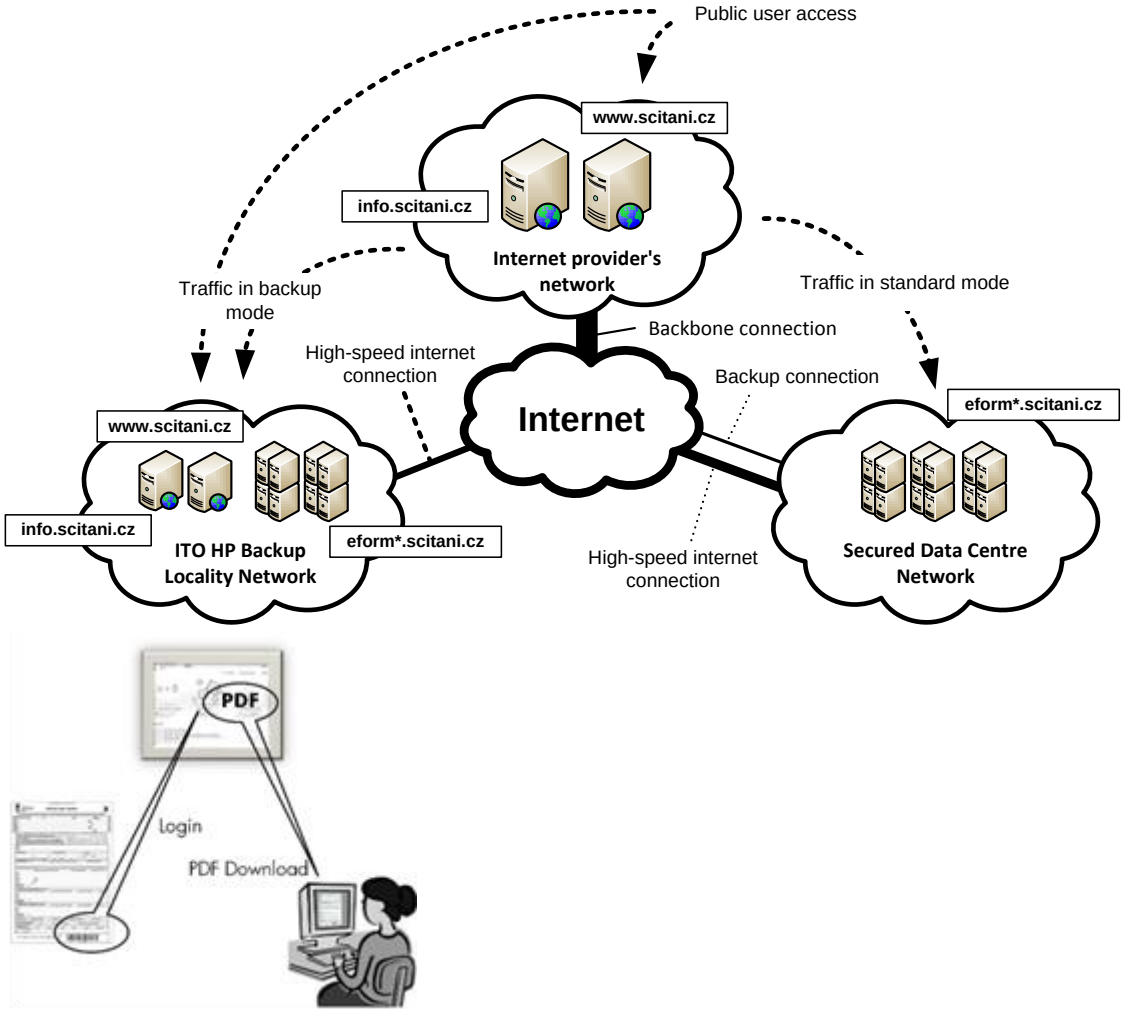
11. To save time, during the distribution phase, the electronic questionnaires were generated as pre-populated before the census was started. Approximately 20 mil [3] questionnaires were proactively pre-populated and saved to a 4-TB database, to be distributed afterwards. The pre-population task was run continuously for several days.

VI. Information Presentation

12. During the census preparation phase, the Office launched a special WWW presentation for the public, with information about the distribution, filling in the forms, and collection of submitted questionnaires.

13. It was anticipated that the public would be interested in the online census so the presentation was divided into two parts – a frontend web server, called “signpost” (website www.scitani.cz) with general information and the second server with detailed presentation (website info.scitani.cz).

3 It was not known which respondents would use the electronic questionnaires, so all forms had been pre- populated.

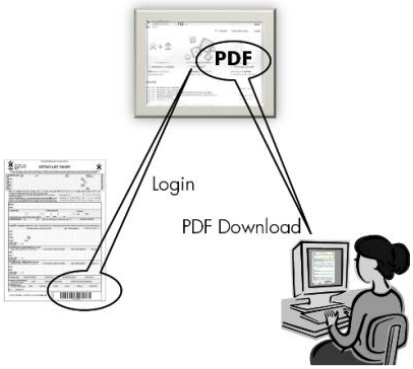


14. The two parts were presented separately, using a hosted infrastructure with multiple localities within the Backbone Internet network, administered by different Internet service providers (refer to the area marked in the image as "Internet provider's network"). By reason of personal data protection, the Office's infrastructure was used for the distribution and collection of the questionnaires.

15. The distribution and the collection were conducted from two geographically separate localities. Due to the above-mentioned mechanism, the load upon the technological devices operated by the CZSO was reduced (the area marked in the image as "Secured Data Centre Network") and a distributed infrastructure (the area marked as "ITO HP Backup Locality Network") was developed for provision of the information, with a robust base to protect the system against failure and hacker attacks targeted at the presentations.

VII. Distribution of Questionnaires

16. In terms of the solution for the distribution of electronic questionnaires, the key issue was to make sure that a citizen will download the form with pre-populated data related just to themselves, their apartment or their house.

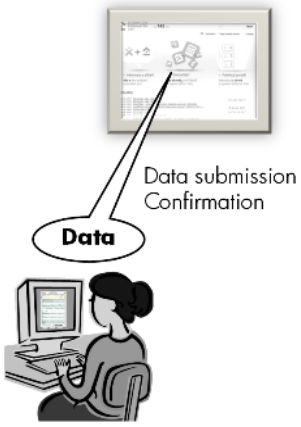


17. It was agreed that the respective mechanism will be based on the fact that each citizen should receive the paper questionnaire. The form comprised the questionnaire ID and a security code. Both codes were chosen by random from a relevant set of codes. The set was large enough so that the codes really used were random, and they were distributed sparsely. This way, prediction of unknown codes based on the codes already known could be avoided.

18. When a citizen wanted to submit the questionnaire online, they used the codes from the paper questionnaires, entered them in the Internet application, downloaded the electronic equivalents of the paper questionnaires into their computers, and filled in the data.

VIII. Collection of Questionnaires

19. With respect to the fact that the size of the data provided by the respondent was, within the context of the form, hundred times smaller than the capacity of the Adobe PDF forms, it was agreed that the CZSO would not receive the completed form but only the data provided by the respondents.



20. For this purpose, a functionality included in the Adobe Reader application was used enabling the data to be sent as an XML document by means of secured (HTTPS) protocol. Thus the data were received as short XML messages, to be further processed.

21. After data submission and successful reception, the functionality asked for the download of another PDF document comprising an electronic confirmation of the data submission. This was implemented by sending a random confirmation code which was connected with the specific questionnaire. The respondent could save the confirmation in the computer or print it out.

IX. Technological Infrastructure

22. The technological infrastructure for distribution and collection of electronic questionnaires was dislocated to three localities:

i. The primary technological infrastructure for applications designed for downloading electronic questionnaires and collecting the data from the questionnaires was operated at a secured Data Centre, hired by the CZSO for implementation of the Population and Housing Census.

ii. Backup technological infrastructure for applications of electronic questionnaires downloading and respondents' data collecting was operated by means of a secured "Cloud" service by the information technologies general supplier.

iii. The infrastructure for public presentations was implemented by hiring the technologies within the Backbone network from the Internet operators.

23. In terms of security, the key element was the primary technological infrastructure. It comprised separate secured zones in compliance with the standard architecture of so-called demilitarised zones, which forms a safe buffer area between the part of the architecture available to the public and the private part of the architecture.

24. The databases and the demilitarised zones were only connected via passages initiated from the database layer. Thus the collected data were "pseudo-online" drawn from the demilitarised zone to internal databases by means of a data pump.

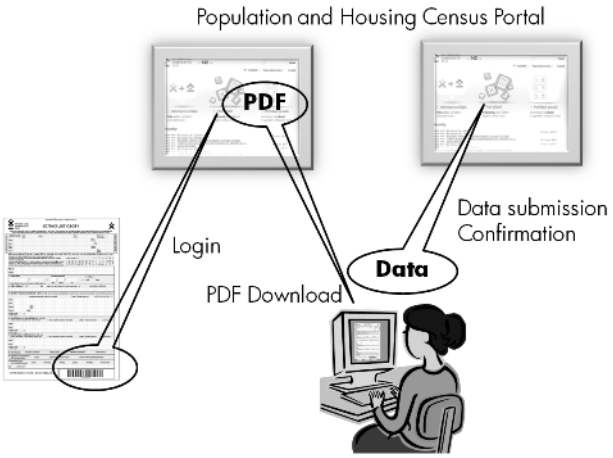
25. The "front-end" of the distribution and questionnaire collection services included a farm comprising five servers. One of the servers was dedicated to deal with requirements coming from IP addresses associated with higher probability of DDoS (Distributed Denial of Service) attacks, mainly from Asia. In case of a successful attack of this type, only one server would have been jammed but the rest of the server farm would continue the distribution and collection services. As the Population and Housing Census was running within the territory of the Czech Republic, the access from the afore-mentioned IPs should be nearly zero. Any higher traffic from these IPs would mean an attempt for attack.

26. Another security level, in terms of unauthorised access, was an activity connected with an external firewall, supported by IDS (Intrusion Detection System) based on the HP Tipping-Point technological platform. The purpose was to provide proactive identification of attacks against the services provided.

27. The backup technological infrastructure was following from the primary infrastructure architecture, except for the fact that virtual technological elements (servers and network devices) were used. In order to reduce the security-related risks, the backup infrastructure was kept in the "hot standby" mode, to be able to redirect the operation to the backup locality with minimum impacts in case the primary infrastructure fails or is overloaded.

X. Recapitulation from the View of the User

28. From the view of the user, the key elements of the complex solution were maximum safety and security, transparency and user-friendliness, both from the view of the complex questionnaire distribution and data collection and from the view of the control of the questionnaire as such.



29. Within the afore-mentioned requirements, the designed system was, in fact, very simple. It means that it was both robust and user-friendly and, from the respondents' view, it included only a few activities as illustrated in the above-mentioned image.

i. There was a PDF electronic form for each printed paper questionnaire. Thus, during the distribution phase, citizens just entered the numbers and security codes from their paper questionnaire, to which the relevant PDF counterpart was matched in the database, and the specific questionnaire was provided to the specific citizen by means of a secured connection (HTTPS).

ii. The citizen subsequently filled in the required data in the pre-populated PDF questionnaire, whose look was nearly identical to the paper one. The form was provided with functionalities to run logic verification of the values entered in selected fields, and other field values that needed to be filled in were offered by drop-down menus.

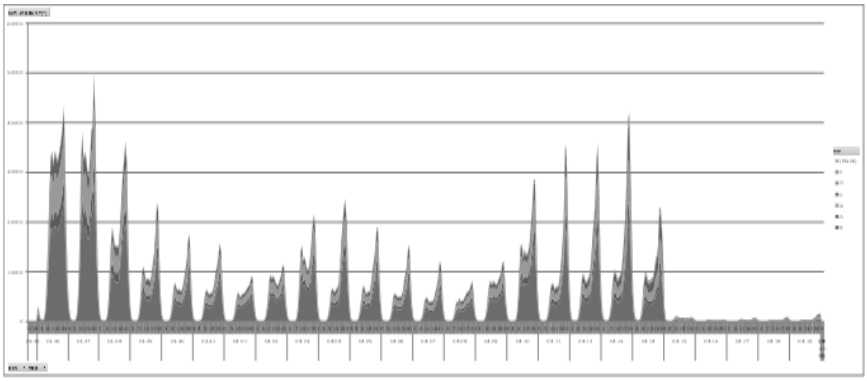
iii. The checked content (not the complete questionnaire) i.e. the very minimum sized data, when compared to the data size of the questionnaire, was sent via a secure (again) channel to the CZSO input database for further processing.

iv. The reception of the data from the submitted questionnaire was confirmed by a new PDF document containing a unique confirmation of the submission.

XI. Electronic Forms – Results and Evaluation

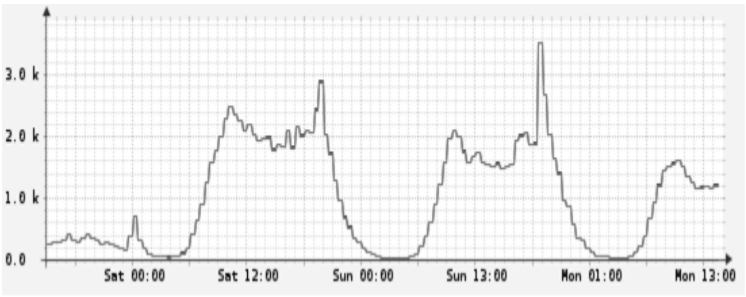
30. During the 2011 Population and Housing Census, for the first time, the citizens of the Czech Republic could fill in electronic forms via the Internet network. In the course of the Population and Housing Census 2011, the CZSO received 4.33 million electronic questionnaires in total, which includes 2.8 million personal pages, 1.05 million household pages and 0.48 million house pages. The share of the e-forms within all submitted questionnaires reached 25.5%.

31. The highest number of e-forms (509 thousand) arrived on the first day of the collection – on Saturday, 26 March. Additional nearly half a million followed on Sunday, 27 March. During the first week, 44% of all questionnaires submitted via the Internet arrived. The fluctuation, in terms of frequency of the submissions related to the time, during the whole collection period, is shown in the following image:



32. The number of the submissions depended on the days of the week. The biggest numbers of the e-forms were received during weekend days (approx. 40% of all received e-forms), mainly on Sundays (968 thousand of the e-forms). On the contrary, the lowest number of the submissions was received on Fridays (200 thousand).

33. The preferred times of the day were the evening hours. Between 6 p.m. and 10 p.m., nearly 40% of all e-forms were delivered – it was over 300 thousand each hour within this interval. The most intensive activity was found between 8 p.m. and 10 p.m., when 939 thousand of the forms were received. The forms were submitted at any time day and night. The typical course of the submission process is shown in the following image:



34. For example, between 1 a.m. and 2 a.m., nearly 13 thousand e-forms were submitted, while between 4 a.m. and 5 a.m. it was more than 2 thousand of them.

35. Nearly half of the personal and household pages were returned back within 10 minutes from the download. In the case of house pages, over 90% of the questionnaires were returned within 10 minutes.

36. From the view of the architecture, the success, in terms of the use of the electronic questionnaires within the 2011 Population and Housing Census, was achieved due to several key factors. It was mainly due to the application of the Adobe PDF technology, which proved to be suitable in terms of meeting the requirements for the data collection, and it was widely acceptable by the public.

37. In connection with e-forms and their possible application, the key questions were the real interest in the electronic enumeration method, its time distribution, and its impact upon the technological infrastructure. At this point, obviously the most important parts were planning and work with media. The highest load upon the systems was seen at times when the media, during the prime time news, informed about the legal obligation to participate in

the census and fill in the questionnaires. However, due to CZSO's controlled work with the media, this was predictable.

38. The results from the monitoring, as regards the load upon the technological infrastructure, show that when it comes to implementation of a task like this, the performance parameters of the computing systems are sufficient.

39. The protection system used during the distribution and collection of the e-forms successfully turned away 3,760 different attempts of attack. The most frequent type of attack was Cross-site scripting (XSS), which is the method of penetration into WWW pages through security errors in the script. In addition, unsuccessful DoS attack attempts were identified as well. In general, the anticipated security-related risks have not appeared, and the public accepted the Internet questionnaires in a positive way. People were aware of the fact that this service saves their time and helps them fulfil their legal obligation easier.